Lab 08 — Machine Learning aplicat pe date omice

1. Introducere

Datele omice, în special datele de expresie genică, sunt caracterizate prin dimensionalitate ridicată și un număr relativ limitat de probe etichetate. Acest tip de date ridică provocări specifice pentru metodele clasice de învățare automată, dar oferă în același timp oportunități pentru identificarea unor tipare biologice relevante.

Scopul acestui laborator este construirea și evaluarea unui pipeline complet de analiză ML aplicat pe o matrice de expresie genică, incluzând:

- clasificare supravegheată,
- analiză nesupravegheată,
- un mini-experiment semi-supervised,
- și interpretarea biologică a rezultatelor.

Datele utilizate constau într-o matrice de expresie genică (1000 probe × 1000 gene), cu etichete pseudo-generate anterior, folosite ca proxy pentru clase biologice distincte.

2. Supervised Machine Learning

2.1 Random Forest

Modelul principal utilizat a fost **Random Forest**, ales datorită capacității sale de a gestiona date de mare dimensionalitate și relații non-liniare între gene.

Pipeline-ul a inclus:

- separarea setului în train/test stratificat (80% / 20%),
- antrenarea modelului Random Forest,
- evaluarea performanței prin `classification_report`,
- analiza matricei de confuzie,
- extragerea importanței genelor.

Rezultate:

- acuratețe = **1.00**
- F1-macro = **1.00**
- matricea de confuzie nu indică erori de clasificare.

Modelul separă perfect cele trei clase, sugerând o separabilitate ridicată în spațiul expresiei genice.

2.2 Logistic Regression (opțional)

Pentru comparație, a fost antrenat și un model **Logistic Regression**, cu normalizare (`StandardScaler`) și clasificare multinomială.

Rezultate:

- acuratețe ≈ **0.99**
- F1-macro ≈ **0.99**

Deși performanța este foarte ridicată, Logistic Regression rămâne ușor inferioară Random Forest, deoarece este un model liniar și nu poate captura interacțiuni complexe între gene.

Comparație RF vs LogReg:

- Random Forest modelează relații non-liniare și interacțiuni între gene.
- Logistic Regression oferă interpretabilitate mai bună, dar flexibilitate mai redusă.
- În acest set de date, relațiile non-liniare par importante.

3. Unsupervised Machine Learning: PCA + KMeans

Pentru explorarea structurii datelor fără a utiliza etichetele, a fost aplicată o analiză nesupravegheată.

PCA

Primele două componente principale explică aproximativ:

- PC1 ≈ 6.0%
- PC2 ≈ 4.2%
  din variația totală, ceea ce este tipic pentru date omice cu zgomot biologic ridicat.

KMeans

A fost ales **k = 3**, pentru a corespunde numărului de clase.

Crosstab Label × Cluster:

Label	Cluster 0	Cluster 1	Cluster 2
0	234	423	0
1	0	0	92
2	250	1	0

Observații:

- o clasă este recuperată aproape perfect într-un cluster separat,
- celelalte două clase prezintă suprapunere parțială,
- PCA reduce dimensionalitatea, dar pierde informație relevantă pentru separarea completă.

Concluzie:
Clusterizarea recuperează **parțial** structura de clase, ceea ce indică faptul că diferențele biologice sunt reale, dar distribuite pe mai multe dimensiuni decât cele capturate de primele două componente PCA.

4. Semi-Supervised Learning

Pentru a simula un scenariu realist de date parțial etichetate:

- 40% din etichetele setului de antrenare au fost eliminate (unknown),
- Random Forest a fost antrenat doar pe probele etichetate,
- s-au generat pseudo-etichete pentru probele unknown,
- modelul a fost reantrenat pe setul complet.

Rezultate:

Model	F1-macro (test)
RF doar pe date etichetate	0.995
RF cu pseudo-labels	0.995

Acuratețea pseudo-etichetelor generate a fost ≈ **98.75%**.

Interpretare:
Performanța nu crește, dar nici nu scade, deoarece modelul inițial avea deja performanță foarte ridicată. Pseudo-labeling-ul confirmă stabilitatea modelului și arată utilitatea metodei în contexte cu mai puține etichete disponibile.

5. Interpretare biologică

Analiza importanței genelor din Random Forest indică gene precum **CTSS, GPX1, S100A9, FCER1G**, care sunt cunoscute pentru roluri în:

- procese imune,
- inflamație,
- stres oxidativ,
- funcții metabolice specifice anumitor tipuri celulare.

Faptul că un număr relativ mic de gene domină clasificarea sugerează existența unor semnături biologice clare, responsabile pentru separarea claselor.

6. Limitări

- Etichetele utilizate sunt **pseudo-labels**, nu adnotări biologice validate experimental.
- Dimensionalitatea mare poate favoriza supra-separabilitatea claselor.
- PCA oferă o vedere limitată asupra structurii reale a datelor.
- Rezultatele pot varia pe seturi de date cu clase mai apropiate biologic.

Bonus — PCA

Eliminarea genelor cu varianță scăzută îmbunătățește ușor vizibilitatea clusterelor în spațiul PCA, reducând zgomotul și accentuând genele informative. Totuși, separarea completă rămâne dependentă de dimensiuni suplimentare din spațiul original.