

# Laborator 2

## Inteligență Artificială

---

Drd.Limboi Sergiu

# Agenda

---

- Introducere în *Învățarea Automată*
- Etape în cadrul procesului de *Învățare Automată*
- Preprocesarea datelor- $\rightarrow$  Normalizarea datelor
- *Discuții Tema 2*

# ÎNVĂȚAREA AUTOMATĂ

---

- Învățarea poate fi explicată ca fiind activitatea de a obține cunoștințe sau de a le înțelege, precum și abilitatea de a-ți însuși noțiuni prin studiu, instruire sau experiență
- Învățarea automată (engleză Machine Learning) se referă la modificări din sisteme care realizează sarcini asociate cu diverse teme, concepte din Inteligență Artificială, sarcini precum diagnoză, predicție, planificare, control sau detectare.

# ÎNVĂȚAREA AUTOMATĂ

---

- Scop->proiectarea și dezvoltarea unor algoritmi și metode utilizate pentru ca un sistem computațional să “învețe”
- De ce să învețe sistemele informatice?
  - Unele sarcini se pot defini doar prin exemple- de aceea trebuie furnizate perechi de intrări-ieșiri, în absența unor legături concrete între datele de intrare și rezultate
  - Este posibil ca anumite informații ascunse, nedescifrate, să reflecte corelații sau legături
  - Cantitatea mare de informații poate fi dificil de codificat de către mintea umană

# ÎNVĂȚAREA AUTOMATĂ

---

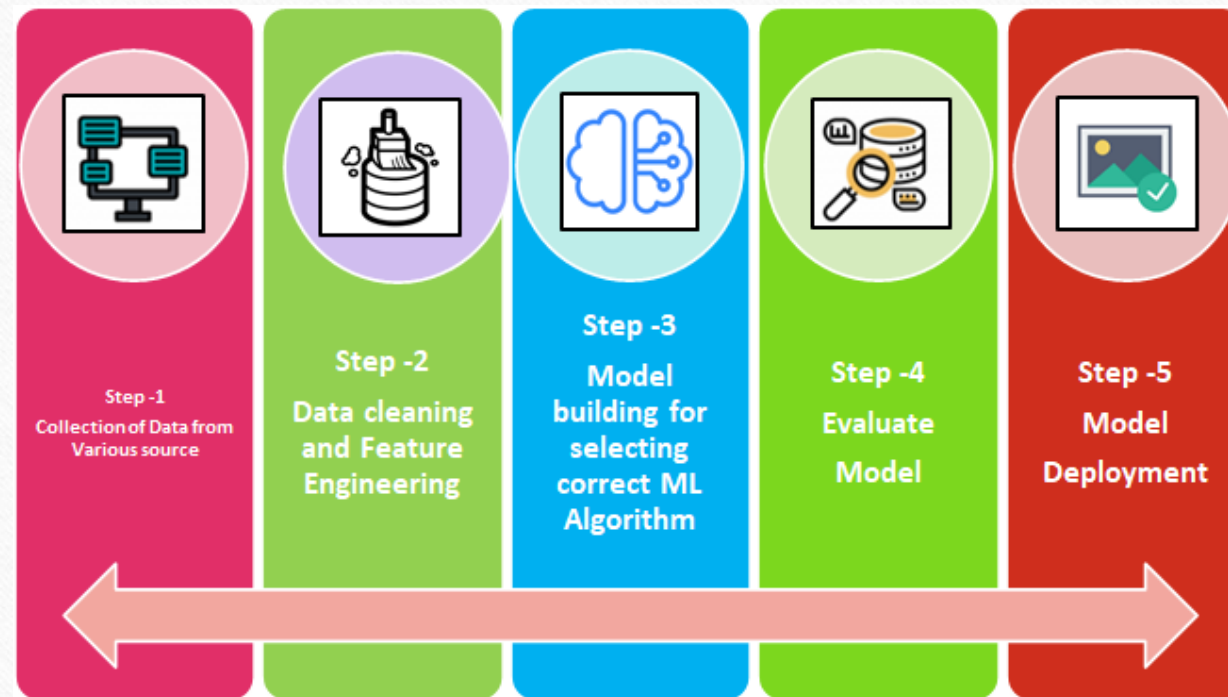
- Pentru a defini conceptele utilizate în problematica învățării se folosește o funcție  $f$ , iar sarcina celui care învață este de a ghici forma funcției.
- Tipuri de învățare
  - Supervizată- știm (uneori doar aproximativ) valorile funcției  $f$  pentru  $m$  cazuri ale unui set de antrenare. Sistemul poate intui, după etapa de antrenare, care ar fi funcția pentru un alt set de date.
  - Nesupervizată-avem doar o mulțime de date pentru care nu știm funcția. Presupune divizarea setului de date în submulțimi sau grupuri. În acest caz valoarea funcției va fi numele subgrupului (clasei) din care setul face parte.
  - Prin întărire (reinforcement learning)- sistemul interacționează cu mediul și poate primi recompense sau penalizări.

# Etape în cadrul unui proces de învățare automată

---

- Colectarea datelor
- Procesarea/Explorarea datelor
- Stabilirea unui model
- Etapa de antrenare (dacă vorbim de învățare supervizată)
- Evaluarea rezultatelor
- Etapa de predicție/clasificare (aplicarea modelului pe un nou set de date)

# Etape în cadrul unui proces de învățare automată



# Preprocesarea/Explorarea datelor

---

- În cadrul acestei etape se analizează datele colectate.
- De exemplu- dacă vorbim de texte este necesară o preprocesare a lor (eliminarea caracterelor speciale, punctuație, repetiție, cuvinte colocviale, etc.).  
Mai multe detalii în laboratoarele următoare
- O sub-etapă importantă este normalizarea datelor (dacă vorbim de valori numerice). Normalizarea implică ajustarea informației la aceeași scală (același numitor comun).
- Să nu comparăm “mere cu pere”



# Normalizarea datelor

---

- De ex. Avem un set de date care conține informații despre temperaturi. O valoare poate varia (ex. -100, 5, 50, 30, 60 grade Celsius).
- Trebuie să aducem valorile la un numitor comun având în vedere paleta largă de valori
- Există diferite tehnici pentru a normaliza datele
  - Clipping
  - Scalare logaritmică
  - Normalizare de tip min-max
  - Standardizare (normalizare Z) -temă

# Metoda de tip Clipping

---

- Definirea unui prag pentru a segmenta datele
- Recomandare de utilizare- atunci când setul de date conține mulți outliers (o instanță a setului este foarte diferită de restul din set- de ex. majoritatea valorilor sunt în intervalul [1,50], iar outliers pot fi 200, 300, 150)

$$x_i^{new} = \begin{cases} x_i, & x_i < threshold \\ threshold, & x_i \geq threshold \end{cases}$$

# Scalare logaritmică

---

- A se utiliza atunci când subsetul/ instanța are mai multă informație sau mai puțină decât restul setului de date.
- “Generally, you use logarithmic scales if the data in your graph represents a large range, such as an exponential growth rate.”
- O scalare logaritmică poate proiecta valori de la 10 la 100.000 de ex
- Ex. multe filme sunt evaluate cu note puține (ratings), iar unele au foarte multe note.

$$x_i^{new} = \log(x_i)$$

# Scalare de tip min-max

---

- Conversia valorilor reale din intervalul lor natural într-unul comun (de obicei  $[0,1]$ )
- De ex. –dacă avem măsurători ale unor rămășițe arheologice umane putem avea valori între 5 și 100. Aceste valori ar oaselor le vom aduce în intervalul  $[0,1]$ .
- A se utiliza când cunoaștem valorile superioare și inferioare ale setului și avem puțini spre deloc outliers. De asemenea, setul de date este aproximativ uniform distribuit în cadrul intervalului de valori.

$$x_i^{new} = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

# Tema 2

---

- Pe baza diferitelor seturi de date (imagini/texte/informații numerice) se dorește prelucrarea informației
  - Valori minime, maxime, medii
  - Vizualizarea datelor/distribuții
  - Normalizarea datelor
  - Etc.
- <https://github.com/lauradiosan/AI-UBB/blob/main/2023-2024/labs/lab02/AI-lab02.md>