UNIVERSITATEA TITU MAIORESCU

FACULTATEA DE INFORMATICĂ

MASTER ȘTIINȚA DATELOR ȘI INTELIGENȚĂ ARTIFICIALĂ
FUNDAMENTELE ȘTIINȚEI DATELOR

CURSUL 1
INTRODUCERE ÎN DATA SCIENCE

Motto
Schimbarea este preţul plătit progresului după cum îndoiala este preţul plătit înţelepciunii.

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

Bucureşti
2024

CUPRINSUL CURSULUI

1. Introducere în Știința Datelor (DS): Definirea conceptului și a modelelor de DS;

2.Tipuri de sarcini în DS:

Clasificare: Predicția apartenenței unei valori la una dintre clasele predefinite folosind seturi de date cunoscute. Algoritmi: Arbori de decizie, Rețele neuronale, Modele bayesiene, Reguli de inducție, k-Nearest Neighbors;

Regresie: Predicția unei etichete numerice a unui punct de date. Algoritmi: Regresie liniară, Regresie logistică;

Detecția anomaliilor: Identificarea valorilor care deviază semnificativ de la restul setului de date. Algoritmi: Metode bazate pe distanță, Metode bazate pe densitate

Serii temporale: Predicția valorilor variabilei țintă pentru un interval de timp pe baza valorilor istorice. Algoritmi: Netezirea exponențială, ARIMA;

Clustering: Identificarea grupurilor într-un set de date. Algoritmi: k-Means, DBSCAN;

Analiza Asocierilor: Identificarea relațiilor dintre elementele unui set de date. Algoritmi: FP-Growth, Apriori;

3. Structura cursului:

Procesul DS: Prezentarea procesului standard folosit în proiectele de DS.

Explorarea Datelor: Analiza inițială a datelor pentru identificarea caracteristicilor esențiale.

Evaluarea modelelor: Metode de evaluare a performanței modelelor predictive.

Algoritmi: Detalii despre algoritmii principali utilizați: Arbori de decizie, Reguli de inducție, k-Nearest Neighbors, Naive Bayesian, RNA, Support Vector Machines, și metode Ensemble.

Regresie: Regresie liniară și logistică.

Analiza Asocierilor: Tehnici precum Apriori și FP-Growth.

Clustering: Algoritmi precum k-Means și DBSCAN.

Mining Text și Previzionarea Seriilor Temporale.

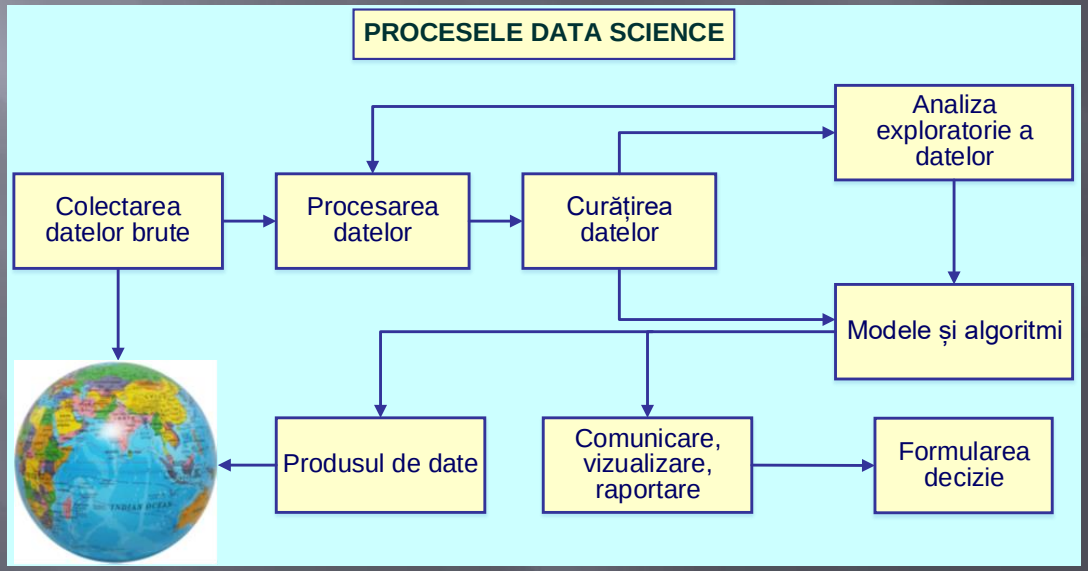
Selecția caracteristicilor: Identificarea și selectarea celor mai relevante variabile.

DEFINIȚII ALE ȘTIINȚEI DATELOR

Știința datelor (Data Science) este un domeniu interdisciplinar care utilizează metode științifice, procese, algoritmi și sisteme pentru a extrage cunoștințe și informații din date structurate și nestructurate. Este un domeniu care combină matematica, statistica, informatica, și cunoștințele de domeniu pentru a analiza și interpreta datele, cu scopul de a obține cunoștințe utile și de a fundamenta deciziile.

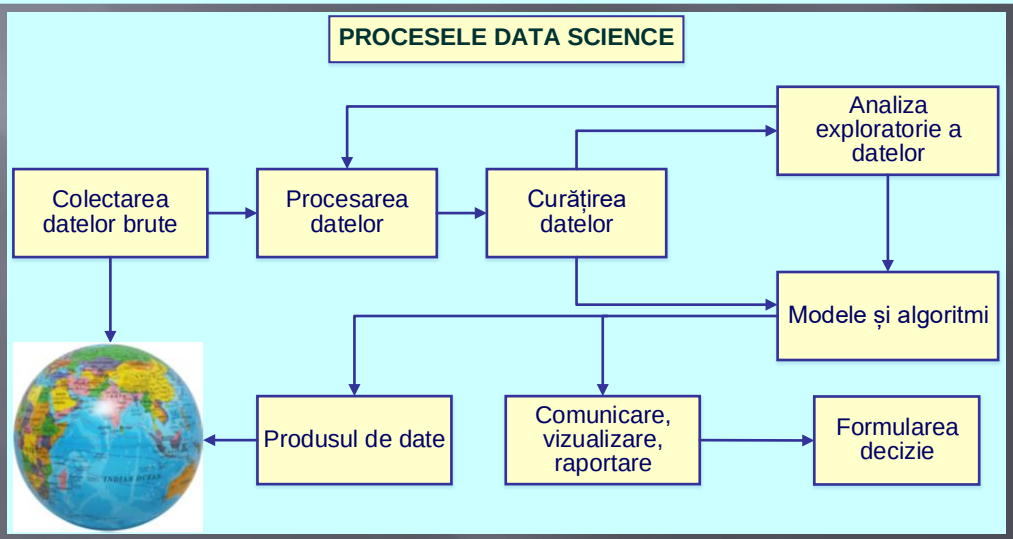
Etapele procesului Data Science sunt:

1.Colectarea și Curățarea Datelor;
2. Explorarea și Vizualizarea Datelor;
3. Modelarea și Analiza Predictivă;
4. Evaluarea și Implementarea Modelului;
5. Comunicarea Rezultatelor;
6. Domenii de utilizare.



ETAPELE PROCESULUI DATA SCIENCE

1. Colectarea datelor brute este prima etapă, în care datele sunt colectate din surse variate ca supoți și formă (text, numere, grafice, imagini, animații, etc.).
2. Procesarea datelor, adică sunt transformarea lor pentru a le face utilizabile;
3. Curățirea datelor pentru a obține un set de date curat, care este gata pentru analize ulterioare.
4. Analiza exploratorie a datelor pentru a înțelege structura și caracteristicile datelor și pentru a identifica tiparele și relațiile din date;
5. Construirea de modele și algoritmi pentru a efectua predicții sau a lua decizii;
6. Utilizare pentru a lua decizii;
7. Rezultatele și deciziile sunt comunicate prin rapoarte și vizualizări;
8. În final, este creat un produs de date care este conectat la comunicare, vizualizare și raportare, și se întoarce la colectarea datelor brute pentru o nouă iterație;
Diagrama subliniază natura iterativă și interconectată a procesului de DS.



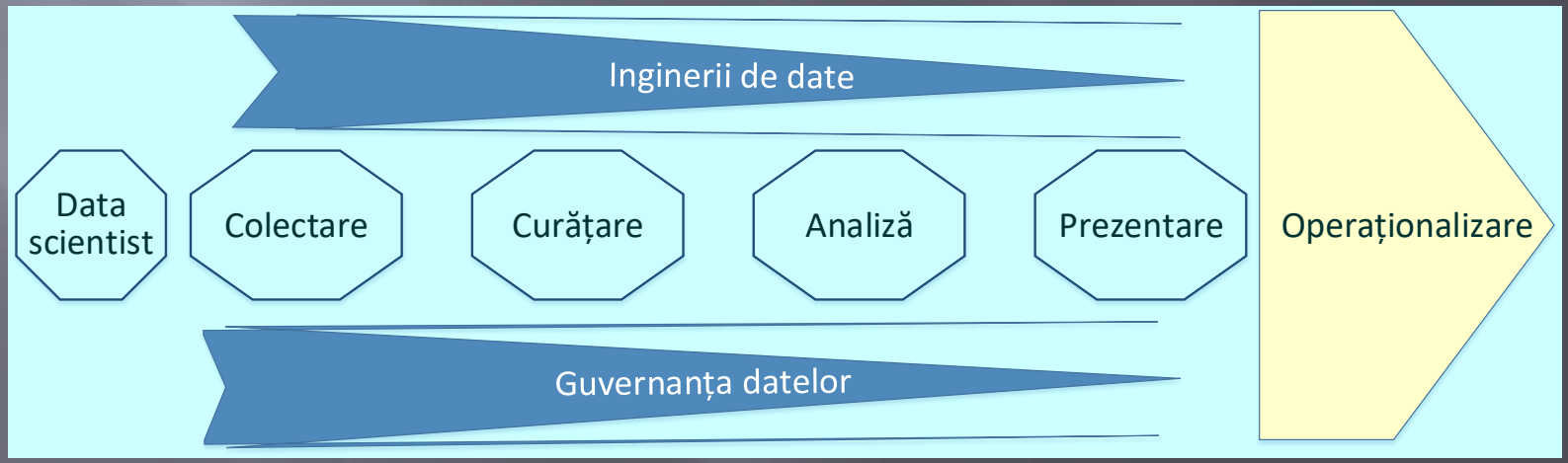
DATA SCIENTIȘTII

Specialiștii în DS se numesc data scientiști. Figura prezintă activitățile principale ale unui data scientist. Activitățile Data Scientist-ului:

- Colectarea datelor - data scientist-ul colectează datele necesare pentru analize ulterioare.
- Datele colectate sunt prelucrate (curățate și transformate) prin: eliminarea valorilor lipsă, corectarea erorilor și transformarea datelor în formate utilizabile.
- Datele sunt analizate identificând tipare și tendințe cu tehnici statistice și algoritmi de ML.
- Rezultatele analizei sunt prezentate folosind vizualizări de date și rapoarte.

1. Contextul mai larg al proiectului de date:

- Deasupra activităților data scientist-ului, este componenta de inginerie care se ocupă cu infrastructura necesară pentru colectarea, stocarea și procesarea datelor. Data scientist-ul trebuie să colaboreze cu inginerii de date.
- Sub activitățile data scientist-ului, guvernanța datelor se referă la politicile și procedurile care asigură utilizarea corectă și etică a datelor.
- Operationalizarea înseamnă integrarea rezultatelor analizei datelor în procesele de afaceri și sisteme operaționale. Data scientist-ul creează valoare din date.



INSTRUMENTE ȘI TEHNOLOGII UTILIZATE ÎN DATA SCIENCE

În Data Science (DS), instrumentele și tehnologiile sunt resursele utilizate pentru a gestiona, analiza și extrage cunoștințe din date.

a. Instrumentele (Tools) în DS sunt software-uri, biblioteci sau aplicații care ajută la efectuarea sarcinilor specifice al DS în colectare, curățare, analiză și vizualizare a datelor.

1. Jupyter Notebooks: Un mediu interactiv foarte utilizat pentru codarea în Python și R, ideal pentru explorarea și documentarea analizei datelor.
2. Pandas: O bibliotecă Python pentru manipularea datelor structurate.
3. NumPy: O bibliotecă pentru calcule numerice și operații pe matrici multidimensionale.
4. Scikit-learn: O bibliotecă Python utilizată pentru învățarea automată (machine learning), oferind instrumente pentru clasificare, regresie, clustering și multe altele.
5. Tableau și Power BI: Instrumente de vizualizare a datelor care permit crearea de rapoarte și dashboard-uri interactive.
6. TensorFlow și PyTorch: Biblioteci folosite pentru învățarea automată, în special pentru rețele neuronale și deep learning.
7. Google Colab: Un mediu bazat pe cloud care permite rularea notebook-urilor Jupyter fără a instala local software-ul.

b. Tehnologiile (Technologies) în Data Science: infrastructura și cadrele de lucru (frameworks) utilizate pentru procesarea, stocarea și gestionarea datelor.

1. Big Data Technologies: Hadoop, Apache Spark, NoSQL Databases;
2. Cloud Technologies: Amazon Web Services (AWS), Microsoft Azure și Google Cloud Platform (GCP): Platforme cloud care oferă resurse și servicii pentru ML și data processing.
3. Data Pipelines: Apache Airflow, Kafka,
4. Machine Learning Technologies: Keras și XGBoost**,** Biblioteci care ajută la crearea și rularea modelelor de ML(machine learning) avansate.

ROLURI ÎN DATA SCIENCE

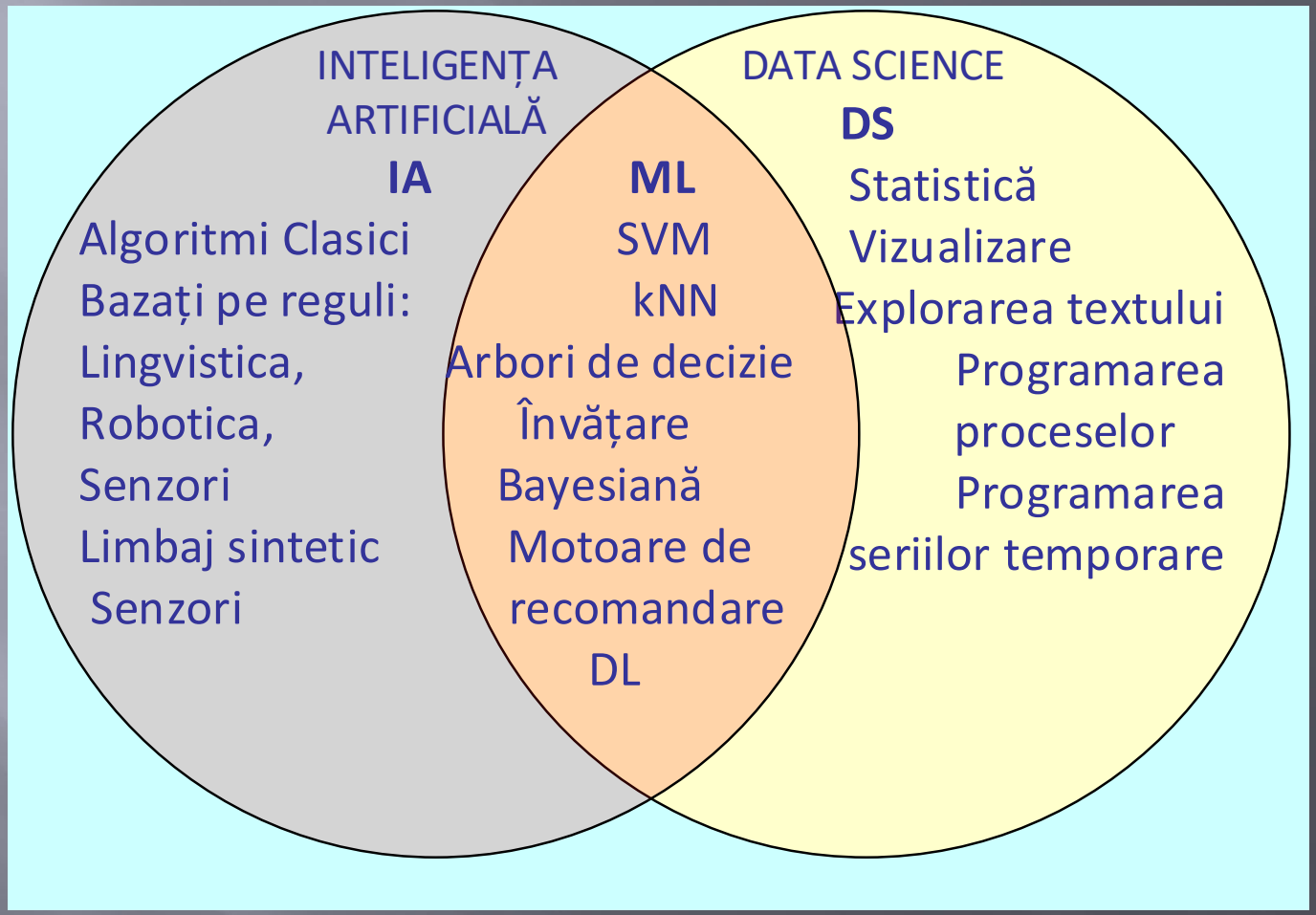
1. Data Scientist: Persoana care combină abilitățile de programare, analiză statistică și cunoștințe de afaceri pentru a analiza datele și a oferi insight-uri.
2. Data Analyst: Se concentrează pe analiza datelor istorice și crearea de rapoarte pentru a răspunde la întrebări specifice.
3. Data Engineer: Creează și gestionează infrastructura necesară pentru colectarea, stocarea și procesarea datelor.
4. Machine Learning Engineer: Specializat în dezvoltarea, implementarea și scalarea modelelor de machine learning.
5. Business Intelligence (BI) Developer: Creează și dezvoltă soluții de raportare și vizualizare a datelor pentru a ajuta organizațiile să ia decizii informate bazate pe date.
6. Data Architect: Proiectează structura și arhitectura bazelor de date și a sistemelor de gestionare a datelor.
7. Statistician: Un expert în metode statistice care aplică modele matematice și tehnici statistice avansate pentru a extrage cunoștințe din date.
8. Data Visualization Expert: Specializat în prezentarea vizuală a datelor.
9. Big Data Engineer: Se ocupă de proiectarea și întreținerea soluțiilor care pot gestiona și analiza seturi mari de date (Big Data).
10. Data Governance Specialist: Asigură respectarea regulilor de guvernanță a datelor în organizații, incluzând politici de confidențialitate și securitate.
11. AI Researcher: Cercetătorii din domeniul AI dezvoltă noi algoritmi și tehnici de ML
Fiecare dintre aceste roluri are o specializare distinctă în cadrul DS.

PROVOCĂRI ÎN DATA SCIENCE

1. Gestionarea volumului uriaș de date (Big Data). Creșterea exponențială a datelor solicită infrastructurile și tehnologiile de stocare și procesare.
2. Curățarea și pre-procesarea datelor. O mare parte din timpul unui Data Scientist este dedicată pregătirii și curățării datelor.
3. Confidențialitatea și securitatea datelor. Protejarea datelor este o preocupare majoră în contextul reglementărilor stricte, precum **GDPR** în Europa sau **CCPA** în California.
4. Explicabilitatea modelelor de M L (Explainable AI). Modelele complexe de ML sunt considerate „cutii negre” din cauza lipsei de transparență în procesul de luare a deciziilor.
5. Lipsa personalului calificat. Cererea de profesioniști în DS depășește cu mult oferta. Se caută oameni cu abilități „hibrid” care să combine statistica, programarea și cunoștințele de afaceri.
6. Bias și echitate în modelele AI. Modelele de AI și ML pot moșteni bias-uri din datele istorice, ceea ce duce la rezultate discriminatorii (de exemplu, în recrutare, acordarea de credite, etc.).
7. Costurile ridicate ale infrastructurii și tehnologiilor. Implementarea și menținerea unor infrastructuri de procesare a datelor este costisitoare pentru companii, în special pentru cele mici.
8. Integrarea datelor din surse diverse (Data Integration). Datele provin din surse multiple și sunt stocate în formate diferite (date structurale, nestructurate, semi-structurate). Integrarea datelor este dificilă și consumatoare de timp.
9. Adaptarea la reglementările și standardele în schimbare. Reglementările legate de date este în evoluție, iar organizațiile trebuie să se adapteze pentru a rămâne conforme cu cerințele juridice.
10. Valoarea practică a datelor și impactul în decizii. Nu toate organizațiile reușesc să transforme datele colectate în insight-uri acționabile. Multe companii întâmpină dificultăți în a crea o cultură bazată pe date și în a lua decizii informate pe baza acestora.

Chiar dacă tehnologiile și datele sunt disponibile, lipsa unei strategii și a unei direcții de acțiune împiedică organizațiile să valorifice avantajele oferite de DS.

RELAȚIA DINTRE DS, ML ȘI IA



Imaginea evidențiază relația dintre Inteligența Artificială (Artificial Intelligence - AI), Învățarea Automată (Machine Learning - ML) și Știința Datelor (Data Science - DS), subliniind cum aceste domenii se intersectează și se completează reciproc.

EXPLICAȚIA RELAȚIEI DINTRE DS, ML ȘI AI (1)

1. Artificial Intelligence (Inteligența Artificială): Reprezintă domeniul care dezvoltă sisteme capabile să îndeplinească sarcini care, în mod normal, necesită inteligență umană. AI include următoarele subdomenii:

- Linguistics (Lingvistică): Procesarea limbajului natural.
- Vision (Viziune): Recunoașterea și interpretarea imaginilor.
- Robotics (Robotică): Dezvoltarea de roboți capabili să execute sarcini fizice.
- Planning (Planificare): Algoritmi care iau decizii în mod autonom.
- Language Synthesis (Sinteză de Limbaj): Generarea de text sau vorbire artificială.
- Sensor (Senzori): Utilizarea senzorilor pentru a capta informații din mediul înconjurător.
- Învățarea Automată (Machine Learning - ML).

CONCLUZII:

1. AI este domeniul care include toate tehnologiile și metodele ce permit unui sistem să simuleze inteligența umană. AI implică atât soluții bazate pe reguli (nonML), cât și soluții bazate pe date și algoritmi de învățare automată (ML).
2. AI cuprinde atât algoritmi tradiționali (care nu necesită date, de exemplu, sistemele expert), cât și algoritmi moderni bazati pe date, cum ar fi Machine Learning și Deep Learning.
3. Exemple: Recunoașterea vocii, asistenți virtuali, roboți autonomi.

EXPLICAȚIA RELAȚIEI DINTRE DS, ML ȘI AI (2)

2. Machine Learning (Învățarea Automată): este o subramură a AI care dezvoltă algoritmi și modele care permit sistemelor să învețe din date, să facă predicții sau să ia decizii fără a fi programate în mod explicit.

În imagine, sunt enumerate câteva tehnici și metode de ML:

- Support Vector Machines (SVM): algoritm utilizat pentru clasificare și regresie.
- kNN (K-Nearest Neighbors): algoritm pentru clasificare bazat pe similitudinea vecinilor apropiați.
- Decision Trees (Arbori de Decizie): Modele utilizate pentru a lua decizii pe baza unor reguli derivate din date.
- Bayesian Learning: Metode de învățare bazate pe teoria probabilităților.
- Deep Learning (Învățare Profundă): O subramură a ML care folosește rețele neuronale cu multe straturi pentru a învăța reprezentări complexe ale datelor.
- Recommendation Engines (Motoare de Recomandare): Sisteme care sugerează produse sau conținut utilizatorilor pe baza datelor anterioare.

CONCLUZII:

1. ML este o subcategorie a AI care se concentrează pe dezvoltarea de algoritmi care permit sistemelor să "învețe" din date și să-și îmbunătățească performanțele fără a fi explicit programate pentru fiecare sarcină.
2. ML se bazează pe date și implică tehnici de analiză statistică și optimizare pentru a construi modele predictive.
3. Exemple: Sistemele de recomandare, recunoașterea imaginilor, clasificarea email-urilor ca spam.

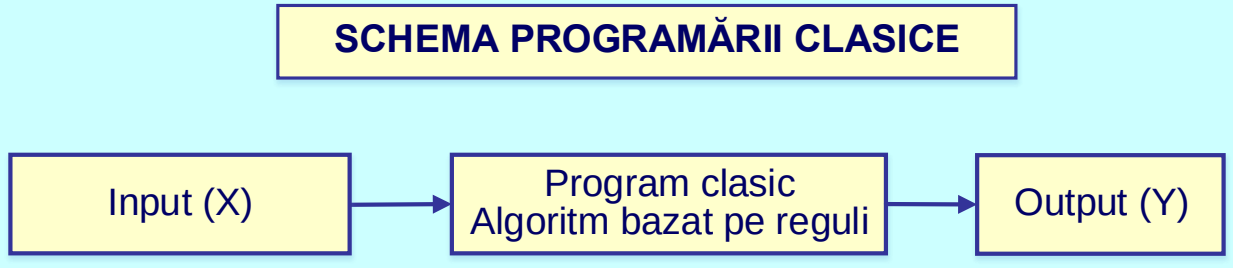
EXPLICAȚIA RELAȚIEI DINTRE DS, ML ȘI AI (3)

3. Data Science (Știința Datelor): Este un domeniu care include Machine Learning, dar acoperă și alte aspecte care nu sunt neapărat legate de AI, cum ar fi:

- Statistics (Statistică): Analiza și interpretarea datelor numerice.
- Visualization (Vizualizare): Crearea de reprezentări grafice ale datelor pentru a comunica informațiile într-un mod clar.
- Text Mining (Explorarea Textului): Analiza datelor nestructurate provenite din text.
- Time Series Forecasting (Prognoza Seriilor Temporale): Predicția valorilor viitoare bazată pe date istorice secvențiale.
- Data Preparation (Pregătirea Datelor): Procesul de curățare, transformare și organizare a datelor înainte de analiză.
- Process Mining (Explorarea Proceselor): Analiza proceselor de afaceri bazată pe date pentru a îmbunătăți eficiența.
- Processing Paradigms (Paradigme de Procesare): Metode și tehnici de gestionare și analiză a datelor, cum ar fi batch processing sau stream processing.
- Experimentation (Experimentare): Testarea ipotezelor și modelelor pentru a evalua eficiența și acuratețea acestora.

Concluzie: Data Science este un domeniu larg care include o serie de tehnici și metode, printre care și Machine Learning. Machine Learning, la rândul său, este o parte a Artificial Intelligence. Toate aceste domenii sunt interconectate și contribuie la crearea de sisteme inteligente care pot învăța din date, analiza, interpreta informații, și formula decizii în mod autonom.

MODELE TRADIȚIONALE DE PROGRAMARE

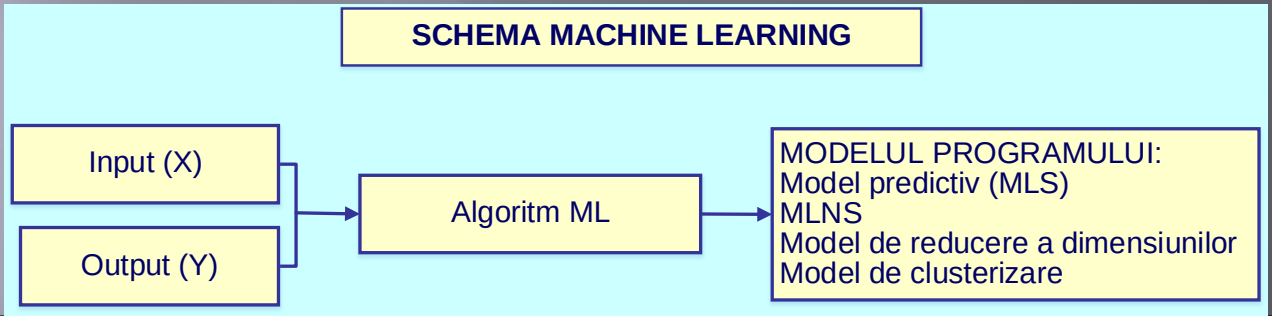


1. Modelul de programare clasică (sau programarea tradițională) se referă la metoda convențională de a scrie cod, în care un programator dezvoltă algoritmi și soluții logice pentru a rezolva o problemă. În acest model, programatorul definește explicit fiecare pas al algoritmului pentru a obține rezultatele dorite.

Caracteristicile modelului de programare clasică sunt:

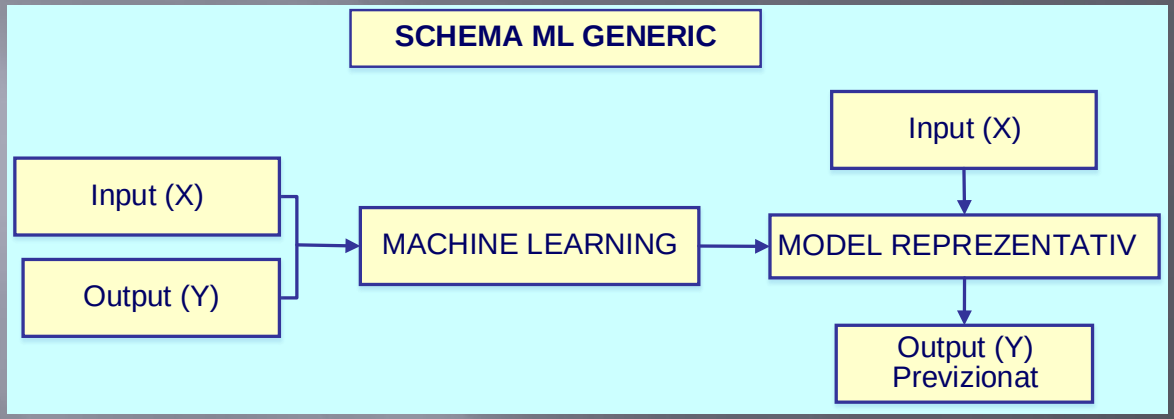
1. Algoritmi expliciți: Programatorul scrie în mod explicit fiecare pas necesar pentru ca un computer să îndeplinească o sarcină.
2. Reguli clare și logice: Codul este scris bazat pe reguli clar definite, iar rezultatele sunt obținute prin urmarea strictă a acestor reguli. Intrările sunt procesate prin intermediul algoritmului, iar rezultatele sunt deterministe – aceleași intrări produc întotdeauna aceleași ieșiri.
3. Control complet asupra codului: Programatorul controlează toate detaliile implementării, de la fluxul logic până la utilizarea resurselor de memorie și procesare.
4. Input/Output clar: Un set de date de intrare este procesat de algoritm pentru a obține un rezultat specific.
5. Exemple de paradigme: Programare procedurală; Programare orientată pe obiecte; Programare funcțională.

MODELE MACHINE LEARNING



2. Învățare Automată (Machine Learning):
Input (X): Setul de date de intrare, similar ca în programarea tradițională.
Output (Y): Rezultatele dorite pentru datele de intrare, cunoscute de obicei în timpul procesului de antrenament.
Machine Learning: În loc să scrie manual regulile pentru a transforma datele de intrare în ieșiri, algoritmii ML învață aceste reguli din date. În acest caz, atât datele de intrare, cât și rezultatele dorite (output) sunt folosite pentru a antrena modelul. Modelul învățat devine „programul” care poate fi folosit pentru a face predicții pe noi seturi de date.
Model of the Program: Rezultatul procesului de învățare automată este un model care poate fi folosit pentru a transforma noi seturi de date de intrare, în predicții.
Concluzie:
Programare Tradițională: Regulile și logica sunt scrise explicit de programatori pentru a produce rezultate specifice.
ML: Modelul învață regulile din datele existente și este capabil să generalizeze pentru a produce rezultate pe noi date. Imaginea subliniază diferența dintre programarea tradițională și ML care permite sistemelor să învețe din date și să genereze reguli automat, făcându-le mai flexibile și adaptabile la noi situații.

CREAREA ȘI UTILIZAREA UNUI MODEL ML



1. Input (X) și Output (y) pentru Învățarea Automată: În partea stângă a imaginii sunt datele de intrare (Input X) și ieșirile dorite (Output y) care sunt introduse în algoritmul ML.

ML: Algoritmul ML folosește aceste date pentru a învăța relațiile dintre intrări și ieșiri. Este etapa de antrenare a modelului, în care algoritmul „învață” pe baza datelor furnizate.

2. Model Reprezentativ (Representative Model): Rezultatul procesului de ML este un model reprezentativ, care capturează relațiile învățate dintre inputuri și outputuri. Acest model poate fi considerat o „regulă” sau un „algoritm” care este folosit pentru a face predicții pe noi date.

3. Predicții pe bază de model:

După ce modelul a fost construit, acesta este aplicat pe noi seturi de date de intrare (Input X). Modelul folosește cunoștințele învățate în timpul antrenamentului pentru a genera Predicted Output (ŷ), care este predicția sa pentru ieșirile corespunzătoare acestor noi intrări.

Concluzie: Imaginea ilustrează clar diferența între faza de antrenament a unui model de învățare automată și utilizarea acelui model pentru a face predicții. În faza de antrenament, modelul învață din datele de intrare și ieșire pentru a crea un model reprezentativ. După ce modelul este construit, acesta poate fi folosit pentru a face predicții pe noi date de intrare.

TEHNICI UTILIZATE ÎN DATA SCIENCE

1. Classification (Clasificare): Procesul de a atribui categorii unui set de date.
2. Regression (Regresie): Tehnică folosită pentru a prezice o valoare numerică continuă pe baza unor variabile de intrare.
3. Clustering (Clustering): Tehnica de grupare a datelor similare împreună într-un cluster, fără a cunoaște în prealabil etichetele grupurilor. Este folosită pentru segmentarea clienților, analiza comportamentală, etc.
4. Association Analysis (Analiza Asocierii): Identificarea regulilor de asociere între variabile într-un set de date. Este utilizată în analizele de coș de cumpărături, cum ar fi „clienții care cumpără produsul X tind să cumpere și produsul Y”.
5. Anomaly Detection (Detectarea Anomaliei): Procesul de identificare a datelor care nu se conformează unui tipar obișnuit. Este utilizată în detectarea fraudei, a erorilor de fabricație.
6. Recommendation Engines (Motoare de Recomandare): Sisteme care sugerează produse, servicii sau informații utilizatorilor pe baza preferințelor și comportamentului lor trecut.
7. Deep Learning (DL): Un subdomeniu al ML care folosește RNN pentru a învăța reprezentări complexe ale datelor. Este utilizată în recunoașterea imaginii, procesarea limbajului natural.
8. Time Series Forecasting (Prognoza seriilor temporale): Tehnica de analiză care face predicții despre viitor. Este folosită pentru prognoza vânzărilor, prognoza cererii de energie, etc.
9. Text Mining (Explorarea Textului): Procesul de analiză a datelor textuale pentru a extrage informații utile. Este utilizată în analiza sentimentelor, extragerea entităților din text.
10. Feature Selection (Selecția Caracteristicilor): Procesul de a alege cele mai relevante variabile (caracteristici) dintr-un set de date pentru a îmbunătăți performanța modelului și a reduce complexitatea acestuia.

SARCINI (TASKS) ÎN DATA SCIENCE

Tasks (Sarcini)	Description (Descriere)	Algorithms (Algoritmi)	Examples (Exemple)
Classification (Clasificare)	Predicția unei clase sau categorii pentru un punct de date, bazată pe un set de date cunoscut.	Arbori de decizie, rețele neuronale, modele bayesiene, k-nearest neighbors (kNN)	Asignarea alegătorilor în grupuri politice cunoscute, clasificarea noilor clienți în categorii existente
Regression (Regresie)	Predicția unei valori numerice țintă pentru un punct de date.	Regresia liniară și regresia logistică	Predicția ratei șomajului pentru anul viitor, estimarea primei de asigurare
Anomaly Detection (Detectarea Anomaliei)	Identificarea punctelor de date care sunt neobișnuite în comparație cu restul dataset-ului.	Metode bazate pe distanță, densitate și LOF (Local Outlier Factor).	Detectarea tranzacțiilor frauduloase pe carduri de credit, detectarea intruziunilor în rețea.
Time Series (Serii Temporale)	Predicția valorilor unei variabile țintă pentru un cadru temporal viitor, bazată pe valorile istorice	Metode precum netezirea exponențială, ARIMA și regresia	Prognoza vânzărilor, prognoza producției, extrapolarea fenomenelor de creștere.
Clustering (Grupare)	Gruparea naturală a datelor în clustere bazate pe proprietăți comune.	K-means și clusteringul bazat pe densitate (DBSCAN).	Identificarea segmentelor de clienți într-o companie pe baza datelor de tranzacții, web și apeluri.
Association Analysis (Analiza Asocierii)	Identificarea relațiilor între articole într-un set de date tranzacțional	Algoritmi precum FP-Growth și Apriori	Identificarea oportunităților de vânzare încrucișată pentru un retailer, bazată pe istoricul tranzacțiilor.

ISTORICUL DATA SCIENCE (1)

Tehnicile statistice moderne, precum regresia liniară, logistică, clustering și învățarea automată (arbori de decizie, rețele neuronale), permit analiza avansată a datelor, iar metode precum PCA ajută la reducerea dimensiunii.

Digitizarea transformă informația din format analog în digital, alimentând fenomene precum IoT, IIoT, Industry 4.0, blockchain și criptografie.

Transformarea digitală, rezultat al digitalizării, afectează modelele de afaceri, structurile socio-economice și organizaționale.

Termenul "Data Science" a apărut în anii 1960 pentru a interpreta datele mari, evoluând pentru a include IA, ML și IoT.

John Tukey, pionier al integrării statisticii cu calculatoarele, a subliniat în 1962 importanța analizei datelor.

În 1974, Peter Naur a definit termenul "Data Science" prin utilitatea datelor în modelarea realității, iar în 1977, Tukey a promovat analiza exploratorie a datelor (EDA).

Prima conferință ACM despre descoperirea cunoștințelor în baze de date a avut loc în 1989.

În 1994, Business Week a evidențiat provocările legate de gestionarea și analiza cantităților mari de date personale colectate de companii.

ISTORICUL DATA SCIENCE (2)

În 2001 și 2002 au avut loc evenimente care au pus bazele dezvoltării moderne a DS și a infrastructurii de procesare a Big Data.

În 2001, William S. Cleveland a avut o contribuție remarcabilă prin definirea unui plan pentru formarea viitorilor specialiști în DS. Acest plan a inclus șase domenii-cheie: statistică, bazele de date și managementul informațiilor, algoritmii și mașinile, vizualizarea și grafica, abilitățile de comunicare și contextul aplicativ al expertizei. Această structură a devenit esențială pentru evoluția DS și pentru formarea profesioniștilor cunoscuți astăzi drept *data scientists*.

În paralel, Consiliul Internațional pentru Știință a început publicarea *Data Science Journal* în 2002, un jurnal dedicat datelor și aspectelor asociate acestora, cum ar fi descrierea sistemelor de date și aspectele juridice ale datelor. Acesta a subliniat nevoia de standarde și evaluare riguroasă în publicarea de cercetări legate de date.

O altă piatră de temelie a fost lansarea în 2006 a versiunii 0.1.0 a Hadoop. Aceasta a marcat începutul unei noi ere în procesarea și stocarea datelor distribuite la scară mare. Hadoop Distributed File System (HDFS) a rezolvat două mari provocări ale Big Data: stocarea volumelor mari de date și procesarea acestora în mod eficient, depășind limitările bazelor de date relaționale tradiționale (RDBMS). Fiind open-source și scalabil, Hadoop a devenit un standard în gestionarea datelor mari, utilizat pe scară largă pentru a răspunde nevoilor de procesare a informațiilor.

Aceste contribuții au pregătit terenul pentru dezvoltări ulterioare în DS și tehnologiile de Big Data.

ISTORICUL DATA SCIENCE (3)

Începând cu 2008 are loc ascensiunea și popularizarea termenului „data scientist”, promovarea sa de către DJ Patil și Jeff Hammerbacher și rolul crucial jucat de aceștia în dezvoltarea domeniului științei datelor în cadrul companiilor de tehnologie. În 2012, Universitatea Harvard a numit „data scientist” cea mai sexy slujbă a secolului XXI.

Termenul NoSQL a reapărut în 2009 pentru a descrie baze de date non-relaționale, iar în 2011 a crescut semnificativ cererea pentru specialiști în știința datelor. În același an, conceptul de „lacuri de date” a fost popularizat, oferind o metodă de stocare a datelor brute într-un mod neclasificat.

În 2013, IBM a raportat că 90% din datele globale au fost create în ultimii doi ani, iar în 2015, tehnologiile de învățare profundă au permis îmbunătățiri semnificative în recunoașterea vorbirii prin Google Voice. Anul 2015 a fost, de asemenea, marcat de creșterea masivă a utilizării AI în cadrul Google, cu peste 2.700 de proiecte care foloseau inteligența artificială.

În concluzie, știința datelor a evoluat rapid în ultimele decenii, devenind esențială pentru guverne, companii și cercetători, extinzându-se în diverse domenii precum traducerea automată, robotică, sănătate, economie și științe sociale. Această evoluție a influențat profund economiile, guvernele și afacerile.

LUCRĂRI DE REFERINȚĂ (4)

Lucrarea „Viitorul analizei datelor” (1962) a lui John Tukey a avut un impact major asupra statisticii, introducând conceptul de „explorare a datelor” și promovând utilizarea metodelor grafice și vizuale pentru analiza interactivă a datelor. Tukey a propus o abordare mai flexibilă a analizei datelor, influențând dezvoltarea statisticii exploratorii.

Contribuțiile lui Dennis V. Lindley la statistica Bayesiană, evidențiate în lucrarea sa „Statistical Inference” (1971), au fost esențiale în promovarea acestei abordări statistice. Lindley a detaliat metodele Bayesiane pentru estimare, testarea ipotezelor și luarea deciziilor, cu aplicații în diverse domenii.

Modelul Cox de regresie proporțională (1972), dezvoltat de David Cox, este o metodă semiparametrică utilizată pentru analiza supraviețuirii. Acest model a devenit esențial în analizarea riscurilor proporționale fără a presupune o formă specifică a distribuției datelor, fiind aplicat în medicină, biologie, inginerie și economie.

Aceste contribuții au avut un impact semnificativ asupra statisticii și cercetării aplicate în multiple domenii.

Bootstrap și MCMC, 1979. 1980 (5)

Bootstrap și MCMC sunt două metode statistice esențiale dezvoltate în anii '70 și '80 care au influențat considerabil analiza datelor și inferența statistică.

Bootstrap, introdus de Efron în 1979, este o metodă de eșantionare prin simulare care permite estimarea preciziei unor statistici fără a face presupuneri despre distribuția datelor. Această metodă a devenit esențială în inferența statistică și este utilizată într-o gamă largă de domenii precum biologia, medicina, economia și științele sociale.

MCMC (Markov Chain Monte Carlo), introdusă în anii '50 și dezvoltată în anii '70-'80, este o metodă bayesiană care permite estimarea parametrilor în modele statistice complexe, atunci când calculele analitice nu sunt posibile. MCMC este acum o metodă standard în inferența bayesiană și este folosită în numeroase domenii precum biologia, fizica și științele sociale.

Ambele metode au revoluționat analiza datelor și au oferit cercetătorilor instrumente puternice pentru abordarea problemelor complexe.

MAȘINI VECTORIALE SUPORT - SVM, 1996 (6)

Mașinile vectoriale de suport (Support Vector Machines, SVM) sunt o metodă de învățare supervizată introdusă de Vapnik și colegii săi în 1996. SVM este o tehnică de clasificare și regresie utilizată în analiza datelor și în machine learning, care se bazează pe conceptul de hiperplanuri de separare optimă între două clase de date.

SVM este cunoscută pentru abilitatea sa de a gestiona atât date liniar separabile, cât și date neliniar separabile, utilizând tehnici de transformare a datelor într-un spațiu de caracteristici de dimensiuni mai mari, denumit și "trucul kernel" (kernel trick). SVM își găsește aplicații într-o gamă largă de domenii, cum ar fi clasificarea imaginilor, recunoașterea textului scris, detectarea fraudelor, medicină și bioinformatică, precum și în multe alte domenii.

SVM a fost un avans semnificativ în domeniul învățării supervizate, oferind o abordare eficientă și flexibilă pentru clasificare și regresie pe baza conceptului de hiperplanuri de separare optimă. De-a lungul anilor, SVM a fost dezvoltată și extinsă, incluzând variante precum SVM cu mașini vectoriale de suport multiple (Multi-Class SVM), SVM cu costuri asimetrice, SVM învățare online și altele.

ANII 2000 ȘI 2010 (7)

În anii 2000, micromatricele și testarea multiplă neuroimagistică au fost două concepte esențiale în neuroștiințe și analiza datelor. Micromatricele, utilizate pentru a măsura expresia genelor, au avut un impact major în biologie moleculară și medicină, oferind informații valoroase despre diferențele genetice și posibile ținte terapeutice. În același timp, testarea multiplă neuroimagistică, folosită în analiza imaginilor cerebrale (fMRI, EEG, MEG), a permis vizualizarea activității cerebrale și a necesitat metode statistice avansate pentru gestionarea testării multiple, cum ar fi corecțiile pentru multiple comparații și machine learning.

În anii 2010, rețelele neuronale au evoluat semnificativ datorită învățării profunde (deep learning). Progresele tehnologice au permis dezvoltarea rețelelor neuronale profunde, cu multiple straturi, care pot învăța reprezentări complexe ale datelor, îmbunătățind performanțele în domenii precum recunoașterea de imagini, prelucrarea limbajului natural și analiza datelor secvențiale. Tehnologii precum CNN și RNN, împreună cu variantele LSTM și atenția neurală, au devenit fundamentale pentru inteligența artificială, influențând sectoare precum medicina, industria auto și finanțele.

ȘTIINȚA DATELOR, 2015 (8)

Anul 2015 a fost un an important pentru domeniul științei datelor, marcat de creșterea interesului și adopției acestei discipline în afaceri, știință și tehnologie. A fost observată o expansiune a tehnologiilor și uneltelor de analiză a datelor, cum ar fi Apache Spark, Hadoop și Tableau. Învățarea automată a avut un avans semnificativ, cu algoritmi precum regresia logistică și rețele neuronale, devenind esențiali pentru analiză și predicție.

Vizualizarea datelor și protecția confidențialității au devenit puncte centrale, iar cererea pentru specialiști în știința datelor a continuat să crească.

Termenul "Data Science" a apărut în anii '60-'70, dar a căpătat amploare în anii 2000, pe măsură ce cantitatea de date a explodat și tehnologiile de analiză au evoluat.

Cărți și articole notabile din 2015 și 2016, precum "Deep Learning" de Ian Goodfellow și "Human-level control through deep reinforcement learning" de Volodymyr Mnih, au avut un impact major asupra domeniului.

CONCLUZII PRIVIND APARIȚIA ȘI CONSACRAREA DS (9)

Termenul "Data Science" a devenit popular în anii 2000, odată cu nevoia de a descrie procesul științific de extragere a informațiilor din date. Știința Datelor s-a dezvoltat din domenii precum statistica, inteligența artificială, învățarea automată și vizualizarea datelor.

Creșterea volumului de date și avansul tehnologic au contribuit la formarea acestei discipline esențiale în domenii precum afaceri, medicină și cercetare.

Deși nu există o lucrare singulară care să marcheze "nașterea" științei datelor, mai multe lucrări importante au influențat dezvoltarea sa. Printre acestea se numără "The Art of Data Science" (2013) și "Data Science for Business" (2013). Contribuții esențiale au venit și din lucrări clasice precum "The Elements of Statistical Learning" (2001) și "Pattern Recognition and Machine Learning" (2006). Știința Datelor continuă să evolueze pe măsură ce apar noi tehnologii și cercetări.

PERSPECTIVA DATA SCIENCE

Se aşteaptă o creştere a cererii de specialişti în DS, deoarece multe industrii și organizaţii vor să se bazeze pe analiza de date pentru a lua decizii fundamentate și a-și îmbunătăţi operaţiunile:

1. Integrarea cu *tehnologiile emergente*: Integrarea DS cu tehnologii precum IoT, 5G, inteligenta artificială și blockchain va juca un rol cheie în dezvoltarea și evoluţia acestei discipline.
2. Analiza de date în timp real: Dezvoltarea de instrumente și tehnologii care să permită analiza de date în timp real va fi esenţială pentru optimizarea proceselor operaţionale și luarea deciziilor în timp real.
3. Protejarea datelor și confidenţialitatea: Protejarea datelor și confidenţialitatea vor fi probleme critice, care vor necesita abordări inovatoare din partea DS pentru a asigura protecţia datelor și securitatea acestora.
4. Analiza de date la scară mare: Datele vor continua să crească exponenţial, astfel ca DS va trebui să dezvolte noi tehnici și instrumente pentru a gestiona și a analiza aceste date la scară mare.
5. Interfeţe vizuale și interactive: Interfeţe vizuale și interactive vor fi tot mai importante, deoarece vor permite oamenilor să interacţioneze cu datele și să înţeleagă rezultatele analizei de date într-un mod mai accesibil și intuitiv.

INSTRUMENTE UTILIZATE ÎN DS

- Limbaje de programare: Python, R, SQL, etc.
- Framework-uri și biblioteci: TensorFlow, PyTorch, Pandas, NumPy, etc.
- Software de analiză de date: Excel, Tableau, Power BI, etc.
- Instrumente de stocare a datelor: Hadoop, Spark, etc.
- Instrumente de gestionare a bazelor de date: MySQL, PostgreSQL, MongoDB, etc.
- Instrumente de analiză statistică: SAS, SPSS, R Studio, etc.
- Instrumente de învățare automata: Jupyter Notebook, Google Colab, etc.
- Software de gestionare a proiectelor (Project Management Software)
- Instrumente de colaborare (Collaboration Tools)
- Software de prezentare (Presentation Software)
- Instrumente de monitorizare a performantei (Performance Monitoring Tools)
- Instrumente de testare (Testing Tools)
- Instrumente de gestionare a versiunilor (Version Control Tools)
- Instrumente de automatizare (Automation Tools)
- Instrumente de machine learning operational (Operational Machine Learning Tools)
- Instrumente pentru implementarea solutiilor (Solution Deployment Tools)
- Instrumente de management al calitatii datelor (Data Quality Management Tools)

TEHNICILE UTILIZATE ÎN DATA SCIENCE

- Analiza exploratorie de date (Exploratory Data Analysis, EDA);
- Regresie;
- Clasificare;
- Clustering;
- Reducerea dimensionalității;
- Algoritmi de învățare automată (Machine Learning) (MLS, MLNS, RL, DL);
- Retele neuronale (Neural Networks);
- Boosting și Bagging;
- Algoritmi de optimizare (Gradient Descent, Evolutionary Algorithms, etc.)
- Procesare de date în timp real (Stream Processing)
- Text Mining și Analiza de discurs
- Recomandare (Recommender Systems)
- Vizualizarea datelor (Data Visualization)
- Minerit de date (Data Mining)
- Modelare predictivă (Predictive Modeling)
- Analiza de date spatio-temporala (Spatio-Temporal Data Analysis)
- Imbinarea datelor (Data Fusion)
- Algoritmi de decizie (Decision Algorithms);
- Analiza de secvențe (Sequence Analysis);
- Analiza de grafuri (Graph Analysis);
- Deep Learning (DL);
- Transfer Learning;
- Reinforcement Learning.

PROCESUL DE LUCRU ÎN DATA SCIENCE

Cele mai utilizate instrumente și tehnologii în DS:

1. Limbaje de programare: Python și R sunt cele mai populare limbaje de programare utilizate în DS, datorită numeroaselor biblioteci și instrumente disponibile pentru analiza de date și ML.
2. Biblioteci și instrumente de analiza de date: Pândaş, NumPy, Matplotlib sunt biblioteci populare pentru prelucrarea și explorarea datelor în Python
3. Algoritmi de machine learning: Algoritmi precum regresia liniara, arborele de decizie, random forest și rețelele neuronale sunt utilizaţi în DS pentru prezicerea sau clasificarea datelor.
4. Sisteme de management de baze de date: MySQL, PostgreSQL, MongoDB sunt sisteme de management de baze de date utilizate pentru stocarea și gestionarea datelor în proiectele de DS.
5. Platforme de analiză de date: Hadoop, Spark, AWS Glue sunt platforme populare utilizate pentru prelucrarea și analiza de date la scară mare.
6. Instrumente de vizualizare de date: Tableau, Power BI, Matplotlib sunt instrumente utilizate pentru crearea de vizualizări interactive și inteligibile.
7. Platforme de colaborare: GitHub, GitLab, Bitbucket.
8. Instrumente de automatizare: Jenkins, Travis CI, CircleCI sunt instrumente utilizate pentru automatizarea și testarea proceselor de integrare și livrare a codului.
9. Sisteme de ML: TensorFlow, PyTorch, scikit-learn sunt biblioteci de ML.
10. Instrumente de prelucrare a limbajului natural: NLTK, spaCy, Gensim sunt biblioteci utilizate pentru prelucrarea textelor și analiza sentimentelor.
11. Platforme de cloud computing: AWS, Google Cloud, Microsoft Azure sunt platforme de cloud computing utilizate pentru stocarea și procesarea de date la scară mare.
12. Instrumente de integrare cu alte sisteme: APIs, integrări cu sisteme CRM sau ERP, integrări cu platforme de analytics sunt instrumente utilizate pentru integrarea datelor și sistemelor în proiecte DS.

CONCLUZII PRIVIND DS

DS se ocupă cu colectarea, stocarea, transformarea și analiza datelor pentru a extrage cunoștințe utile în luarea deciziilor. În era modernă, DS este esențială datorită cantităților mari de date generate, fiind comparată cu resursele clasice de producție, cum ar fi munca sau capitalul. DS permite companiilor să optimizeze procesele, să dezvolte produse personalizate și să îmbunătățească eficiența în diverse domenii precum sănătatea, energia și securitatea.

De-a lungul timpului, DS a evoluat cu începuturi în timpul celui de-al Doilea Război Mondial, când matematicienii analizau date pentru rezolvarea problemelor de război.

În anii '60 și '70 au apărut primele baze de date și software-uri de analiză (Excel, SAS, SPSS), iar în anii '80-'90, ML și analiza de date avansată au extins capacitatea de a explora date.

În prezent, DS se bazează pe tehnologii moderne precum Big Data, Cloud Computing și DL. Acesta are aplicații vaste, de la afaceri și sănătate la guvernare și cercetare științifică. Procesul de lucru în DS este iterativ și cuprinde pași precum definirea problemei, curățarea datelor și interpretarea rezultatelor.

Instrumentele utilizate includ limbaje de programare (Python, R), baze de date (SQL, NoSQL), platforme de analiză (Azure, AWS) și pachete de analiză precum TensorFlow sau Scikit-learn. Statistica joacă un rol important în DS, ajutând la analiza datelor și construirea de modele.

Pe viitor, DS va continua să evolueze, fiind strâns legată de IA, Big Data și automatizare, având potențialul de a transforma economia și de a rezolva probleme complexe într-un mod etic și responsabil.