CURS 1: EXTRACȚIA ȘI TRANSFORMAREA DATELOR

1: Extracția și transformarea datelor

Extracţia şi transformarea datelor

Curs 1 - Web mining și data analysis Durată: 1 oră și 20 minute

2: Importanța datelor în AI

- Modelele AI se bazează pe date de calitate pentru acuratețe.
- Garbage in -> Garbage out: datele slabe duc la modele slabe.
- Datele sunt necesare pentru:
  - Antrenarea modelelor de învățare automată
  - Realizarea de predicții
  - Înțelegerea tendințelor și comportamentelor
- Exemplu: Un model de detectare a spamului eșuează atunci când este antrenat pe etichete de e-mail incorecte

3: Tehnici de colectare a datelor

API-uri (Interfețe de Programare a Aplicațiilor) - Oferă acces direct la date structurate din diverse surse. - Exemplu: Utilizarea API-ului Coindesk pentru a colecta titluri despre bitcoin pentru analiza sentimentului

Web Scraping - Extrage informații de pe site-uri web atunci când API-urile nu sunt disponibile. - Necesită parsarea HTML. - Unelte: BeautifulSoup, Scrapy. - Exemplu: Scraping al prețurilor produselor de pe un site de e-commerce

Baze de date - Bazele de date SQL și NoSQL stochează date structurate. - Exemplu: Interogarea unei baze de date Sqlite pentru a obține date despre tranzacțiile utilizatorilor

4: Prezentare generală a preprocesării datelor

- Datele nu sunt rareori curate; preprocesarea este esențială.
- Pași importanți:
  - Gestionarea valorilor lipsă
  - Normalizare și scalare
  - Codificarea variabilelor categorice
  - Detectarea și eliminarea valorilor aberante
- Exemplu: Curățarea unei baze de date de clienți prin eliminarea înregistrărilor duplicate

5: Gestionarea valorilor lipsă

- Datele lipsă pot afecta performanța modelului.
- Strategii:
  - Eliminarea rândurilor/coloanelor cu prea multe valori lipsă.
  - Imputarea valorilor lipsă folosind media, mediana sau moda.
  - Prezicerea valorilor lipsă folosind modele de învățare automată.
- Exemplu: Gestionarea valorilor lipsă pentru vârstă într-un set de date despre supraviețuirea pe Titanic

6: Normalizare și scalare

- Normalizarea scalează datele la un interval fix (ex. 0-1).
- Standardizarea centrează datele în jurul valorii zero.
- Esențială pentru modelele care se bazează pe distanțe (ex. k-NN, SVM).
- Exemplu: Scalarea caracteristicilor numerice într-un set de date despre prețurile locuințelor

7: Codificarea cariabilelor categorice

- Modelele de învățare automată necesită date numerice.
- Strategii:
  - One-hot encoding (pentru date nominale)
  - Label encoding (pentru date ordinale)
- Exemplu: Codificarea genului (masculin/feminin) în valori binare

8: Ingineria și selecția caracteristicilor

- **Ingineria Caracteristicilor**: Crearea unor noi caracteristici semnificative din date existente.
- **Selecția Caracteristicilor**: Eliminarea caracteristicilor irelevante sau redundante.
- Tehnici:
  - Analiza corelației
  - Eliminarea recursivă a caracteristicilor (RFE)
  - Analiza Componentelor Principale (PCA)
- Exemplu: Extragerea cuvintelor cheie din recenziile produselor pentru îmbunătățirea analizei sentimentului

9: Gestionarea seturilor de date dezechilibrate

- Problema: O clasă domină setul de date.
- Soluții:
  - Resampling (supra-eșantionarea minorității, sub-eșantionarea majorității).
  - Utilizarea funcțiilor de pierdere ponderate în antrenarea modelului.
  - Generarea de date sintetice (tehnica SMOTE).
- Exemplu: Gestionarea dezechilibrului de clasă în seturile de date pentru detectarea fraudelor

10: Rezumat & concluzii

- Datele de înaltă calitate sunt esențiale pentru succesul AI.
- Metode de colectare a datelor: API-uri, web scraping, baze de date.
- Pași de preprocesare: curățare, gestionarea valorilor lipsă, scalare, codificare.
- Ingineria și selecția caracteristicilor îmbunătățesc performanța modelului.
- Gestionarea dezechilibrului de clasă este crucială pentru predicții corecte.

11: Conținutul seminariului

- Extracţia datelor dintr-un API (ex. Twitter, OpenWeather).
- Exemplu de web scraping: extragerea detaliilor produselor de pe un site de e-commerce.
- Preprocesarea şi curăţarea unui set de date din lumea reală.
- Gestionarea valorilor lipsă şi ingineria caracteristicilor.
- Aplicarea tehnicilor de scalare şi codificare în practică.

12: Întrebări & discuții

- Sesiune deschisă pentru întrebări şi discuţii.