Proiecte Data Mining (2023-2024)

Temele sunt grupate în 3 categorii:

- A: Proiecte orientate către algoritmi
- B: Proiecte orientate către seturi de date
- C: Participare la o competiție de analiză a datelor
    - **FedCSIS 2024 Challenge: Predicting stock trends**
    - Task description, enrollment and data files:
        - https://knowledgepit.ai/fedcsis-2024-challenge/

Obs.

1. Bibliografia de start este disponibilă pe Classroom sau la link-urile specificate pentru fiecare temă.
2. Proiectele pot fi realizate individual sau în echipe de 2 studenți (cu specificarea clară a contribuției fiecăruia)

A. Proiecte orientate către algoritmi

Proiectele de tip A constau în:

- Un **raport** (cca 6-8 pagini) în care sunt descrise particularitățile problemei abordate (clasificare, grupare, regresie, analiza asocierii, prelucrare serii temporale), este prezentat cel puțin un algoritm de rezolvare (folosind bibliografia de start și eventual alte lucrări) și sunt prezentate rezultatele obținute aplicând algoritmul implementat (pentru seturi de date la alegere).
- Structura raportului:
    - Abstract (2-3 paragrafe): se descrie pe scurt care sunt obiectivele proiectului si care sunt principalele rezultate
    - Introducere: se descrie problema abordată și abordările existente (pe baza bibliografiei); se prezintă pe scurt ideea de rezolvare și modul în care e structurat raportul
    - Descrierea metodei: se descrie metoda (pe baza bibliografiei)
    - Descrierea implementării: se furnizează detalii de implementare inclusiv prin referire la codul sursă
    - Prezentarea rezultatelor testării pe un set de date (la alegere)
    - Concluzii: descrierea principalelor provocări întâlnite la implementare și a eventualelor direcții de îmbunătățire
- **Implementarea** de la zero a unui algoritm (limbajul de programare este la alegere – Python, R, Java, C++ etc).

Tematici pentru proiecte de tip A

1. Algoritmi pentru selecția atributelor (implementarea algoritmului Relief sau a unui algoritm greedy de tip forward). Biblio: FeatureSelection
2. Algoritmi pentru discretizarea atributelor (implementarea algoritmului Holte 1R). Biblio: FeatureDiscretization
3. Algoritmi pentru pre-procesarea seturilor de date debalansate (de exemplu SMOTE - https://arxiv.org/pdf/1106.1813.pdf, https://jmlr.org/papers/v18/16-365)
4. Algoritmi pentru construirea arborilor de decizie oblici. Biblio: DecisionTree archive, https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/ , https://www.ijcai.org/Proceedings/2020/0750.pdf , https://github.com/aia-uclouvain/pydl8.5
5. Algoritmi de acoperire cu reguli (implementarea algoritmului PRISM + Random PRISM – comparație cu implementarea de la https://github.com/dahvreinhart/Rule-Based-PRISM/blob/master/prism.py). Biblio: CoveringAlgorithms archive
6. Clasificator Naïve Bayes (implementarea unui algoritm pentru clasificare multiplă – comparație cu implementarea de la https://machinelearningmastery.com/classification-as-conditional-probability-and-the-naive-bayes-algorithm/ + analiza extinderii in cazul clasificării multi-etichetă - https://github.com/adhiraj/naivebayes). Biblio: NaiveBayes
7. Perceptron multinivel antrenat cu Backpropagation (implementarea unei rețele neuronale feedforward, cu unul sau doua nivele ascunse si antrenată cu backpropagation – testare pentru o problemă de clasificare sau regresie). Biblio: MLP+BP archive
8. Rețea neuronală de tip RBF – Radial Basis Function (implementarea unei rețele RBF și unui algoritm de învățare bazat pe estimarea separată a centrilor, parametrului funcției radiale și a ponderilor – testare pentru o problemă de regresie neliniară). Biblio: http://mccormickml.com/2013/08/15/radial-basis-function-network-rbfn-tutorial/ + RBF archive
9. Fuzzy c-means (e.g. implementarea variantei standard propusă de Bezdek și testarea pentru o problemă de grupare la alegere). Biblio: FuzzyCMeans archive
10. Algoritmi aglomerativi de grupare (e.g. implementarea variantei cu complete-linkage). Biblio: HierarchicalAlgorithms archive
11. DBSCAN (e.g. implementarea unei variante a algoritmului DBSCAN). Biblio: DBSCAN archive
12. Algoritmi de clustering bazați pe descompunerea matricii de similaritate (implementarea unui algoritm pentru spectral clustering). Biblio: SpectralClustering archive
13. Algoritmul Apriori (e.g. implementarea unei variante simple a algoritmului Apriori). Biblio: Apriori archive
14. [Bioinfo – proiect de echipă – 2 studenți] Studiu comparativ al algoritmilor de biclustering si ilustrare rezultate pentru date de tip microarray (analiza expresiei genice). Variante de algoritmi: Church&Cheng, Murali&Kasif, Bimax, Plaid Model, Spectral Biclustering. Biblio: Biclustering archive.

B. Proiecte orientate inspre date

- **seturi de date de la UCI Machine Learning Repository**
- **seturi de date de la https://www.kaggle.com**
- **un set de date la alegere asociat unei probleme reale (cu argumentarea alegerii)**

Proiectele de tip B constau in:

- **Un raport** (cca 6-8 pagini) in care:
  - este descris setul de date si problema care urmează a fi rezolvată,
  - este descrisă metoda/metodele utilizate (pe baza lucrărilor menţionate în descrierea setului din UCI Machine Learning Repository sau pe baza descrierii și/sau a implementarilor disponibile pe Kaggle)
  - sunt prezentate si discutate rezultatele
- Structura raportului:
  - Abstract (2-3 paragrafe): se descrie pe scurt care sunt obiectivele proiectului si care sunt principalele rezultate
  - Introducere: se descrie problema abordată și abordările existente (pe baza bibliografiei); se prezintă pe scurt ideea de rezolvare și modul în care e structurat raportul
  - Descrierea setului de date: se prezintă caracteristicile setului de date, inclusiv analiza statistică: nr inregistrari, nr atribute, tipurile atributelor, distributia valorilor atributelor (in functie de tip: histograma, medie, mediana, moda, abatere standard), vizualizarea datelor (daca e cazul), procentul de valori absente, gradul de debalansare (dacă e cazul)
  - Descrierea fluxului de prelucrări: se descriu prelucrările efectuate, cu motivarea selecției metodelor utilizate și cu detalii de implementare (se accentuează contribuțiile proprii)
  - Prezentarea rezultatelor, compararea cu cele prezentate în alte lucrări sau pe Kaggle (dacă este cazul)
  - Concluzii: prezentare succintă a observațiilor rezultate în urma implementării și a aplicării fluxului de prelucrări asupra setului de date; identificarea avantajelor și dezavantajelor (sau limitărilor) abordării propuse
- **Implementarea** fluxului de prelucrări (etapele de prelucrare aplicate asupra setului de date), valorile parametrilor si rezultatele obţinute aplicând un instrument de data mining (la alegere – poate fi R, Scikit-learn, Weka sau altă platformă). Alegerea modelelor/metodelor/etapelor de pre-procesare trebuie argumentată.

Exemple de tematici pentru proiecte de tip B:

15. Microblog PCU data set (http://archive.ics.uci.edu/ml/datasets/microblogPCU). **Scop**: indentificarea spammer-ilor (clasificare binară)
16. Absenteism at work (https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work). **Scop:** predictia numărului de ore de absență de la locul de muncă în funcție de diferite cause (medicale sau de altă natură) (regresie)
17. GPS trajectories (http://archive.ics.uci.edu/ml/datasets/GPS+Trajectories). **Scop**: identificarea grupurilor de traiectorii similare (clustering)
18. Blog feedback dataset (http://archive.ics.uci.edu/ml/datasets/BlogFeedback). **Scop**:predicţia numarului de comentarii in următoarele 24h (regresie)
19. Online news popularity (http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity). **Scop**: predicția numărului de partajări ale stirilor (regresie)
20. AAAI2013 Accepted Papers Dataset (http://archive.ics.uci.edu/ml/datasets/AAAI+2013+Accepted+Papers) . **Scop**: clustering bazat pe cuvinte cheie
21. Air pollution (https://www.kaggle.com/prakaa/air-quality-data-earlwood-nsw-australia) - analiza influenței factorilor de mediu (wind direction, wind speed, temperature, humidity) asupra concentrației diferitelor substanțe chimice din atmosferă (regresie, analiză serii temporale)
22. Cryptocurrency Historical Prices (https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory). **Scop:** construirea unui model de predicție
23. Intruder detection (https://www.kaggle.com/hassan06/nslkdd). **Scop:** construirea unui clasificator și identificarea atributelor relevante.
24. Google Stock Prediction (https://www.kaggle.com/shreenidhihipparagi/google-stock-prediction). **Scop:** construirea și analiza comparativă a unor modele pentru prognoză uni și multi-dimensională – prelucrarea seriilor temporale.
25. Credit Card Approval (https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data). **Scop:** clasificare/ identificare atribute relevante pentru acordarea unui imprumut
26. Walmart Sales Forecast (https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast/code). **Scop:** analiza influenta zile de sarbatoare asupra volumului vanzarilor
27. Online Payment Fraud Detection (https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset). **Scop:** clasificare (frauda / non-frauda)
28. CO2 emission (https://www.kaggle.com/datasets/rinichristy/2022-fuel-consumption-ratings). **Scop**: estimare emisie CO2 (regresie)
29. Parkinson disease classification (https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification). **Scop**: diagnostic boala Parkinson pe baza unor înregistrări vocale (clasificare binară)
30. Activity recognition (https://archive.ics.uci.edu/ml/datasets/Activity+recognition+using+wearable+physiological+measurements#). **Scop:** identificarea tipului de activitate desfășurată în funcție de semnalele înregistrate de la senzori plasați pe subiecți (clasificare multiplă)

31. **Bioinfo:** mice protein expression (https://www.kaggle.com/ruslankl/mice-protein-expression)
32. **Bioinfo**: genetic variant classification (https://www.kaggle.com/kevinarvai/clinvar-conflicting)
33. **Bioinfo:** Genome wide peak detection problem (http://archive.ics.uci.edu/ml/datasets/chipseq).
    Scop: clasificare binară