Analiza exploratorie a datelor (EDA)

Day 1 - Mihai Trăscău

Cuprins

EDA! Cum?

- **Orice proiect** din ML **începe cu Analiza Exploratorie a Datelor**

„Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone – as the first step.” (John W. Tukey, Exploratory Data Analysis, 1977)

- Explorarea **datelor** pleacă de la înțelegerea **task**-ului. Câteva exemple:
  - Pentru clasificare supervizată, verificăm cât de balansate sunt clasele
  - Pentru regresie, verificăm distribuția exemplele în raport cu valoare țintă
  - ...
- Statisticile și metricile extrase au rolul de a scoate în evidență
  - Anomalii
  - Erori
  - Trenduri
  - Distribuții
  - ...
  - Indicii despre ce modele ar fi mai potrivite

ML mantra: ***„Garbage in, garbage out.”***

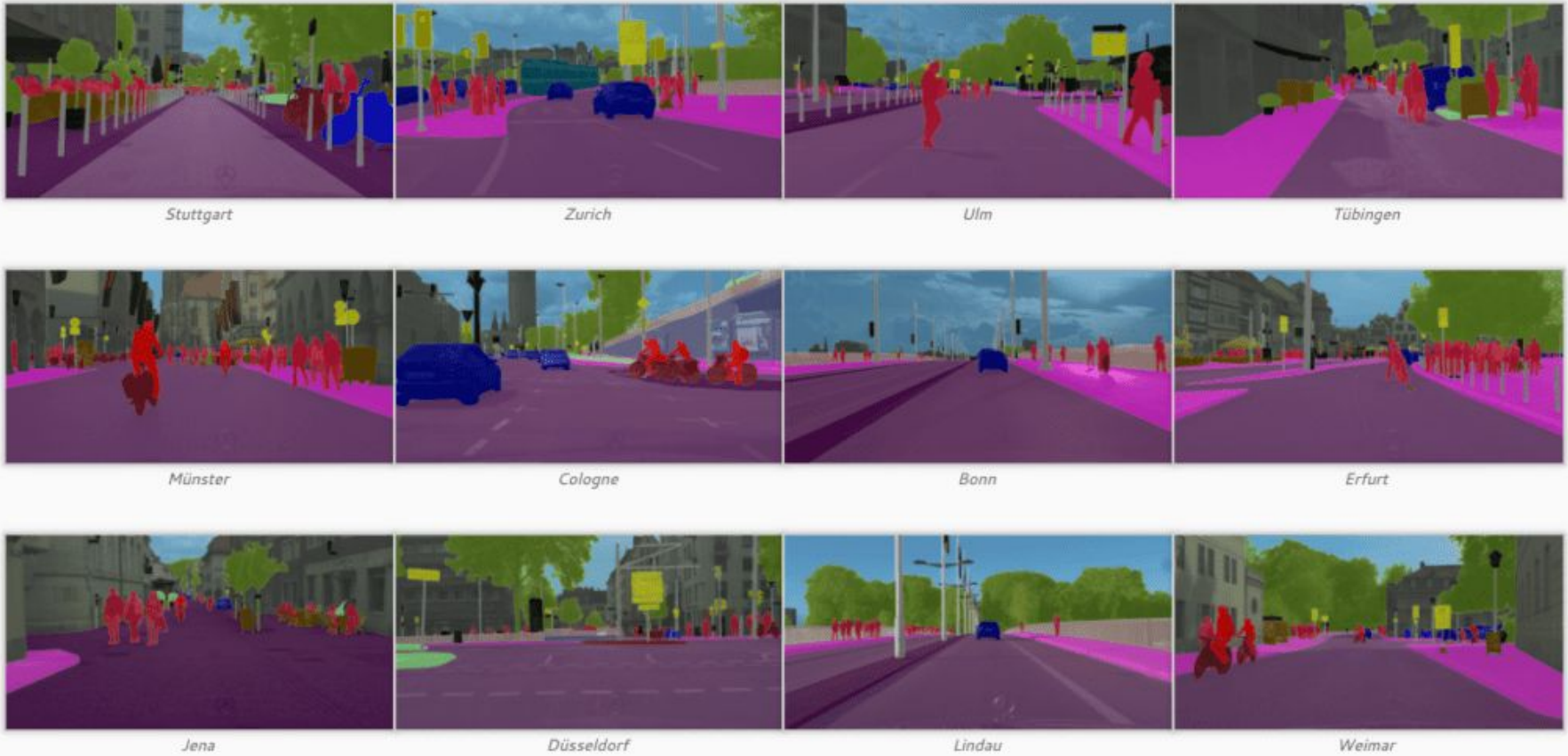
Tipuri de date în ML (I)

- 14 exemple din 2 clase (*Yes/No*)
- 4 atribute cu domeniile de valori asociate lor:

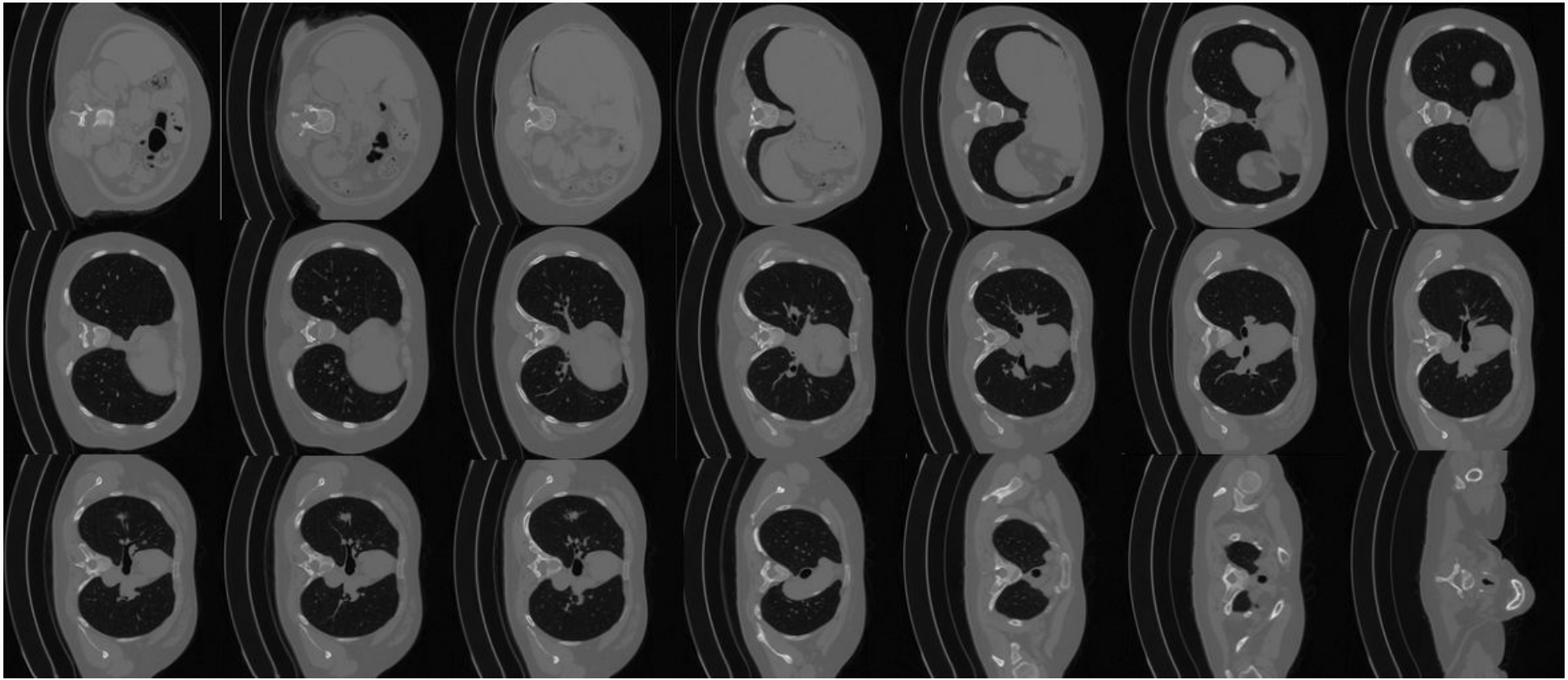
Outlook: { *Sunny*, *Overcast*, *Rain* }
Temperature: { *Hot*, *Mild*, *Cool* }
Humidity: { *High*, *Normal* }
Wind: { *Strong*, *Weak* }

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Tipuri de date în ML (II)



Tipuri de date în ML (III)



Covarianță

- Definiție: Măsură numerică ce descrie relația statistică dintre două variabile numerice pentru care exprimă direcția de variabilitate comună ale acestora.

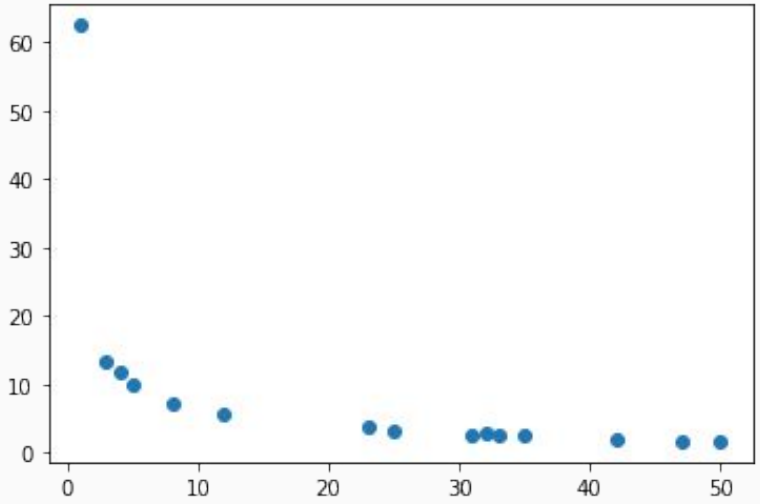
- Pe scurt:
  - Când valorile uneia dintre variabile *cresc* iar valorile celeilalte variabile *cresc* de asemenea, ele au o ***covarianță pozitivă***
  - Când valorile uneia dintre variabile *cresc* iar valorile celeilalte variabile *scad*, ele au o ***covarianță negativă***
  - Când valorile celor două variabile *nu manifestă niciun fel de relație statistică* valoarea de ***covarianță este 0***

$$\operatorname{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y)).$$

- Deși utilă, valoarea de covarianță este greu de interpretat (în relație cu alte valori de covarianță)
- Intervalul valorilor este *(-∞, +∞)*
- Presupune relaționare liniară între variabile
- Este utilizată ca instrument ce face parte din diverse alte metode (de exemplu, PCA)

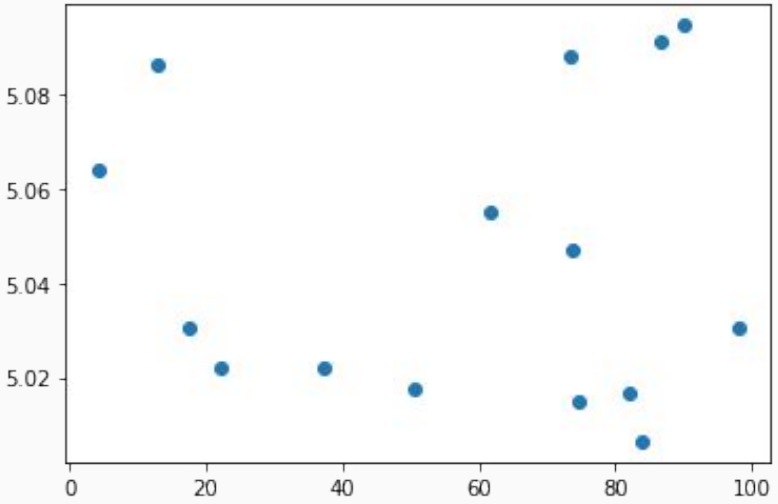
Covariantă negativă

x	y	$x - \bar{x}$	$y - \bar{y}$	Cov
1	62.48347	-22.4	53.56717	-79.9936
3	13.326	-20.4	4.409698	-5.99719
4	11.86829	-19.4	2.95199	-3.81791
5	10.05277	-18.4	1.136469	-1.39407
8	7.286517	-15.4	-1.62979	1.673246
12	5.604744	-11.4	-3.31156	2.516784
23	3.710914	-0.4	-5.20539	0.13881
25	3.365941	1.6	-5.55036	-0.59204
31	2.570636	7.6	-6.34567	-3.21514
32	2.766448	8.6	-6.14985	-3.52592
33	2.681775	9.6	-6.23453	-3.9901
35	2.497095	11.6	-6.41921	-4.96419
42	2.028371	18.6	-6.88793	-8.54103
47	1.790305	23.6	-7.126	-11.2116
50	1.711248	26.6	-7.20505	-12.777
				-135.691



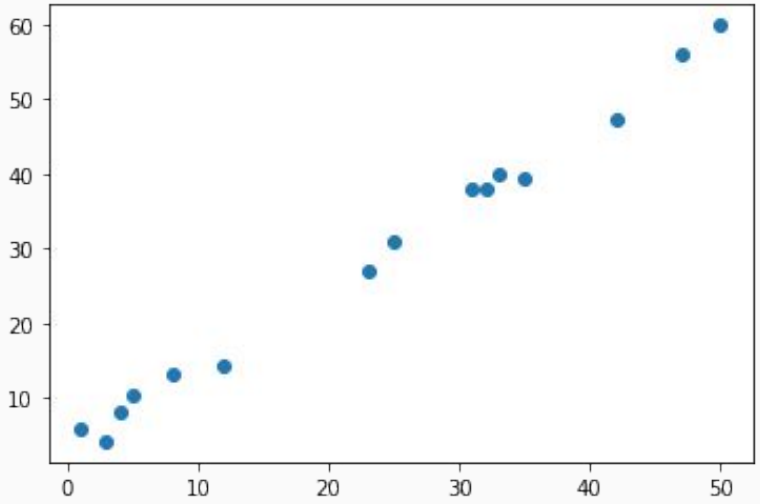
Covariantă zero

x	y	$x - \bar{x}$	$y - \bar{y}$	Cov
22.104	5.022076	-35.8388	-0.02365	0.056505
4.205967	5.063946	-53.7368	0.01822	-0.06527
90.13109	5.09447	32.18829	0.048744	0.1046
12.93337	5.086122	-45.0094	0.040396	-0.12121
73.42461	5.087704	15.48181	0.041979	0.043327
73.82211	5.046932	15.87932	0.001206	0.001277
37.34157	5.022141	-20.6012	-0.02359	0.032392
81.93614	5.016565	23.99335	-0.02916	-0.04664
61.59471	5.055092	3.651916	0.009366	0.00228
17.70031	5.030458	-40.2425	-0.01527	0.040962
74.71551	5.014836	16.77272	-0.03089	-0.03454
83.87204	5.006528	25.92925	-0.0392	-0.06776
50.6937	5.01768	-7.2491	-0.02805	0.013554
98.06425	5.030388	40.12145	-0.01534	-0.04102
86.60257	5.090949	28.65977	0.045223	0.086406
				0.004847



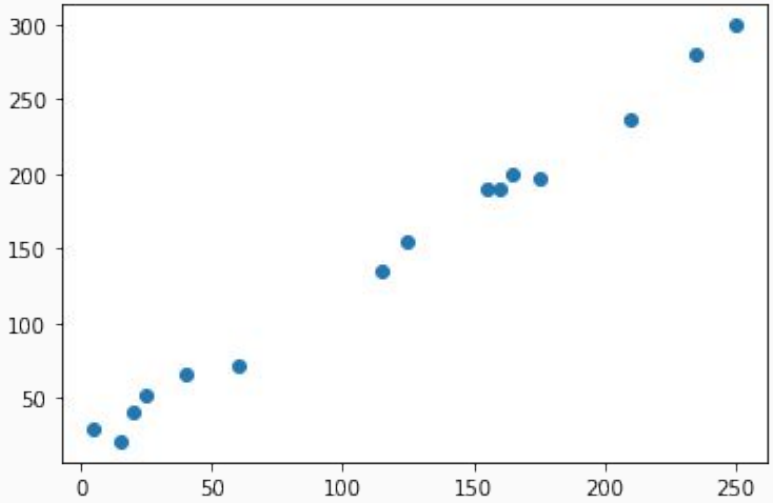
Covariantă pozitivă

x	y	$x - \bar{x}$	$y - \bar{y}$	Cov
1	5.824876	-22.4	-23.031	34.39294
3	4.212997	-20.4	-24.6429	33.51429
4	8.089474	-19.4	-20.7664	26.85786
5	10.37351	-18.4	-18.4823	22.67168
8	13.31944	-15.4	-15.5364	15.95072
12	14.26162	-11.4	-14.5942	11.09162
23	27.12442	-0.4	-1.73144	0.046172
25	31.04781	1.6	2.191947	0.233808
31	37.9035	7.6	9.047636	4.584136
32	37.90247	8.6	9.046608	5.186722
33	39.88369	9.6	11.02783	7.057814
35	39.53995	11.6	10.68409	8.262361
42	47.28982	18.6	18.43396	22.85811
47	56.13006	23.6	27.2742	42.9114
50	59.93427	26.6	31.07841	55.11237
				290.732



Covariantă pozitivă (2)

x	y	$x - \bar{x}$	$y - \bar{y}$	Cov
5	29.12438	-112	-115.155	859.8234
15	21.06498	-102	-123.214	837.8574
20	40.44737	-97	-103.832	671.4465
25	51.86756	-92	-92.4117	566.792
40	66.59721	-77	-77.6821	398.7681
60	71.30809	-57	-72.9712	277.2906
115	135.6221	-2	-8.65718	1.154291
125	155.239	8	10.95973	5.845191
155	189.5175	38	45.23818	114.6034
160	189.5123	43	45.23304	129.668
165	199.4185	48	55.13917	176.4454
175	197.6997	58	53.42044	206.559
210	236.4491	93	92.16982	571.4529
235	280.6503	118	136.371	1072.785
250	299.6713	133	155.392	1377.809
				7268.3



- Deși variabilele sunt doar rescalate și variază la fel una față de cealaltă, pentru covariantă obținem valori cu totul diferite

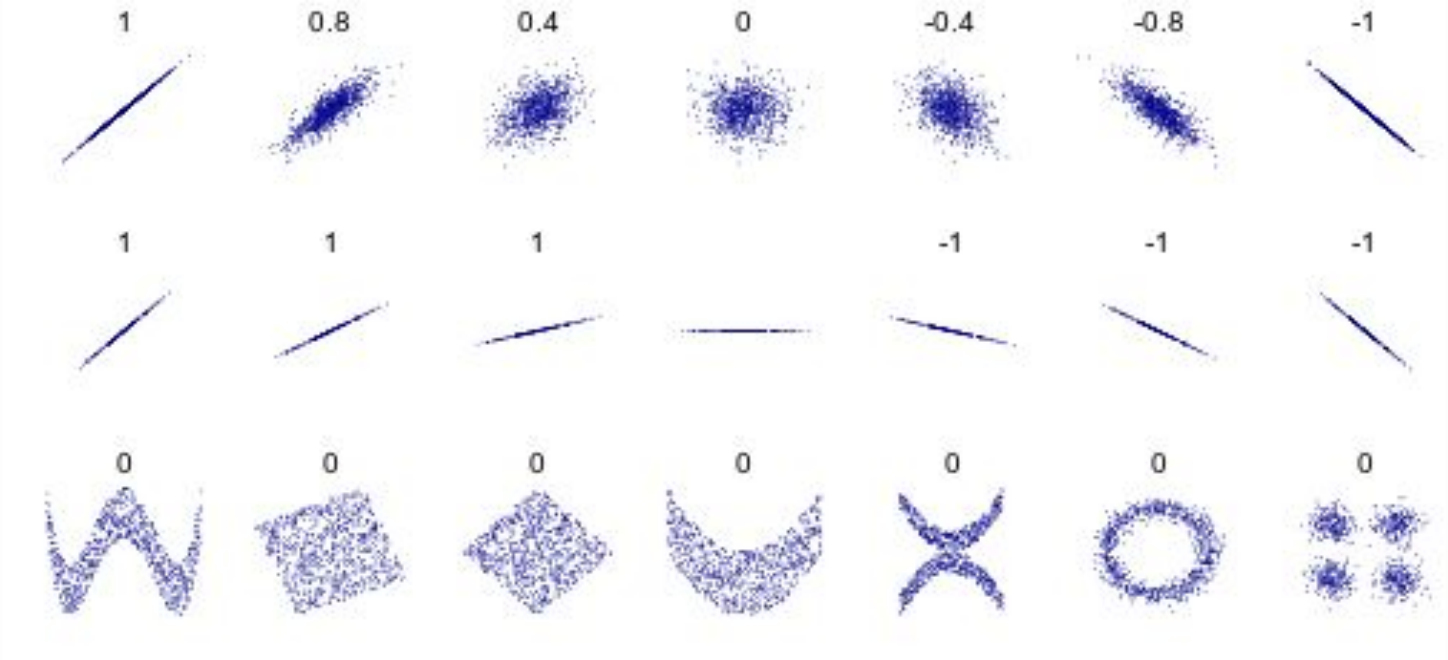
Coeficient de corelație

- Definiție: Măsură numerică care descrie relația statistică dintre două variabile care exprimă direcția și tăria variabilității comună ale acestora.

- Există mai multe tipuri de coeficienți de corelație care au fost propuși
- Diferă după tipul de date și de relații între variabile asupra cărora se aplică:
  - Coeficient Pearson - pentru variabile numerice cu corelație liniară
  - Coeficient Spearman - pentru variabile de rang cu funcție monotonă de corelație

- Ca domeniu de valori se situează în intervalul *[-1, +1]*
  - Valori apropiate de *-1* descriu o corelație negativă (când valoarea unei variabile crește, cealaltă scade)
  - Valoarea *0* denotă lipsa vreunei corelații
  - Valori apropiate de *1* descriu o corelație pozitivă (ambele variabile cresc în același timp)

Coeficient de corelație



Exemple de vizualizări pentru perechi de variabile pentru care a fost calculat coeficientul de corelație Pearson (afișat deasupra graficului)

Coeficient de corelație Pearson

- Definiție: Covarianța a două variabile numerice normalizată prin produsul deviațiilor standard ale acestora.

- Coeficientul de corelație Pearson se notează cu *r* sau cu *ρ*

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Coeficient de corelație Pearson

- Limitări:
  - Nu distinge între cele 2 atribute care este „*cauza*” și care este „*efectul*”
  - Operează doar pe variabile (atribute) continue
  - Nu poate fi utilizat pentru a analiza relații non-lineare între atribute
  - Distribuția valorilor pentru atribute este (aproape) normală (Gaussiană)
  - Nu indică „panta” corelației (ci doar dispersia în jurul trendului)
  - Datele trebuie să fie *IID* (Independent and Identically Distributed)
  - Trebuie tratate valorile lipsă
  - Presupune absența datelor cu anomalii

Coeficient de corelație Spearman

- Definiție: Măsură numerică care descrie relația statistică *monotonă* dintre două variabile *de rang* și care exprimă direcția și tăria variabilității comună ale acestora
- Formula este similară cu cea a coeficientului de corelație Pearson
- Calculul coeficientului de corelație Spearman se realizează astfel
  - Se sortează crescător valorile din setul de date pentru prima variabilă (X_i)
  - Se notează rangul pentru fiecare exemplu în parte (de la *1* la *N*, unde *N* este numărul de exemple); dacă fiecare exemplu are un rang unic, putem utiliza formula (2)
  - Se repetă procedeul și pentru a doua variabilă (Y_i)
  - Se calculează diferența de rang între cele două variabile (d_i) și pătratul acesteia (d_i^2)
  - Calculăm coeficientul de corelație Spearman folosind formula (2)

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\operatorname{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}}, \quad (1)$$

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad (2)$$

IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
86	2	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Selecție de atribute

- Definiție: Procesul prin care selectăm și păstrăm doar *atributele cele mai relevante* din setul de date dat în vederea obținerii unor *performanțe mai bune*

- Atributele cele mai relevante se determină prin diverse metode:
  - Filtrare
    - Utilizăm diverse metrici pentru a determina o ordonare a atributelor
    - Păstrăm doar primele *k* atribute
    - Exemplu de metrică: coeficientul de corelație Pearson (între atribut și valoarea țintă)
    - Eliminare de colinearități (atribute care au aceeași contribuție, deci corelație foarte mare)
  - Evaluare
    - Selecția celor mai bune *k* atribute este descrisă ca o problemă de căutare
    - Combinația cea mai bună de atribute este descoperită prin evaluarea modelelor
    - Există metode aditive (forward selection) sau subtractive (backward selection)

Selecție de atribute

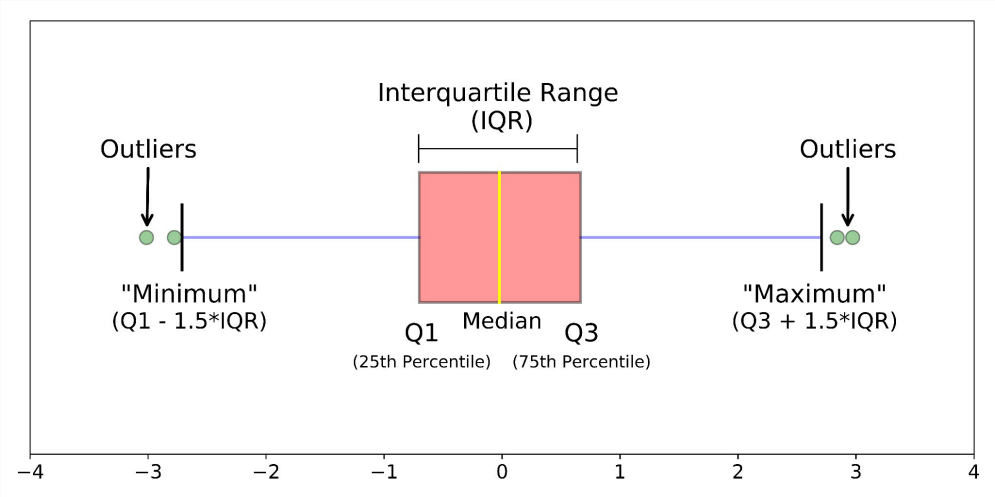
- Atributele cele mai relevante se determină prin diverse metode:
  - Regularizare
    - În timpul antrenării algoritmului, la fiecare iterație se aplică metode prin care sunt potențate atributele relevate și suprimate cele nerelevante
    - De exemplu, putem modifica funcțiile de cost ale algoritmilor prin adăugarea de noi termeni de penalizare
  - Reproiectare
    - Transformarea spațiului într-unul cu mai puține dimensiuni
    - Exemplu, algoritmul de analiza componentelor principale (PCA)

- Performanțe mai bune:
  - Acuratețe mai mare
  - Timp de execuție mai mic
  - Interpretabilitate mai bună

Date cu anomalii (outliers)

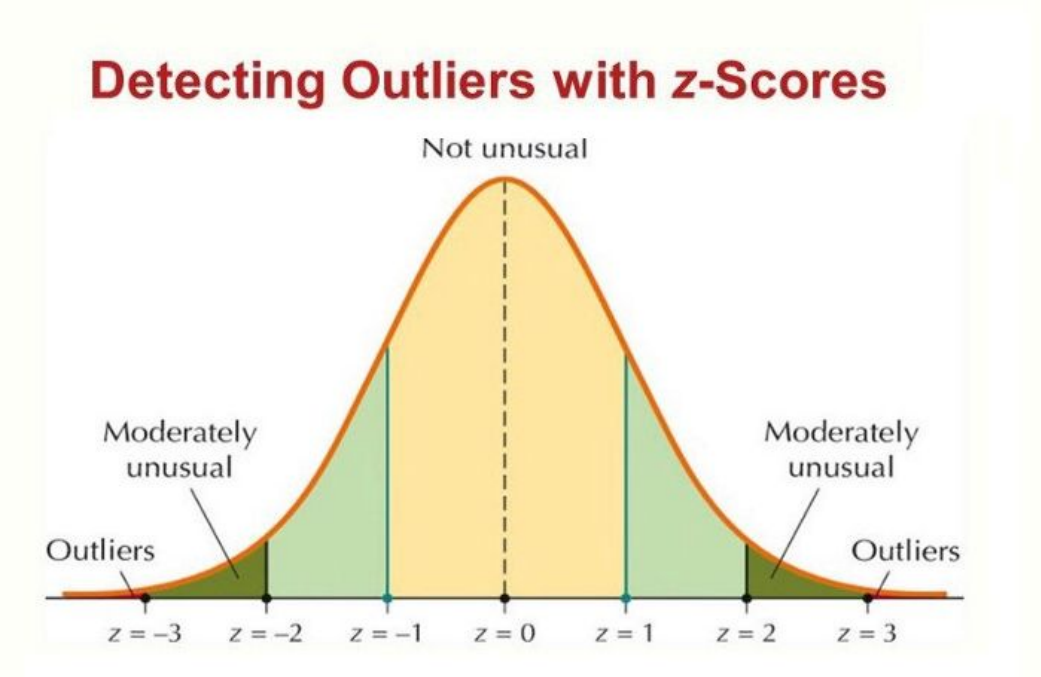
- Definiție: O observație dintr-un set de date este considerată o anomalie dacă valoarea ei este îndepărtată de valorile celorlalte observații

- Sursele de erori care generează anomalii pot fi variate și multiple
- Datele cu anomalii pot influența negativ performanțele modelelor aplicate
- Pentru a identifica datele cu anomalii putem aplica diverse metode:
  - Vizualizare cu „Box Plots” (și calcularea quartilelor)



Date cu anomalii (outliers)

- Pentru a identifica datele cu anomalii putem aplica diverse metode:
  - Calculul scorului Z
    - Valorile care depășesc 3 deviații standard sunt considerate anomalii
    - Presupune distribuție normală (Gaussiană) a valorilor pentru atributul analizat



Rescalarea atributelor

- Definiție: Metode prin care domeniile de valori ale atributelor din setul de date sunt standardizate

- Foarte mulți algoritmi sunt sensibili la scara de valori pe care atributele o folosesc
- Atribute cu valori foarte mari pot ajunge să domine procesele de decizie, de exemplu, din problemele de clasificare sau regresie
- Scalare Min-Max (normalizare)
  - Transformarea intervalului de valori pe baza valorii minime și maxime
  - Algoritmii aplicați nu au presupuneri legate de distribuția valorilor atributelor

$$x' = (x - x_{min})/(x_{max} - x_{min})$$

- Scalare prin logaritmare
  - Utilizată când mare parte din exemple sunt concentrate pe un subdomeniu mic de valori

$$x' = log(x)$$

- Scalare prin calculul scorului z (standardizare)
  - Presupune distribuție normală (Gaussiană) a valorilor atributului

$$x' = (x - \mu)/\sigma$$

Din practică

- Definiție: Metode prin care domeniile de valori ale atributelor din setul de date sunt standardizate

- Foarte mulți algoritmi sunt sensibili la scara de valori pe care atributele o folosesc
- Atribute cu valori foarte mari pot ajunge să domine procesele de decizie, de exemplu, din problemele de clasificare sau regresie
- Scalare Min-Max (normalizare)
  - Transformarea intervalului de valori pe baza valorii minime și maxime
  - Algoritmii aplicați nu au presupuneri legate de distribuția valorilor atributelor

$$x' = (x - x_{min})/(x_{max} - x_{min})$$

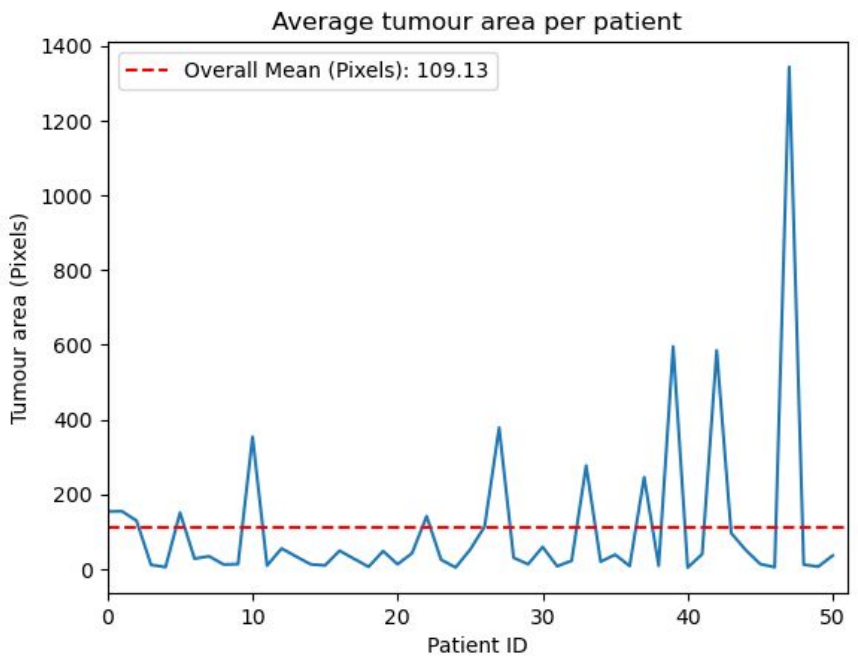
- Scalare prin logaritmare
  - Utilizată când mare parte din exemple sunt concentrate pe un subdomeniu mic de valori

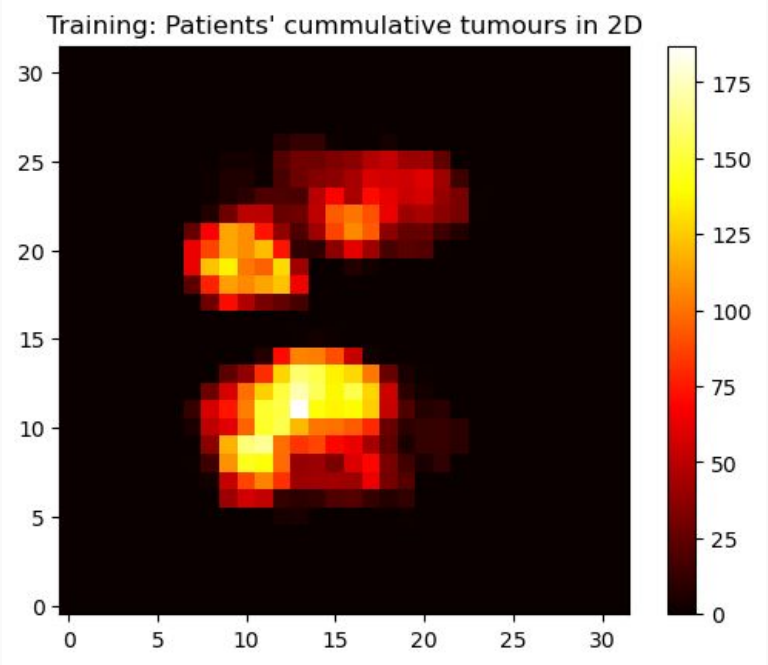
$$x' = log(x)$$

- Scalare prin calculul scorului z (standardizare)
  - Presupune distribuție normală (Gaussiană) a valorilor atributului

$$x' = (x - \mu)/\sigma$$

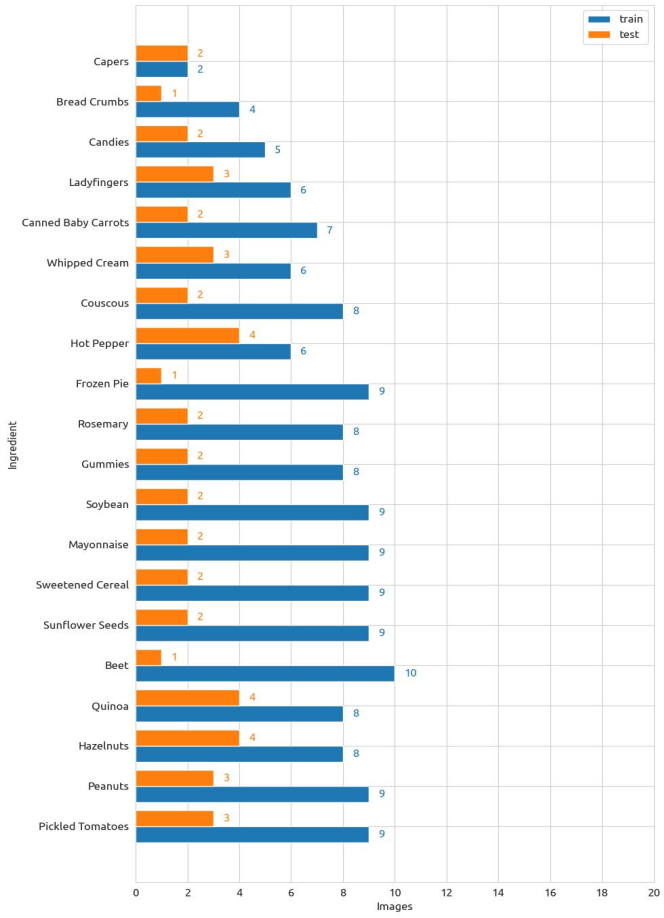
Lecții învățate din practică (I)

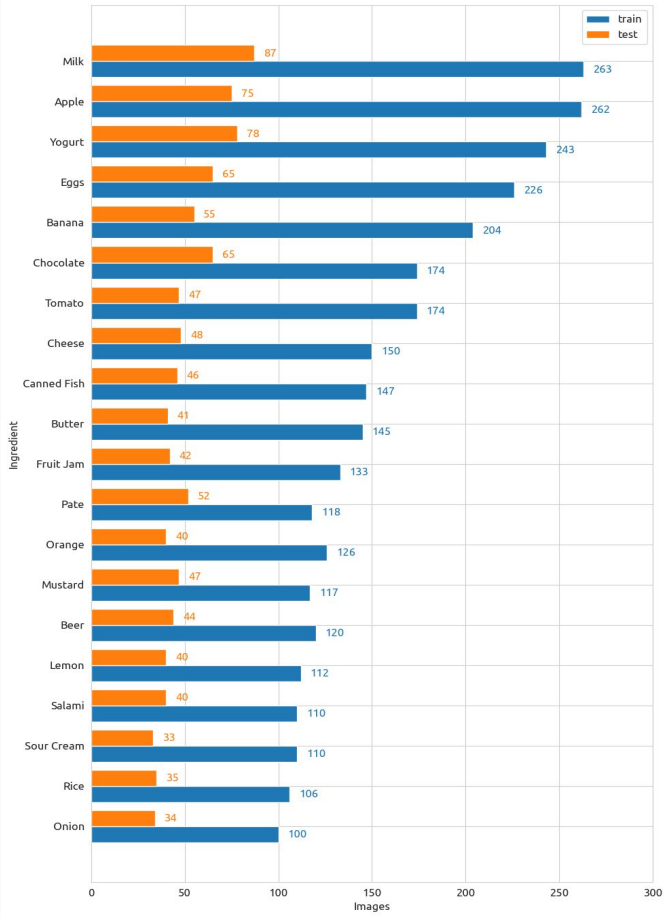




Dimensiunea și poziționarea țesutului tumoral în scan-uri CT. Tumorile sunt de dimensiuni mici și există un bias pentru centralitate ⇒ Redimensionarea imaginii și decuparea sunt constrânse

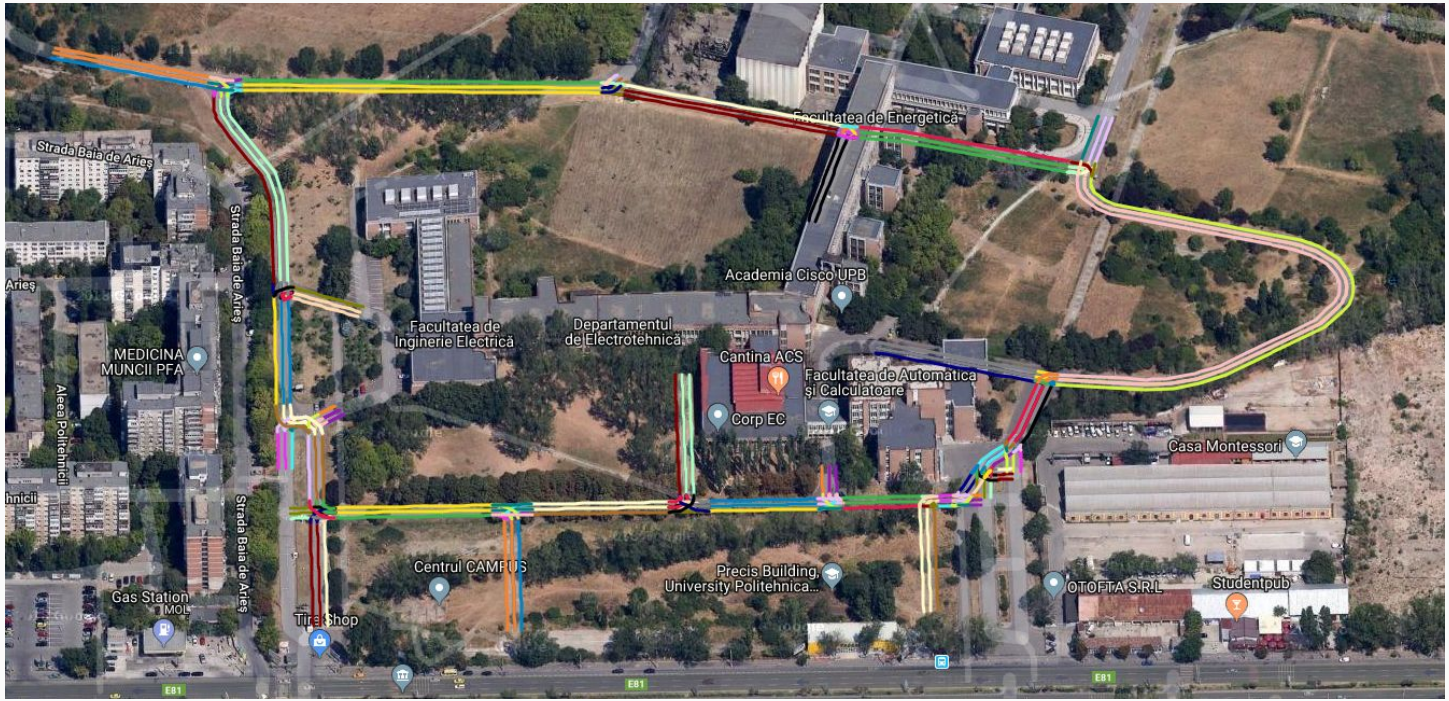
Lecții învățate din practică (II)





Distribuțiile „deformate” (în cazul acesta, de tip *long tailed*), conduc la multe erori în clasificare (segmentare de instanțe)

Lecții învățate din practică (III)



În setul de date UPB Campus, există un bias pentru curbe la dreapta. Funcția de eroare poate fi / trebuie ajustată sau introducem mecanisme care să nu penalizeze predicția virajelor.