Machine Learning

Partea I. Introducere în Machine Learning

Ce ne așteaptă?

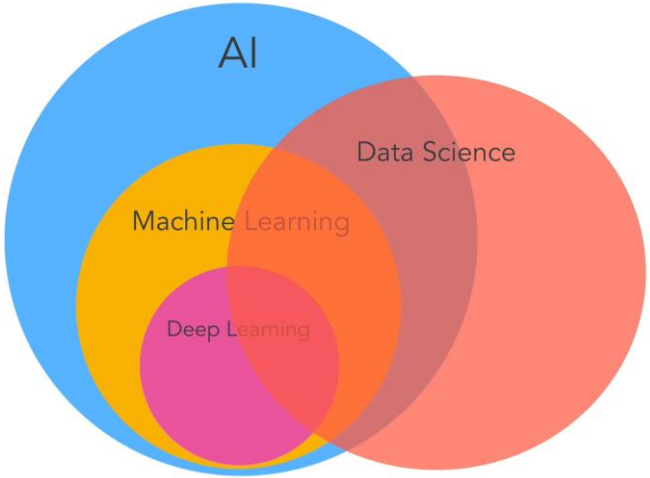
1. Ce este Machine Learning
2. Machine Learning și Data Science
3. Tipuri de Machine Learning
4. Machine Learning supravegheat
5. Machine Learning nesupravegheat

1. Ce este Machine Learning

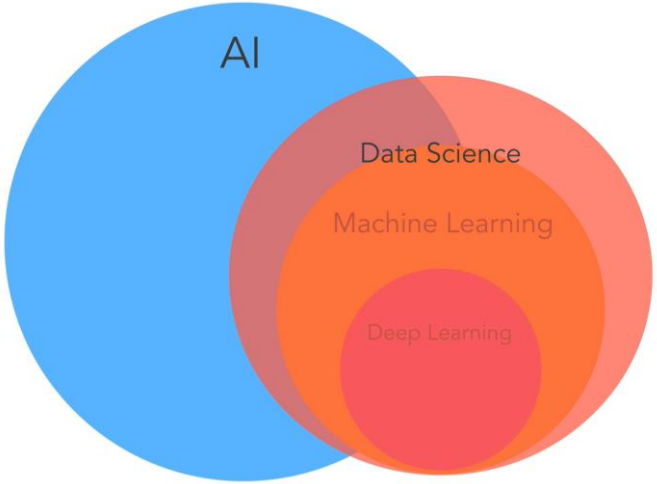
Ce reprezintă Machine Learning

- **Machine Learning (ML) – o componentă a Inteligenței Artificiale (AI) care utilizează diverși algoritm pentru crearea unui model pe baza unor date și utilizarea ulterioară a acestui model pentru realizarea predicțiilor de viitor pe date similare.**
- **Inteligența artificială (Artificial Inteligence - AI) reprezintă o tehnologie de crearea inteligînței pentru mașini ce imită inteligența umană.**
- **ML poate fi interpretat și ca o componentă a Data Science.**
- **În funcție de sarcina ce trebuie soluționată, ML se bazează pe diferiți algoritmi iar o sub-ramură a ML ce utilizează rețelele neuronale se numește Deep Learning (DL)**

AI vs ML vs DL vs Data Science

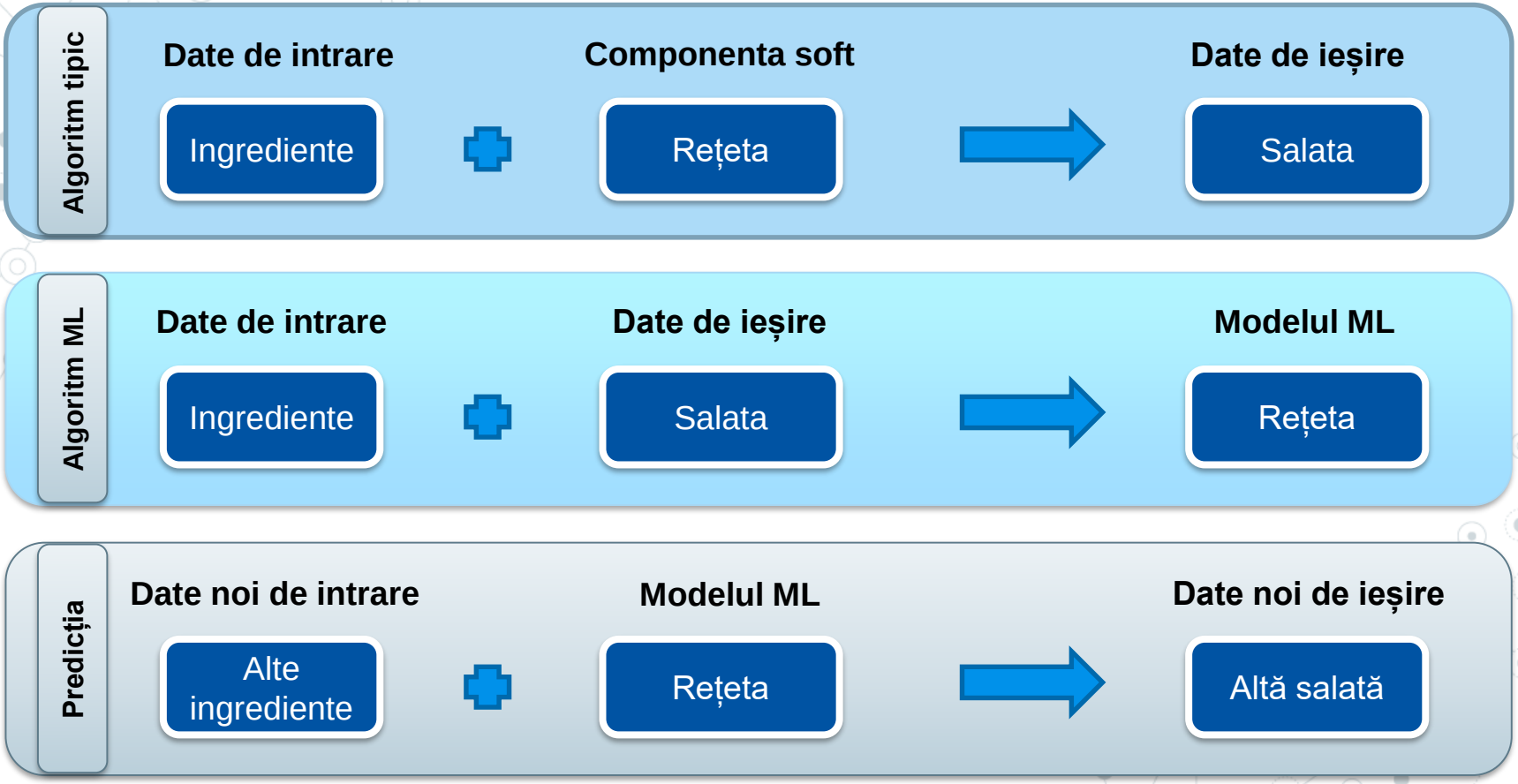


ML ca componentă AI



ML ca componentă Data Science

Algoritm tipic vs algoritm ML



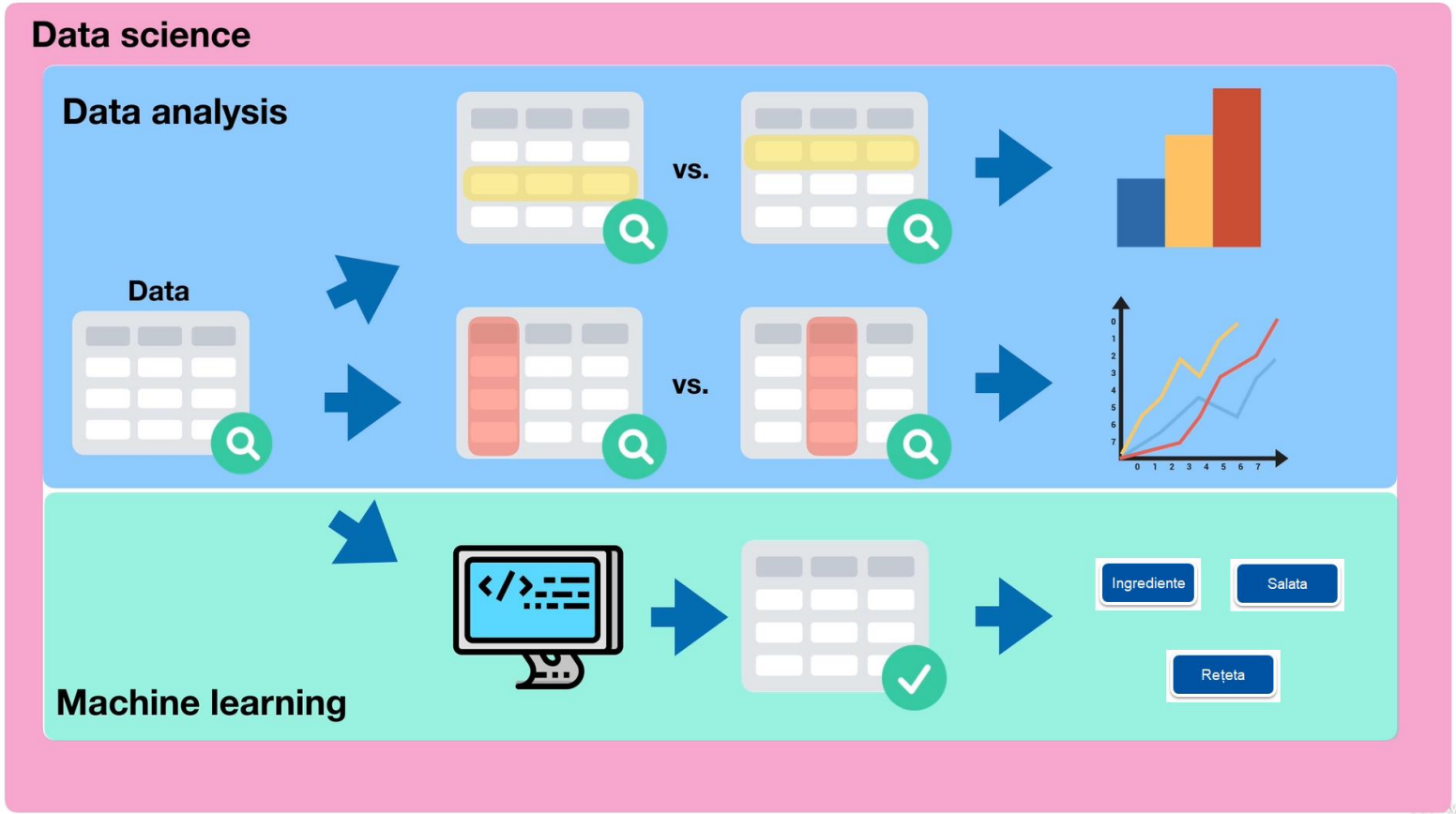
Algoritmi ML

- ML în general poate fi considerat drept un algoritm statistic de calcul a cărui performanțe se îmbunătățesc în mod automat pe baza datelor
- Spre deosebirea de algoritmii tipici care se bazează pe date de intrarea ale omului pentru a adopta o careva abordare, algoritmul ML desinestătător deduce cea mai bună abordare pe baza datelor istorice.
- Algoritmii ML nu sunt programați în mod explicit cu privire la ce decizie să ia ci sunt proiectați să deducă cea mai optimă abordare pe baza datelor
- De exemplu, algoritmului ML de detecție a mesajelor spam nu îi este specificat de către om caracteristicile unui mesaj spam ci el singur pe baza mai multor mesaje care au fost clasificate ca spam sau nu, determină aceste caracteristici.

Modelul ML

- Algoritmii ML detectează automat caracteristicile datelor și importanța acestora în cadrul datelor existente
- Multe dintre caracteristicile datelor detectate de algoritmii ML nu sunt perceptibili pentru om, în special în cadrul datelor nestructurate
- Pe baza caracteristicilor detectate și a ieșirilor datelor istorice algoritmul ML generează legea dintre aceasta care este "împachetată" într-un model numit și estimator
- Un model efectiv se bazează pe un algoritm corespunzător problemei și pe un set cât mai numeros de date cât mai reprezentative

Machine Learning vs Data Science



2. Machine Learning și Data Science

Datele și lumea reală

- **În lumea reală datele se utilizează pentru a soluționa probleme sau pentru a răspunde la întrebări**

Lumea reală

Probleme de soluționat

Cum modific sau repar X?

Răspuns la întrebări

Cum modificările lui X vor afecta Y?

- **Pentru aceasta, pe baza datelor din lumea reală se creează Produse de Date (Data Product) și se realizează Analiza Datelor (Data Analysis)**

Lumea reală

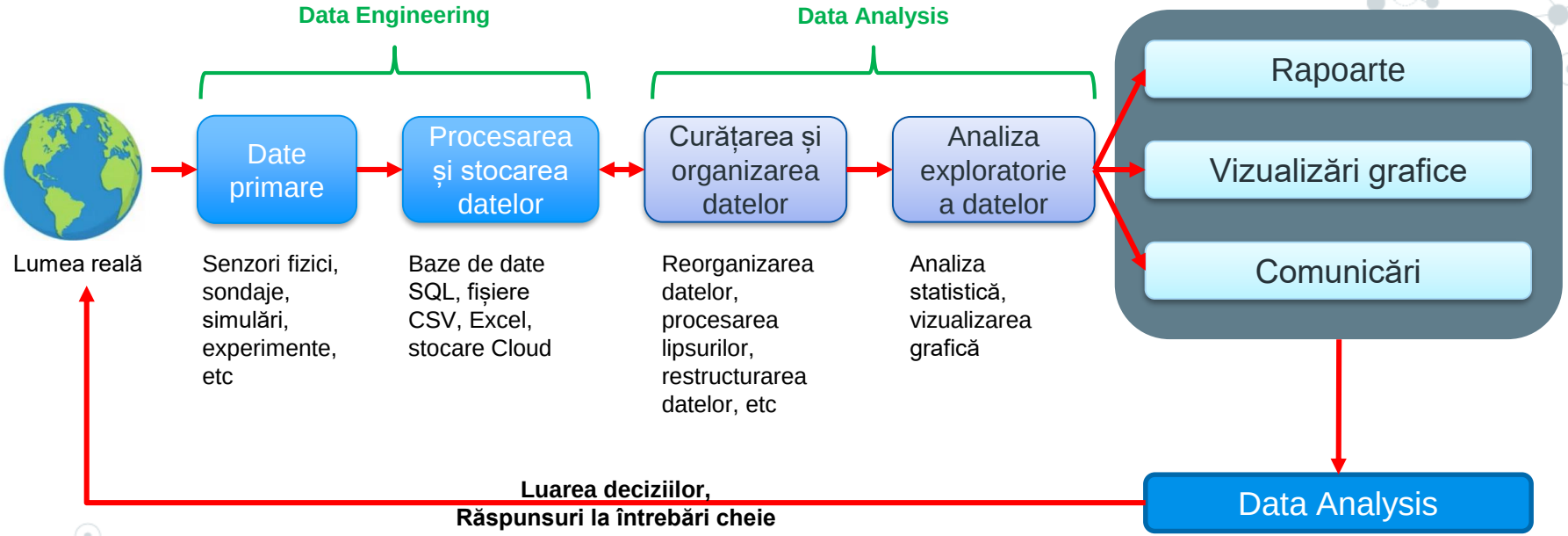
Data Product

Aplicații mobile, Servicii Web ,etc

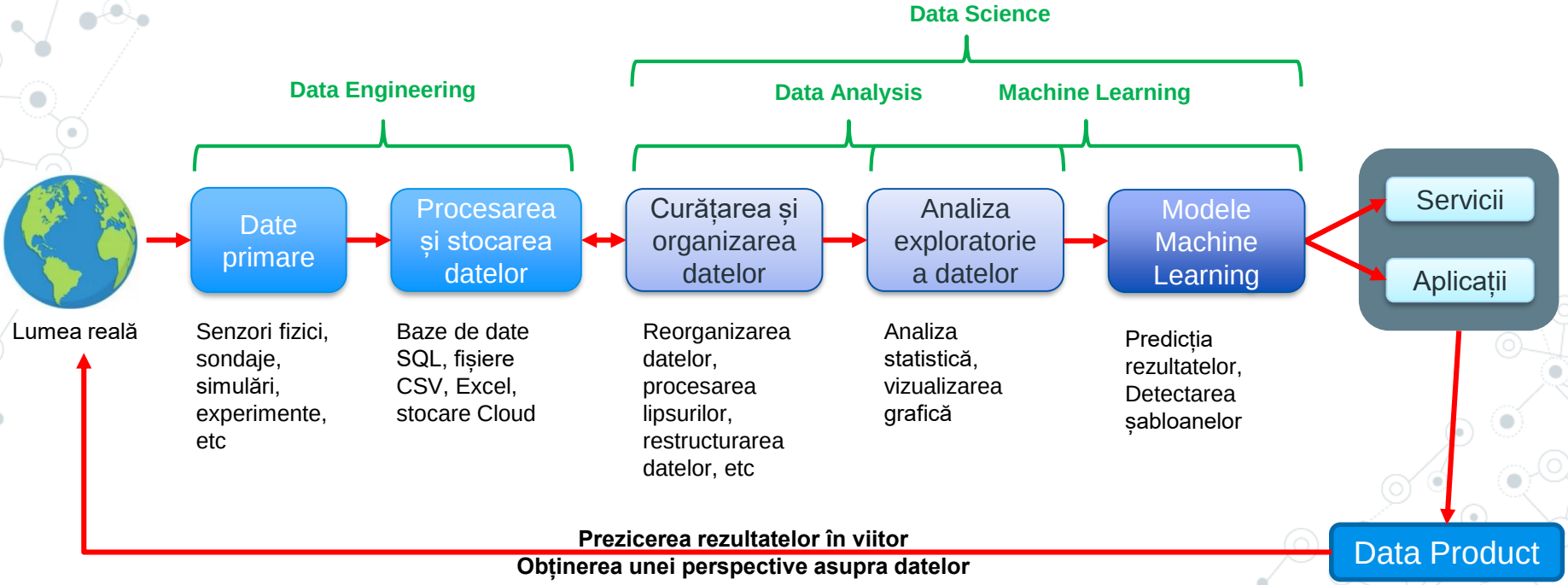
Data Analysis

Rapoarte, grafice, comunicări, etc

Etapele Data Analysis



Etapele Machine Learning



3. Tipuri de Machine Learning

Clasificarea ML

- În funcție de conținutul datele istorice, ML se clasifică în două mari categorii

- ML supravecheat utilizează date istorice cu rezultate (labeled data) și permite elaborarea modelelor ce vor fi utilizate în viitor pentru predicția rezultatelor altor date similare

- ML nesupravecheat utilizează date istorice fără rezultate (unlabeled data) și permite elaborarea modelelor ce vor fi asigura detectarea unor șabloane în date similare

ML supravegheat

- **ML supravegheat în funcție de tipul datelor rezultatelor asigură soluționarea a 2 tipuri de sarcini:**
- **Sarcini de clasificare – rezultatele datelor istorice sunt de tip categoriale și se elaborează modele ce ar prezice categoria noilor rezultate**
- **Exemplu sarcini de clasificare cu clase binară:**
  Predicția îmbolnăvirii unui pacient
  Predicția unui spam
- **Exemplu sarcini de clasificare cu clase multiple:**
  Predicția rasei unui câine
  Predicția literei scrisă de mână
- **Sarcini de regresie – rezultatele datelor istorice sunt de tip numeric și se elaborează modele ce ar prezice valoarea numerică a noilor rezultate**
- **Exemple**

  Predicția prețului unei case
  Predicția punctajului unui test

ML nesupravegheat

- **Asigură soluționarea sarcinilor legate de gruparea și interpretarea datelor fără rezultate**

- **Exemplu**

  Gruparea clienților unui magazin în funcție de produsele cumpărate în ultima perioadă cu scopul de a le transmite mesaje informative corespunzătoare

- **Unul dintre neajunsurile algoritmilor ML nesupravegheat este lipsa mecanismelor de evaluarea a performatelor acestora din cauza lipsei “ răspunsurilor corecte” a datelor istorice**

4. Machine Learning supravegheat

Colectarea datelor

- **Concretizarea sarcinii și particularitățile produsului rezultat**
- **Colectarea și organizarea setului de date istorice, de obicei, cea mai de durată în cadrul ML**
- **Exemplu**

  Sarcina constă în elaborarea unui model de predicție a prețului caselor într-o anumită zonă. Un set organizat de date istorice referitoare la caracteristicile și prețul caselor de locuit în acea zonă este prezentat în tabel

Suprafața (m^2)	Nr. dormitoare	Nr. garaje	Prețul (€)
200	3	2	250 000
190	2	1	225 000
230	3	3	325 000
180	1	1	200 000
210	2	2	275 000

Setul de date X și setul y

- **Divizarea setul de date istorice în setul de date X și setul de date y**
- **Setul de date X reprezintă caracteristicile datelor istorice sau datele istorice de intrarea**
- **Setul de date y reprezintă etichetele datelor istorice sau datele istorice de ieșire**
- **După elaborarea modelului ML, pe baza caracteristicilor similare a datelor noi se va prezice eticheta acestora:**

Setul X			Setul y
Suprafața (m^2)	**Nr. dormitoare**	**Nr. garaje**	**Prețul (€)**
200	3	2	250 000
190	2	1	225 000
230	3	3	325 000
180	1	1	200 000
210	2	2	275 000

Datele de training și datele de test

- **Divizarea setului de date în date de training (70-80%) și date de test (20-30%)**
- **Datele de training vor permite crearea modelul ML**
- **Datele de test vor permite evaluarea performanțelor modelului ML**

	Suprafața (m^2)	Nr. dormitoare	Nr. garaje	Prețul (€)
Datele de training	200	3	2	250 000
	190	2	1	225 000
	230	3	3	325 000
Datele de test	180	1	1	200 000
	210	2	2	275 000

Subseturi ale datelor

- În rezultatul divizării datelor în subseturi se obțin:
  - Subsetul X de training (X_{train}) – conține caracteristicile datelor ce se vor utiliza la crearea modelului
  - Subsetul y de training (y_{train}) – conține etichetele datelor ce se vor utiliza la crearea modelului
  - Subsetul X de test (X_{test}) – conține caracteristicile datelor ce se vor utiliza la evaluarea modelului
  - Subsetul y de training (y_{test}) – conține etichetele datelor ce se vor utiliza la evaluarea modelului

	Suprafața (m^2)	Nr. dormitoare	Nr. garaje	Prețul (€)	
X_{train}	200	3	2	250 000	y_{train}
	190	2	1	225 000	
	230	3	3	325 000	
X_{test}	180	1	1	200 000	y_{test}
	210	2	2	275 000	

Crearea și trainingul modelului

- **Selectarea algoritmului statistic - în funcție de sarcina propusă se selectează un algoritm de soluționare**
- **Crearea modelului – odată selectat algoritmul se creează un model ML "ne antrenat"**
- **Trainingul modelului ML – învățarea modelului pe baza datelor de training (X_{train} și y_{train})**
- **Învățarea modelului presupune determinarea legității dintre caracteristicile datelor (X_{train}) și etichetele corespunzătoare lor (y_{train})**

Suprafața (m^2)	Nr. dormitoare	Nr. garaje	Prețul (€)
200	3	2	250 000
190	2	1	225 000
230	3	3	325 000

X_{train} y_{train}

Evaluarea modelului

- **Realizarea predicției – modelul creat și antrenat este utilizat pentru realizarea predicției pe datele X_{test} creându-se datele de predicție y_{pred}**
- **Evaluarea modelului – prin diverse mecanisme matematice se determină gradul de similitudine dintre datele etichetă de test (y_{test}) considerate și valori adevărate și datele de predicție (y_{pred})**

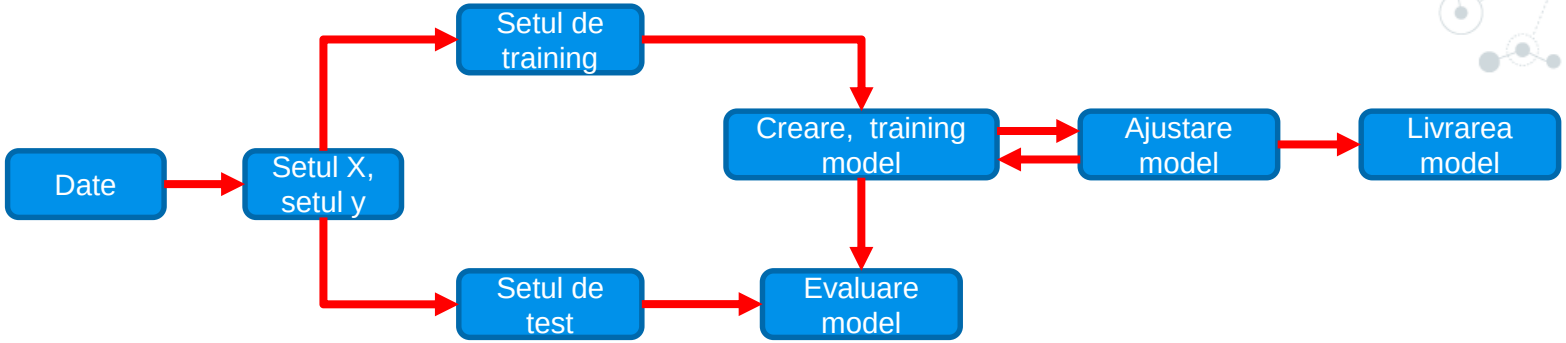
Suprafața (m^2)	Nr. dormitoare	Nr. garaje	Prețul adevărat (€)	Prețul prezis (€)
180	1	1	200 000	210 000
210	2	2	275 000	260 000
	X_{test}		y_{test}	y_{pred}

Ajustarea modelului

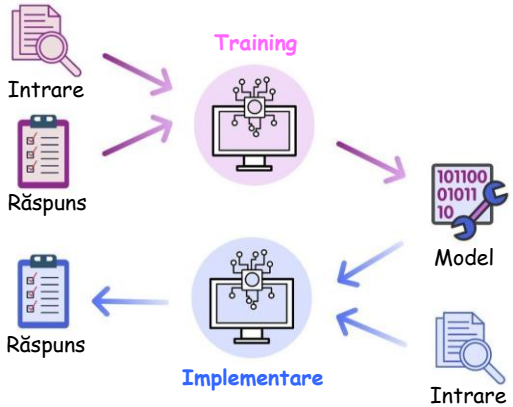
- Ajustarea modelului - modificarea valorilor hiper-parametrilor algoritmului în cazul performanțelor nesatisfăcătoare a modelului
- Hiper-parametrii sunt niște parametri ai algoritmului ce pot fi modificați de către utilizator
- Numărul și tipul hiper-parametrilor depind de algoritmul utilizat
- După ajustarea modelului se repetă procedurile de creare, training și evaluarea a modelului până se ajunge la performanțele dorite
- Dacă performanțele dorite nu pot fi atinse cu acest algoritm se purcede la îmbunătățirea datelor și/sau selectarea altui algoritm

Totalizarea procedurilor

- **Elaborarea modelului**



- **Utilizarea modelului**



5. Machine Learning nesupravegheat

Structura datelor

- ML supravegheat reprezintă o mixtiune dintre matematică și artă, ML nesupravegheat tinde a fi mai mult artă deoarece nu va fi complet clar dacă răspunsul obținut este corect sau nu

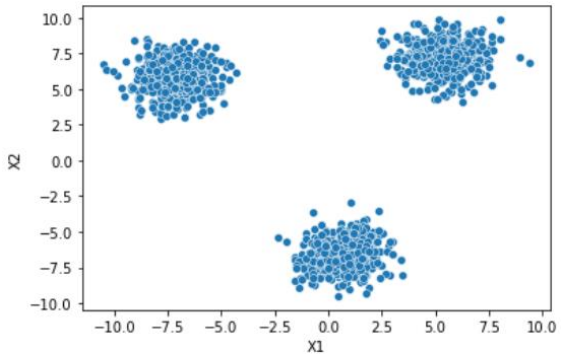
- ML nesupravegheat utilizează date istorice care nu conțin datele de ieșire, adică nu conțin etichetele datelor, deci nu poate fi definit setul y și toate datele se vor considera set X

- În lipsa etichetelor datelor nu este posibilă evaluarea modelului și deci nu are sens crearea setului datelor de test, deci toate datele vor fi considerate date de training.

Vizualizarea datelor

- **ML nesupravegheat utilizează doar setul X de date, deci poate opera doar cu caracteristicile datelor și toate sarcinile se concentrează pe analiza acestor caracteristici.**
- **Concentrarea sarcinilor pe analiza caracteristicilor datelor necesită și o mai bună cunoaștere a acestora de către utilizator inclusiv prin intermediul instrumentelor de vizualizare**
- **Exemplu de vizualizarea a datelor**

	X1	X2
0	4.645333	6.822294
1	4.784032	6.422883
2	-5.851786	5.774331
...	...	...
1497	3.180138	6.608660
1498	5.454552	6.461246
1499	-7.769230	7.014384

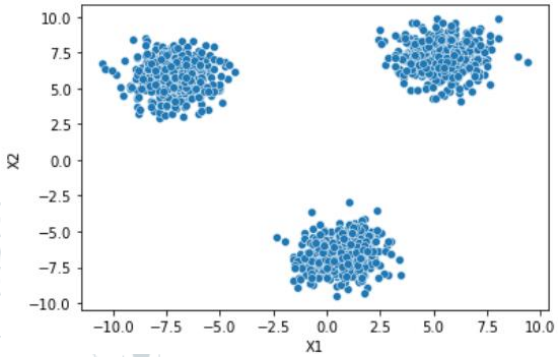


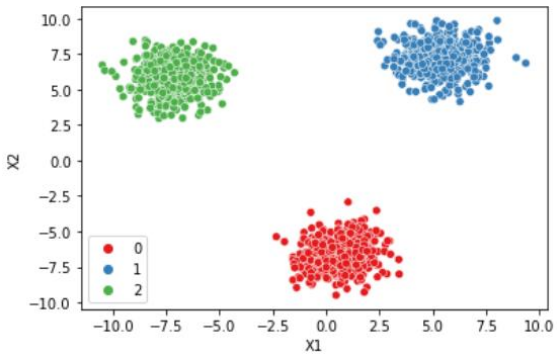
Sarcini ML nesupravegheat

- ML nesupravegheat se utilizează pentru descoperirea șabloanelor și grupurilor în date dar și caracteristicilor cele mai semnificative
- Sarcinile ML nesupravegheat pot fi:
  - Sarcini de clusterizare (clustering) – utilizând caracteristicile grupează liniile datelor în clustere distincte
  - Sarcini de reducere caracteristicilor (dimensionality reduction) – utilizând caracteristicile determină cum acestea interacționează și le reduce în mai puține componente

Clusterizarea

- Clusterizarea permite detectarea unor șabloane în cadrul caracteristicilor datelor și formarea unor clustere pe baza acestor șabloane
- Într-o anumită măsură procedura de clusterizare poate fi considerată ca procedura de atribuirea a unor etichete mai puțin confirmate pentru datele istorice
- Fiecare cluster poate fi considerat ca un tip de etichetă a datelor istorice și respectiv aceste noi date pot fi aplicate în ML supravegheat
- Exemplu grafic de clusterizare





Dimensionality Reduction

- Dimensionality Reduction permite descoperirea corelațiilor între caracteristicile datelor și crearea unor noi caracteristici prin combinarea celor primare
- Noile caracteristici includ și informația referitoare la semnificația lor în setul de date și deci cele mai puțin semnificative pot fi excluse
- Dimensionality Reduction nu reprezintă o simplă reducere a caracteristicilor primare existente și a componentelor obținute în procesarea acestor caracteristici