DATA MINING 2 – Project Guidelines

Andrea Fedele
andrea.fedele@phd.unipi.it

a.a. 2024/2025

UNIVERSITÀ DI PISA

Project Modules

There are **4 modules** for this project:

- **Module 0:** Data Understanding & Preparation
- **Module 1:** Advanced Data-Preprocessing
  - Imbalanced Learning, Dimensionality Reduction, Anomaly Detection
- **Module 2:** Advanced ML & XAI
  - Logistic Regression, Support Vector Machines, Neural Networks, Ensamble Methods, Gradient Boosting, Rule-based Classifiers
- **Module 3:** Time Series Analysis
  - Time Series Similarity, Approximation, Motif, Shapelets, Classification, Clustering

Project Datasets

1. **Tabular** dataset [module 0, 1, 2]
   - Extended version of the DM1 dataset: ~150k records, additional info about titles (i.e., cast number, n. of companies that worked on it).
2. **Time series** dataset [module 0, 3]
   - Represents the domestic daily gross income of movies at the box office, starting from their release date.
     - ~1k time series, each associated with the movie's genre and rating (avg and category);

Module 0

Data Understanding and Preparation

- *Tabular Dataset*
  - Explore and analyse the dataset to understand its structure and key characteristics (describe them); Conduct data pre-processing (i.e., encoding categorical variables, feature scaling); Create new variables if needed. You may take inspiration from existing notebooks but should define your own research prospective.
- *Time-series* Dataset
  - Perform exploratory analysis and describe your findings; pre-process the dataset to prepare it for tasks such as time-series clustering, motif/anomaly detection, and classification. If the dataset is too large for these tasks, consider using approximation (i.e., SAX, PAA).

Module 1

Outliers

- Identify the top 1% outliers: adopt *at least three* different methods from *different* families (i.e., density-based, angle-based…) and compare the results.
- Visualize the outliers in a 2 or 3d scatter plot using at least one dimensionality reduction technique.
- Deal with the outliers in a way you see fit, e.g. by removing them from the dataset or by treating the anomalous variables as missing values and employing replacement techniques. In this second case, you should check that the outliers are not outliers anymore. Justify your choices in every step.

Module 1

Imbalanced Learning

- Define one simple unbalanced classification tasks and solve it either with a Decision Tree or KNN.
- If the dataset is already unbalanced, leave it as it is; otherwise, modify it to create an imbalanced version (e.g., 96% - 4%, for binary classification).
- Solve the classification task using the Decision Tree or KNN by adopting *at least 2* techniques of imbalanced learning (Undersampling, Oversampling).
- Analyse and discuss the impact of these techniques, focusing on class distribution and detailed performance metrics.

Module 2

Advanced Classification

- Solve a multi-class classification task (i.e., predict rating) using the classification methods covered in the course: Logistic Regression, Support Vector Machines, Neural Networks, Ensemble Methods, Gradient Boosting Machines.
  - The target variable must have more than 5 classes at this stage. You are encouraged to explore *additional target variable* (after rating) that might provide meaningful insights, guided by previous analyses or new hypotheses.
- **Always** perform hyper-parameter tuning phases and justify your choices. Explain which parameters you tested, why you selected them, and which ones performed best.
- Evaluate **each** classifier with the techniques presented in DM1, including accuracy, precision, recall, F1-score, and ROC curve (or lift, precision-recall curves).
- Besides the numerical evaluation draw your conclusions about the various classifiers (e.g. for Neural Networks: what are the parameter sets or the convergence criteria which avoid overfitting? For Ensemble classifiers how the number of base models impact the classification performance? What is revealing the feature importance of Random Forests? Is there a specific class that performs bad in each classifier but one and why?)

Module 2

Advanced Regression

- Define a *multiple* regression task, i.e., using more than one input feature, and solve it using 2 advanced regression approaches (not linear).
- Compare and evaluate the approaches using appropriate metrics. Discuss the result.

Explainability

Try to use one or more explanation methods (e.g., TREPAN, LIME, LORE, SHAP, Counterfactual Explainers, etc.) to illustrate the reasons for the classification in one of the steps of the previous tasks.

Module 3

Motifs/Discords

Analyse the dataset for finding motifs and/or anomalies. Visualize and discuss them and their relationship with shapelets.

Clustering

- Use at least two clustering algorithm on time series using an appropriate distance.
- Analyse the clusters and highlight similarities and differences and visualize the clusters using at least 2 dimensionality reduction techniques. Discuss the result.

Module 3

Classification

Define one (or more) classification task and solve it using:

- KNN with at least two distances
    1. Euclidean or Manhattan
    2. DTW
- Shapelets: Analyse the shapelets retrieved
- At least one other method (rocket, muse, cnn, rnn etc)

Sequential Pattern Mining (optional)

Discretize the time-series to run sequential pattern mining (e.g., identify frequent pattern or trends within the time series).