Proiecte Data Mining (2018-2019)

Temele sunt grupate în 2 categorii:

- Proiecte orientate către algoritmi
- Proiecte orientate către seturi de date

A. Proiecte orientate către algoritmi

Proiectele de tip A constau în:

- Un raport care in care sunt descrise particularitățile problemei abordate, 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 simple de date).
- Implementarea unui algoritm (limbajul de programare este la alegere – R, Python, Java).

Tematici pentru proiecte de tip A:

1. Algoritmi pentru selecția atributelor (implementarea algoritmului Relief sau a unui algoritm greedy de tip forward). Biblio: FeatureSelection folder
2. Algoritmi pentru discretizarea atributelor (implementarea algoritmului Holte 1R). Biblio: FeatureDiscretization folder
3. Algoritmi pentru construirea arborilor de decizie (implementarea algoritmului ID3 - http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm). Biblio: DecisionTree folder
4. Algoritmi de acoperire cu reguli (implementarea algoritmului PRISM). Biblio: CoveringAlgorithms folder
5. K-Nearest Neighbor (implementarea algoritmului kNN bazat pe distanta euclidiana). Biblio: kNN folder
6. Clasificator Naïve Bayes (implementarea unui algoritm pentru date cu atribute discrete). Biblio: NaiveBayes folder
7. Perceptron multinivel antrenat cu Backpropagation (implementarea unei rețele neuronale feedforward, cu un nivel ascuns si antrenată cu backpropagation – testare pentru o problemă de clasificare). Biblio: MLP+BP folder
8. Rețea neuronală de tip RBF – Radial Basis Function (implementarea unei rețele RBF și un algoritm de învățare – testare pentru o problemă de regresie neliniară). Biblio: http://mccormickml.com/2013/08/15/radial-basis-function-network-rbfn-tutorial/ + RBF folder
9. Fuzzy c-means (e.g. implementarea variantei standard propusă de Bezdek). Biblio: FuzzyCMeans folder
10. Algoritmi aglomerativi de grupare (e.g. implementarea variantei single-linkage variant). Biblio: HierarchicalAlgorithms folder
11. DBSCAN (e.g. implementarea unei variante a algoritmului DBSCAN). Biblio: DBSCAN folder

12. DENCLUE (e.g. implementarea unei variante de algoritm de grupare bazat pe funcții de densitate). Biblio: DENCLUE folder
13. Algoritmi de clustering bazați pe descompunerea matricii de similaritate (implementarea unui algoritm pentru spectral clustering). Biblio: SpectralClustering folder
14. Algoritmul Apriori (e.g. implementarea unei variante simple a algoritmului Apriori). Biblio: Apriori folder
15. [Bioinfo – proiect de echipă – 2-5 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 folder.

B. Proiecte orientate inspre date

- **seturi de date de la UCI Machine Learning Repository**
- **seturi de date de la https://www.kaggle.com**

Proiectele de tip B constau in:

- Un raport in care este descris setul de date, problema care urmează a fi rezolvată, metoda/metodele utilizate (pe baza lucrărilor menţionate în descrierea setului din UCI Machine Learning Repository sau pe baza descrierii și/sau a kernel-urilor de la Kaggle)
- Descrierea 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, Weka, Scikit-learn sau altă platformă). Alegerea modelelor/metodelor/etapelor de pre-procesare trebuie argumentată.

Exemple de tematici pentru proiecte de tip B:

16. DBWorld e-mails data set (http://archive.ics.uci.edu/ml/datasets/DBWorld+e-mails). **Scop**: clasificarea e-mailurilor in 2 categorii: anunturi de conferinte vs alte mesaje (clasificare binară)
17. Microblog PCU data set (http://archive.ics.uci.edu/ml/datasets/microblogPCU). **Scop**: indentificarea spammer-ilor (clasificare binară)
18. SMS Spam Collection (http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection). **Scop**: clasificarea SMS-urilor in spam/ham (clasificare binară)
19. Energy efficiency data set (http://archive.ics.uci.edu/ml/datasets/Energy+efficiency). **Scop**: predicția consumului de energie într-o clădire pe baza altor caracteristici (regresie)
20. GPS trajectories (http://archive.ics.uci.edu/ml/datasets/GPS+Trajectories). **Scop**: identificarea grupurilor de traiectorii similare (clustering)
21. Blog feedback dataset (http://archive.ics.uci.edu/ml/datasets/BlogFeedback). **Scop**:predicţia numarului de comentarii in următoarele 24h (regresie)
22. Online news popularity (http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity). **Scop**: predicția numărului de partajări ale stirilor (regresie)

23. Student performance dataset (http://archive.ics.uci.edu/ml/datasets/Student+Performance). **Scop**: predicția notei (la matematică, portugheză sau nota finală)
24. AAAI2013 Accepted Papers Dataset (http://archive.ics.uci.edu/ml/datasets/AAAI+2013+Accepted+Papers) . **Scop**: clustering bazat pe cuvinte cheie
25. News aggregator dataset (http://archive.ics.uci.edu/ml/datasets/News+Aggregator). **Scop**: gruparea știrilor pe categorii (clustering)
26. Bioinfo: Genome wide peak detection problem (http://archive.ics.uci.edu/ml/datasets/chipseq). **Scop:** clasificare binară
27. House price prediction (https://www.kaggle.com/c/house-prices-advanced-regression-techniques). **Scop:** estimarea preţului casei pornind de la caracteristici (regresie)
28. Credit card fraud detection (https://www.kaggle.com/dalpozz/creditcardfraud). **Scop:** predicţie fraudă pe baza caracteristicilor de utilizare (clasificare)
29. Gender recognition by voice (https://www.kaggle.com/primaryobjects/voicegender). **Scop:** identificarea sexului unei persone pe baza caracteristicilor vocii (clasificare)
30. Paper clustering (https://www.kaggle.com/benhamner/nips-2015-papers). **Scop:** gruparea lucrărilor pe baza similarităţii între conţinut (clustering)
31. **Kaggle competition**: CERN particle tracking (https://www.kaggle.com/c/trackml-particle-identification).
32. **Kaggle competition**: Data Science for good: DonorsChoose (https://www.kaggle.com/donorschoose/io)
33. **Kaggle competition**: Predict future sales (https://www.kaggle.com/c/competitive-data-science-predict-future-sales)
34. **Kaggle dataset:** Births in Poland (https://www.kaggle.com/mknorps/births/data)
35. **Kaggle dataset:** Air pollution (https://www.kaggle.com/prakaa/air-quality-data-earlwood-nsw-australia)
36. **Kaggle dataset:** 80 cereals - regresie (https://www.kaggle.com/crawford/80-cereals)
37. **Kaggle dataset:** Cryptocurrency Historical Prices (https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory)
38. **Kaggle dataset:** Chocolate Bar Ratings (https://www.kaggle.com/rtatman/chocolate-bar-ratings)
39. **Bioinfo – Kaggle dataset:** mice protein expression (https://www.kaggle.com/ruslankl/mice-protein-expression)
40. **Bioinfo – Kaggle dataset**: genetic variant classification (https://www.kaggle.com/kevinarvai/clinvar-conflicting)

Obs: poate fi utilizat orice alt set de date de la UCI Machine Learning, de la alte competiții Kaggle sau de la alte resurse.