Data Mining Projects. (2017-2018)

There are 2 types of projects:

- **Oriented towards algorithms**
- **Oriented towards data**

A. Projects oriented towards algorithms

Projects of type A consist of:

- A report describing the particularity of the addressed problem (classification, clustering, regression, association etc) and at least one of the algorithms which can solve that problem (based on the starting bibliography or on other related works) and presenting the results obtained by applying the implemented algorithm(s).
- An implementation from scratch of an algorithm (the programming language is at your choice – R, Python, Java, C etc).

Topics for projects of type A:

1. Algorithms for feature selection (e.g. implementation of Relief algorithm or of a greedy-like forward algorithm). Biblio: FeatureSelection folder
2. Algorithms for feature discretization (e.g. implementation of Holte 1R discretizer). Biblio: FeatureDiscretization folder
3. Algorithms for decision trees induction (e.g. implementation of the ID3 algorithm - http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm). Biblio: DecisionTree folder
4. Covering algorithms (e.g. implementation of PRISM algorithm). Biblio: CoveringAlgorithms folder
5. K-Nearest Neighbor (e.g. implementation of the classical kNN based on Euclidean distance). Biblio: kNN folder
6. Naïve Bayes classifier (e.g. implementation of a classifier for data with discrete attributes). Biblio: NaiveBayes folder
7. Multilayer perceptron and backpropagation (e.g. implementation of a one hidden-layer network trained with standard backpropagation tested for XOR). Biblio: MLP+BP folder
8. Fuzzy c-means (e.g. implementation of the standard version proposed by Bezdek). Biblio: FuzzyCMeans folder
9. Hierarchical agglomerative clustering algorithm (e.g. implementation of single-linkage variant). Biblio: HierarchicalAlgorithms folder
10. DBSCAN (e.g. implementation of a variant of the DBSCAN algorithm). Biblio: DBSCAN folder
11. DENCLUE (e.g. implementation of a clustering algorithm based on density functions). Biblio: DENCLUE folder
12. Apriori algorithm (e.g. implementation of a simple variant of Apriori algorithm). Biblio: Apriori folder

B. Projects oriented toward data

- datasets from UCI Machine Learning Repository)
- datasets from https://www.kaggle.com

Projects of type B consist of:

- A report describing the dataset, the problem to be solved and the used method (mainly based on the papers referred in the dataset description from UCI Machine Learning Repository).
- Description of the processing workflow (the processing steps applied to the dataset), the parameter values which have been used and the results obtained by applying a data mining tool (at your choice – it could be an R library, Weka, a Python library or another platform) to the dataset.

Topics for projects of type B:

13. DBWorld e-mails data set (http://archive.ics.uci.edu/ml/datasets/DBWorld+e-mails). **Aim**: classify the e-mails in two categories: conference announcements vs other messages (binary classification task)
14. Microblog PCU data set (http://archive.ics.uci.edu/ml/datasets/microblogPCU). **Aim**: identify spammers (binary classification task)
15. SMS Spam Collection (http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection). **Aim:** classification of SMS messages in spam/ham (binary classification task)
16. Energy efficiency data set (http://archive.ics.uci.edu/ml/datasets/Energy+efficiency). **Aim:** predict heating and cooling load in a building (based on a set of other characteristics)
17. GPS trajectories (http://archive.ics.uci.edu/ml/datasets/GPS+Trajectories). **Aim:** identify clusters of similar trajectories (clustering)
18. Blog feedback dataset (http://archive.ics.uci.edu/ml/datasets/BlogFeedback). **Aim:** prediction of the number of comments in the following 24 h) (regression)
19. Online news popularity (http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity). Aim: prediction of the number of shares of the news (regression)
20. Student performance dataset (http://archive.ics.uci.edu/ml/datasets/Student+Performance). **Aim:** prediction of one of the grades (math, Portuguese or final)
21. AAAI2013 Accepted Papers Dataset (http://archive.ics.uci.edu/ml/datasets/AAAI+2013+Accepted+Papers) .**Aim:** clustering based on keywords
22. News aggregator dataset (http://archive.ics.uci.edu/ml/datasets/News+Aggregator). **Aim:** group news by category
23. House price prediction (https://www.kaggle.com/c/house-prices-advanced-regression-techniques). **Aim:** house price estimation starting form some characteristics (regression)
24. Credit card fraud detection (https://www.kaggle.com/dalpozz/creditcardfraud). **Aim:** fraud prediction based on user actions (classification)

25. Gender recognition by voice (https://www.kaggle.com/primaryobjects/voicegender). **Aim:** prediction of the gender of a person based on the voice characteristics (classification)
26. Paper clustering (https://www.kaggle.com/benhamner/nips-2015-papers). **Aim:** grouping papers submitted to a conference based on the similarity between their content (clustering)
27. **Kaggle competition**: Data Science for good: DonorsChoose (https://www.kaggle.com/donorschoose/io)
28. **Kaggle competition**: Predict future sales (https://www.kaggle.com/c/competitive-data-science-predict-future-sales)
29. **Kaggle dataset:** Births in Poland (https://www.kaggle.com/mknorps/births/data)
30. **Kaggle dataset:** Air pollution (https://www.kaggle.com/prakaa/air-quality-data-earlwood-nsw-australia)
31. **Kaggle dataset:** 80 cereals - regresie (https://www.kaggle.com/crawford/80-cereals)
32. **Kaggle dataset:** Cryptocurrency Historical Prices (https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory)
33. **Kaggle dataset:** Chocolate Bar Ratings (https://www.kaggle.com/rtatman/chocolate-bar-ratings)

Remark: proposals of other problems or datasets which can be solved by using data mining methods are also accepted