Data Mining Projects (2024-2025)

There are three types of projects (only one project should be prepared – at your choice):

- **A: Oriented towards algorithms**
- **B: Oriented towards data**
- **C**: **Participation to a Data Mining/Machine Learning competition**
    - Kaggle ongoing competitions: https://www.kaggle.com/competitions
    - FedCSIS 2025 Challenge: Predicting Chess Puzzle Difficulty https://knowledgepit.ml/predicting-chess-puzzle-difficulty-2/

Remarks.

1. The starting bibliography is available on Classroom or at the links available for each topic.
2. The projects can be prepared individually or in teams of 2 students (with a clear statement of the contribution of each student).

A. Projects oriented towards algorithms

Projects of type A consist of:

- A **report** (around 6-8 pages) describing the particularity of the addressed problem (classification, clustering, regression, association, time series processing etc) and at least one of the algorithms which can solve that problem (based on the starting bibliography and/or on other related works) and presenting the results obtained by applying the implemented algorithm(s) on test data (the test data are at your choice).

  Structure of the report:
    - *Abstract* (2-3 paragraphs): briefly describe which the objectives of the project are and which are the main results (e.g. performance of your implementation with respect to existing implementations).
    - *Introduction*: describe the addressed problem and the existing approaches/algorithms (based on the bibliography – you can start from the papers provided for each topic, but you should also search for more recent publications); briefly presents the idea of the solution and how the report is structured
    - *Description of the method*: describe the method (based on the bibliography)
    - *Description of the implementation*: provide implementation details including reference to the source code
    - *Presentation of test results on a data set*: the datasets are at your but the results should include performance values (depending on the problem to be solved)
    - *Conclusions*: description of the main challenges encountered in the implementation and possible directions for improvement

- An **implementation from scratch of an algorithm** (the programming language is at your choice – Python, Java, C, R etc) – full access to the implementation and to the data used for training and testing should be provided (GitHub, Zenodo, Colab).

Topics for projects of type A:

1. Algorithms for **feature discretization** (e.g. implementation of Holte 1R discretizer). Biblio: FeatureDiscretization archive
2. Algorithms for **pre-processing imbalanced data sets** (e.g. SMOTE - https://arxiv.org/pdf/1106.1813.pdf , see also https://jmlr.org/papers/v18/16-365)
3. Algorithms for **data dimensionality reduction for visualization**: (i) tSNE (https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf) (ii) UMAP (https://arxiv.org/pdf/1802.03426.pdf)
4. Algorithms for **oblique 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 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. **Covering algorithms** (e.g. implementation of PRISM algorithm + Random PRISM). Comparative analysis wrt the implementation from https://github.com/dahvreinhart/Rule-Based-PRISM/blob/master/prism.py. Biblio: CoveringAlgorithms archive
6. **Naïve Bayes classifier** (implementation of an algorithm for **multiple classification** of data with **continuous attributes** – comparison with the implementation from https://machinelearningmastery.com/classification-as-conditional-probability-and-the-naive-bayes-algorithm/ and analysis of the extension for multi-label classification - https://github.com/adhiraj/naivebayes). Biblio: NaiveBayes archive
7. **Multilayer perceptron and backpropagation** (e.g. implementation of a one or two hidden-layers network trained with standard backpropagation and tested a classification or regression problem). Biblio: MLP+BP archive.
8. **Radial Basis Neural Network** (implementation of a RBF network and of a learning algorithm based on the separate estimation of the centres, width parameters and weights – testing for a nonlinear regression problem). Biblio: http://mccormickml.com/2013/08/15/radial-basis-function-network-rbfn-tutorial/ + RBF archive
9. **Fuzzy c-means** (e.g. implementation of the standard version proposed by Bezdek + testing for a clustering problem). Biblio: FuzzyCMeans archive
10. **Hierarchical agglomerative clustering algorithm** (e.g. implementation of complete-linkage variant). Biblio: HierarchicalAlgorithms archive
11. **DBSCAN** (e.g. implementation of a variant of the DBSCAN algorithm). Biblio: DBSCAN archive
12. **Affinity Propagation Clustering**. Biblio: AffinityPropagationClustering archive
13. **Apriori** algorithm (e.g. implementation of a simple variant of Apriori algorithm). Biblio: Apriori archive

B. Projects oriented toward data

- datasets from UCI Machine Learning Repository)
- datasets from https://www.kaggle.com
- a dataset at your choice, corresponding to a real problem (with a motivation of the choice).

Projects of type B consist of:

- A **report** (around 6-8 pages) containing the description of the dataset, of the processing steps and of the results

  Structure of the report
  - *Abstract* (2-3 paragraphs): briefly describe which the objectives of the project are and which are the main results.
  - *Introduction*: describe the addressed problem, the motivation of choosing the dataset(s) and summarizes existing results on the same dataset(s) with clear references to state of the art publications; briefly presents the idea of the solution and how the report is structured.
  - *Description of the dataset*: the characteristics of the dataset should be presented, including statistical analysis: no. of records, no. of attributes, types of attributes, distribution of attribute values (by type: histogram, mean, median, mode, standard deviation), data visualization (if applicable), percentage of missing values, degree of imbalance (if applicable).
  - *Description of the processing flow:* describe the processing carried out, giving reasons for the selection of the used methods and details of implementation (own contributions should be emphasized)
  - *Presentation of results*, comparison with those presented in other papers or on Kaggle (if applicable)
  - *Conclusions:* brief presentation of the observations resulting from the implementation and application of the processing flow on the dataset;
- **Implementation 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:

14. **Human Activity Recognition-1** (UCI HAR Dataset). Aim: (a) multiple classification (prediction of the activity type; (b) clustering (unsupervised identification of clusters).
15. **Human Activity recognition-2** (https://archive.ics.uci.edu/ml/datasets/Activity+recognition+using+wearable+physiological+measurements#). **Aim**: identification of the activity type based on recorded signals from sensors. (multiple classification)

16. **Census income dataset** (Adult - UCI Machine Learning Repository). Aim: Predict whether annual income of an individual exceeds $50K/yr based on census data (binary classification) + identify the most important input attributes (additional task: use explainability methods to provide explanations for the decision).
17. **Microblog PCU data set** (http://archive.ics.uci.edu/ml/datasets/microblogPCU). **Aim**: identify spammers (binary classification)
18. **Mushroom classification** (https://www.kaggle.com/datasets/vishalpnaik/mushroom-classification-edible-or-poisonous) **Aim:** Classify a mushroom as poisoneous or edible and identify the most important attribute(s) (binary classification + explainability tools)
19. **Credit Card Approval** (Credit Approval - UCI Machine Learning Repository). **Aim:** classification/ identification of relevant attributes for credit card approval (classification)
20. **Bioresponse** (OpenML) **Aim:** Predict a biological response of molecules from their chemical properties (binary classification, attribute selection)
21. **Hill-valley** (OpenML) **Aim:** classify a sequence of Y coordinates as defining a hill or a valley (binary classification)
22. **Electricity** (OpenML) **Aim:** predict if the price will increase or decrease (binary classification) – large dataset
23. **GPS trajectories** (http://archive.ics.uci.edu/ml/datasets/GPS+Trajectories). **Aim:** identify clusters of similar trajectories (clustering)
24. **AAAI2013 Accepted Papers Dataset** (http://archive.ics.uci.edu/ml/datasets/AAAI+2013+Accepted+Papers). **Aim:** clustering based on keywords (clustering) – it requires some text mining tools
25. **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) - it requires some text mining tools
26. **Grocery Basket Analysis** (InstaCart Online Grocery Basket Analysis Dataset) Aim: analyze shopping patterns and predict which previously purchased products will be in a user's next order (association rules, clustering).
27. **Engine Remaining Useful Life** (https://phm-datasets.s3.amazonaws.com/NASA/6.+Turbofan+Engine+Degradation+Simulation+Data+Set.zip) . Aim: predict the number of remaining operational cycles before failure for some engines (regression, based on multivariate timeseries).
28. **Blog feedback dataset** (http://archive.ics.uci.edu/ml/datasets/BlogFeedback). **Aim:** prediction of the number of comments in the following 24 h) (regression)
29. **Online news popularity** (http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity). Aim: prediction of the number of shares of the news (regression)
30. **Google Stock Prediction** (https://www.kaggle.com/shreenidhihipparagi/google-stock-prediction). **Aim:** time series forecasting (one and multi-dimensional).
31. **CO2 emission** (https://www.kaggle.com/datasets/rinichristy/2022-fuel-consumption-ratings). **Aim:** CO2 emission estimation + explanation of the influence of attribute values (regression + explainability tools)

32. **SmartGrid Energy Forecasting** (Individual Household Electric Power Consumption - UCI Machine Learning Repository). Aim: predict the consumption based in historical data (timeseries forecasting) – multivariate, some missing values
33. **Log files analysis** (GitHub - logpai/loghub: A large collection of system log datasets for AI-driven log analytics [ISSRE'23], 2008.06448). Aim: analyze the structure of logfiles, clustering or anomaly detection
34. **Phishing websites (**Phishing Websites - UCI Machine Learning Repository) Aim: predict if a website is phishing or not and identify the most predictive attributes (binary classification)
35. **OnlineShoppersIntention** (OpenML) **Aim:** predict online shoppers behavior (classification or regression)
36. **CERT Insider Threat (**CERT Insider threat) Aim:
37. **Cellular Localization Sites of Proteins (**Yeast - UCI Machine Learning Repository). Aim: predict localization site based on numerical attributes (multiple classification)
38. **Sleep Deprivation and Cognitive Performance (**OpenML). Aim: predict stress level (regression)
39. **Cybersecurity Attacks (**OpenML) Aim: estimate anomaly scores (regression)/ attack type (classification)
40. **Multispectral Data Analysis – Indian Pines (**OpenML). Aim: land-use types (8 categories – multiple classification)

Remark: Another dataset corresponding to a real-world problem can be chosen but the choice should be motivated.