DATA MINING – Project Guidelines

Andrea Fedele
andrea.fedele@phd.unipi.it

a.a. 2024/2025

UNIVERSITÀ DI PISA

General Information

Who?

- Groups of min 2, max 3 people (ideally, heterogeneously distributed… e.g., 2 DS + 1 InfoUma);
- Insert your group at the following link.

What?

- A Project:
  - Data Understanding & Preparation;
  - Clustering;
  - Classification;
  - Pattern Mining.
- A Report:
  - max 20 pages of text, including tables and figures;

only the report will be evaluated!

General Information

When?

There are two options:

[1 week before the oral exam]: complete project → NO BONUS POINTS

OR

[15/11/2024]: mid-draft (Data Understanding & Preparation + Clustering) → 0.5 points

[31/12/2024]: complete project → 0.5 points

So delivering the mid-draft and the complete project within the deadlines → 1 point

Deadlines can change, so you should refer to the main page of the course;

Where?

- send the report to **BOTH** andrea.fedele@phd.unipi.it & riccardo.guidotti@unipi.it
  - Object: [DM1 Project 24/25]
  - Report title: Project_Surname1_Surname2….pdf

How?

- Code language: Python (suggested);
- Report: written in LaTex (Overleaf) (suggested), or Office/OpenOffice…;

The Project

Dataset: The IMDb Dataset contains data about movies, TV shows, and other forms of visual entertainment, along with their ratings, which is generated by the internet community. Each record includes key information such as the original title, release year, runtime, and the number of user votes. Additionally, the dataset provides insights into critical aspects like awards, nominations, and user reviews, as well as statistical ratings, from the best and worst ratings to the total number of critic reviews. It also includes metadata like country of origin, number of images and videos, and the title's genre. The dataset is updated as of September 1, 2024.

Links to the dataset will be provided on the main page of the course!

The Report

Structure

- Title page and index are not counted for the 20 pages limit;
- Only PDF are allowed, no doc, jupyter notebooks, python code;
- It is better to use font size higher than 9pt;
- Multiple columns are allowed;

Content

- You must justify every choice (from the variables management to the parameters you tune);
- Discuss every result; even if some of them don’t convince you, be fair and try to discuss the possible limitations (they can be imputed to the dataset, to an algorithm that does not fit with the dataset, etc…);
- Plots and tables without any comment are useless;
- Nice and readable plots make your analysis more understandable ;
- Even if you find a top configuration for your algorithm (e.g., k-means, k=5) you must list which are the different parameters you tested and justify your choice;

The Tasks

- Data Understanding & Preparation;
- Clustering;
- Classification;
- Pattern Mining;

The next slides provide several analytical suggestions, but:

- You are allowed to organize the content of the complete project as you prefer;
- You are allowed to identify the classification task as you prefer;
- You are allowed to explore tools and methodologies not introduced during the lectures (e.g., feature selection methods, new plots, algorithms), but it is suggested to write me an email before;

Data Understanding & Preparation (30 pts)

- Data Semantics
    - Introduce the variables with their meaning and characteristics;
- Distribution of the variables and statistics
    - Explore (single, pairs of…) variables quantitatively (e.g., statistics, distributions);
- Assessing data quality
    - Are present errors, outliers, missing values, semantic inconsistencies, etc?
- Variable transformations
    - Is it better to use for further modules transformed variables (e.g., log-transformated)?
- Pairwise correlations and eventual elimination of variables
    - Matrix correlation (analyse high correlated variables);

Clustering (30 pts)

- Analysis by centroid-based methods
  - K-Means (mandatory), Bisecting K-Means (optional), X-Means (optional);
  - Choice the attributes, identify the best value of k, discuss the clusters.
- Analysis by density-based clustering
  - DBSCAN (mandatory), OPTICS (optional);
  - Choice the attributes, identify the best parameter configuration, discuss clusters.
- Analysis by hierarchical clustering
  - Choice the attributes, the distance function, analyse several dendrograms.
- Final discussion
  - Which is the *best* algorithm? Remember that *best* is studied w.r.t. several aggregate statistics, cluster distributions and w.r.t. the typology of algorithm used for that particular dataset;

Classification (30 pts)

- Classification of **at least 1** target variable of your choice:
  - by Decision Trees;
  - by KNN;
  - by Naive Bayes.

You should discuss the choice of the attributes and identify the best parameter configurations (e.g. gain criterion for trees, best k for KNN etc.).

- Discussion
  - Evaluate the quantitative performance of the algorithms w.r.t. confusion matrix, accuracy, precision, recall, F1, ROC curve
  - Discuss some insight (e.g. try to interpret the tree(s))
  - Which is the *best* algorithm? *Best* can be studied w.r.t. the performance evaluation or other preferred point of view;

Pattern Mining (and Regression) (20 + 10 pts)

- Frequent Pattern extraction
  - Using different values of support, etc;
- Discuss Frequent Pattern
  - Including qualitative and quantitative analysis, e.g., how the number of patterns w.r.t k min_sup changes;
- Association Rules extraction
  - Using different values of confidence, etc;
- Discuss Association rules
  - Including qualitative and quantitative analysis, e.g., how the number of rules w.r.t k min_conf changes, histograms of rules' confidence and lift;
- Exploit the most useful extracted rules
  - E.g., use them to replace missing values or to predict the target variable;

- Regression: univariate and multivariate regression:
  - Choosing 2 or more continous variables and using different regressors (linear, ridge, lasso, Decision Tree, KNN)

Bonus & Other

- You can get 3 additional extra points in the final mark w.r.t. the following criteria:
  - Innovation (0.5 pts)
  - Experimentation (0.5 pts)
  - Performance (0.5 pts)
  - Appearance, Summary, Organization (0.5 pts)
  - Mid-draft and complete project within time (0,5 + 0,5 = 1 point)

- **Project Mark**: average of the previous modules + 3 bonus points;