

**UNIVERSITATEA TITU MAIORESCU**

**FACULTATEA DE INFORMATICĂ**


**MASTER ȘTIINȚA DATELOR ȘI INTELIGENȚĂ ARTIFICIALĂ  
FUNDAMENTELE ȘTIINȚEI DATELOR**

**CURSUL 2  
FUNDAMENTELE MATEMATICE ȘI STATISTICE  
ALE DATA SCIENCE**


**Autor:  
Prof.univ.dr.ing.Titi PARASCHIV**

**București  
2024**


# CUPRINS

- ▶ 1. Matematica
  - ▶ 2. Algebra liniară
  - ▶ 3. Statistica
  - ▶ 4. Calcul diferențial și integral
  - ▶ 5. Algoritmi de optimizare
  - ▶ 6. Teoria grafurilor
  - ▶ 7. Teoria generală a sistemelor
  - ▶ 8. Statistica pentru Data Science
  - ▶ 9. Modelare statistică
  - ▶ 10. Tehnici de optimizare
  - ▶ 11. Tehnici de învățare automată
  - ▶ 12. Concluzii
- 
- Several white lines of varying lengths and orientations are drawn in the bottom right corner of the slide, creating a modern, abstract graphic element.


# MATEMATICA ÎN DATA SCIENCE

- ▶ **Cunoștințe esențiale:**  
Algebra liniară, statistică, calcul diferențial și integral, algoritmi de optimizare, teoria grafurilor, teoria sistemelor.
  - ▶ **Importanță:**  
Fundamentele matematice sunt esențiale pentru înțelegerea și aplicarea tehnicilor de Data Science.
- 
- Several white lines of varying lengths and angles are drawn in the bottom right corner of the slide, creating a modern, abstract graphic element.


# ALGEBRA LINIARĂ

- ▶ • Vectori: Definiție, adunare, produs scalar și vectorial.
  - ▶ • Matrici: Definiție, operații, proprietăți ale matricilor.
  - ▶ • Ecuații liniare: Soluții prin metode de substituție, eliminare și utilizarea matricilor.
- 
- A series of white diagonal lines of varying lengths and thicknesses, located in the bottom right corner of the slide.


# STATISTICA

- ▶ • Definiție: Colectarea, organizarea, analiza și interpretarea datelor.
  - ▶ • Indicatori: Tendința centrală (media, mediana, modulul), dispersia și variabilitatea datelor.
  - ▶ • Analiză: Corelație, regresie, probabilitate, inferență statistică.
- 
- A series of white diagonal lines of varying lengths and thicknesses, located in the bottom right corner of the slide, creating a modern, abstract graphic element.


# CALCULUL DIFERENȚIAL ȘI INTEGRAL

- ▶ • Diferențiere: Calcularea ratei de schimbare a unei funcții.
  - ▶ • Integrare: Calculul ariei sub o curbă și determinarea primitivelor unei funcții.
  - ▶ • Aplicații: Fizică, economie, inginerie, științe sociale.
- 
- A series of several parallel white lines of varying lengths and slopes, located in the bottom right corner of the slide, creating a modern, abstract graphic element.

# ALGORITMI DE OPTIMIZARE


- ▶ • Tipuri de algoritmi: Gradient descent, căutare pe linie, algoritmi genetici, căutare bazată pe judecată.
  - ▶ • Aplicabilitate: Probleme de optimizare în inginerie, economie, Data Science.
- 
- Several white lines of varying lengths and angles are drawn in the bottom right corner of the slide, creating a modern, abstract graphic element.

# TEORIA GRAFURILOR


- ▶ • Concepte: Noduri, muchii, grafuri orientate și neorientate, arbori.
  - ▶ • Aplicabilitate: Rețele de comunicații, optimizarea traseelor, științe sociale.
- 
- A series of white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.



# TEORIA SISTEMELOR

- ▶ • Definiție: Studiul sistemelor complexe și interacțiunilor dintre componentele acestora.
  - ▶ • Concepte: Feedback, emergență, modelare și simulare.
  - ▶ • Aplicații: Biologie, economie, inginerie, sociologie.
- 
- A series of white lines of varying lengths and orientations are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

# STATISTICA ÎN DATA SCIENCE

- ▶ • Particularități: Volumul și varietatea datelor, integrarea tehnologiei, abordarea interdisciplinară.
  - ▶ • Tehnici avansate: Rețele neuronale, învățare supervizată și nesupervizată, învățare prin transfer.
- 
- A series of white diagonal lines of varying lengths and thicknesses, located in the bottom right corner of the slide, creating a modern, abstract graphic element.

# MODELARE STATISTICĂ ÎN DS

Construirea unui Model:

Modelarea începe prin definirea unui model statistic care descrie relația dintre VI (predictori) și VD (de interes).

Estimarea Parametrilor:

Următorul pas este estimarea parametrilor care controlează relațiile dintre variabile.

Evaluarea Modelului:

Se face prin metrici de evaluare, cum ar fi eroarea pătratică medie (MSE), coeficientul de determinare ( $R^2$ ), precizia,

Validarea și Îmbunătățirea Modelului:

Se face folosind un set de date independent pentru a testa capacitatea sa de generalizare.

Aplicarea Modelului:

Modelele statistice dezvoltate sunt apoi utilizate pentru a face predicții asupra datelor noi.

Tipuri de Modele Statistice în DS:

Regresia liniară;

Regresia logistică;

Arborele de decizie;

Clustering;

Modele de rețele neuronale.

# TEHNICI DE OPTIMIZARE ÎN DS

**Gradient Descent: (Căderea Gradientului)** este una dintre cele mai populare tehnici de optimizare utilizate în machine learning și deep learning.

**Conjugate Gradient** este o metodă folosită pentru optimizarea funcțiilor neconstrânse de dimensiuni mari, fiind o tehnică de compromis între exactitate și viteză.

**Simulated Annealing** este o metodă probabilistică de optimizare care imită procesul de răcire a materialelor în metalurgie.

**Algoritmi Evolutivi si Genetici:**

**Algoritmii Genetici** sunt metode de optimizare inspirate din procesele evoluției biologice, cum ar fi selecția naturală, încrucișarea și mutația.

**Algoritmi Evolutivi** extind ideile algoritmilor genetici și includ tehnici precum evoluția diferențială și algoritmii de enjambament (crossover).

**Optimizarea Neliniară:**

**Programarea Liniară și Neliniară:**

**Metoda Monte Carlo:**

**Optimizarea funcțiilor de pierdere în învățarea automată;**

**Optimizarea hiperparametrilor;**

**Optimizarea proceselor.**

# TEHNICI DE ÎNVĂȚARE AUTOMATĂ

## 1. Învățare Supervizată (Supervised Learning)

**Regresie Liniară:** Folosită pentru a prezice valori numerice continue.

**Regresie Logistică:** Utilizată pentru clasificarea binară.

**Arbori de Decizie (Decision Trees):** Modele care fac predicții prin învățarea unor reguli decizionale simple.

**Random Forest:** Un ansamblu de arbori de decizie pentru a îmbunătăți acuratețea și generalizarea.

**Support Vector Machines (SVM):** Utilizate pentru clasificarea liniară și neliniară.

**Rețele Neuronale:** Modele inspirate de structura creierului uman.

## 2. Învățare Nesupervizată (Unsupervised Learning)

**Clustering (Clusterizare):** Împărțirea datelor în grupuri bazate pe similaritate.

**Analiza Componentelor Principale (PCA):**

**Asociații (Association Rules):**

## 3. Învățare Semi-Supervizată (Semi-Supervised Learning)

**Self-Training:** Modelul se antrenează inițial pe datele etichetate, apoi folosește propriile predicții pentru a eticheta datele neetichetate.

**Co-Training:** Folosirea a două modele antrenate pe subseturi diferite de date.


## 4. Învățare prin Recompensă (Reinforcement Learning)

**Q-Learning:** Algoritm pentru învățarea unei politici care maximizează recompensa.

**Online SVM:** Adaptarea suport vector machines pentru actualizare continuă.

## 5. Învățare Profundă (Deep Learning)

# CONCLUZIE

- ▶ • Importanța Matematicii și Statisticii: Bazele teoretice pentru dezvoltarea și aplicarea tehnicilor de Data Science.
  - ▶ • Evoluția: Continuu în contextul noilor tehnologii și metode de analiză a datelor.
- 
- Several white lines of varying lengths and orientations are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.