Project on Data Analysis and Mining
part I

Main Goal

The aim of this project is to analytically explore a real-world data set.

You are encouraged to explore a meaningful dataset of your choice and, in accordance with the obtained findings, adequately apply the proposed methods and techniques for the available data.

You should analyse and critically interpret your results, hypothesize causes for the eventual limited efficacy of the applied methods, and identify opportunities to improve the Data Analysis and Mining process.

Students may choose the mining programming language to apply, among *Python*, *R*, *MatLab*.

Obs.: Due to the limited time available to this project the proposed tasks have to be very focused.

1. Define your Project

Select a data set of your preference from a Machine Learning / Data Science data repository satisfying the following conditions: (i) number of entities not less than 100; (ii) number of features/variables not less than 10; (iii) **all features numerical**.

Hint: To make computations and explanations constructive do not take too big data sets.

Data repositories

- UCI ML repository
- Kaggle datasets
- Machine Learning and AI Datasets, Carnegie Mellon University, https://guides.library.cmu.edu/c.php?g=844845&p=6191907
- StatLib Datasets Archive, Carnegie Mellon, http://lib.stat.cmu.edu/datasets/

Start writing a report which must contain:

- Project title page
- Your motivation explaining the choice of the data set
- Descriptive information of the data set: description of features; number of entities; source address
- Summarize examples of problems to be addressed from the knowledge domain and point out which one you propose to explore with your data and why.

Consider the recommendations on writing your report at the end of this document.

2. Exploratory Data Analysis

Inspect your data focused on the problem you want to address[1].
Apply Python libraries like: pandas, matplotlib, seaborn, plotly, numpy, scipy.

a) Data Summarization & Statistics
  - *Suggestion*: **df.info(), df.describe()** for quick insights
  - Checking distributions using **histograms & box plots**
  - Understanding central tendency & dispersion
b) Data Visualization
  - Univariate analysis: **histograms, bar charts, box plots**
  - Bivariate analysis: **scatter plots, pair plots, heatmaps**
  - Correlation analysis using **seaborn.heatmap()**

---

[1] **Exploratory Data Analysis Tutorials**
https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/
https://www.analyticsvidhya.com/blog/2021/08/how-to-perform-exploratory-data-analysis-a-guide-for-beginners/

3. Fuzzy Clustering with Anomalous Patterns

a) Apply two types of normalization to your data: by *range* and *z-score*.

b) Study the Fuzzy *c*-Means (FCM) program in the software package of your choice[2].

Apply the program to your dataset at the same hyperparameter *c* with random seeds. Do this raging the number of clusters $c = c\_min, ..., c\_max$. Plot the FCM clustering criterion (i.e. the *cost function*) as a function of *c*. Analyse the graphic and comment if any number of clusters better fits your data than the others.
Hint: If, at a given c, the fuzzy *c*-means converges to the same result at any initialization, then it is likely that parameter *c* is correct.

c) Study the (Iterative) Anomalous Pattern (IAP) clustering algorithm. Test the implementation with the benchmark data sets provided to you.

d) Take the Anomalous Clustering as the initialization algorithm to the fuzzy *c*-means and apply the Anomalous Patterns FCM to your data set. Discuss the option taken for setting its stop condition. Present and visualize the found fuzzy partitions of AP-FCM taking advantage of the PCA visualization (check PCA tutorial).

e) Discuss the results obtained by Anomalous Patterns FCM for your data case respecting the following: (i) location of the initial prototypes; (ii) number of clusters.

f) Apply, at least, three validation indices, like the Adjust Rand Index (ARI), Xie-Beni index, Silhouette index to access the quality of the fuzzy *c*-partitions obtained in a). Compare these results with the one of Anomalous Patterns FCM getting in c).

g) Make interpretation of the found clusters (***after defuzzification***) for your data, as discussed in classes.

4. Principal Component Analysis

Select a subset of 3 to 6 features related to the same aspect of the phenomenon your data set relates to. Explain your choice.

a) Visualize the data over these features in 2D/3D PC plane using the two types of normalization as before

b) Choose between conventional PCA or SVD for the visualization. Make a comment whether one of the normalizations is better and why.

c) At these visualizations, use a distinct shape/colour for data points representing a pre-specified group of objects. Comment on the choice of your groups.

d) Calculate and make a graphical presentation of the "quality" of the PC projection of your data. Discuss your results.

---

[2] **Suggestions**
Python tollboxes: https://github.com/scikit-fuzzy/scikit-fuzzy ; https://pypi.org/project/scikit-fuzzy/
R toolboxes: https://cran.r-project.org/web/packages/ppclust/vignettes/fcm.html ; https://rpubs.com/rahulSaha/Fuzzy-CMeansClustering
Toolbox for internal validity indices: https://pypi.org/project/Clusters-Features/ ; https://github.com/clslabMSU/clustGUI

Recommendations for the Report Writing

As Part of the team

- **Title** to be chosen depending on the meaning of your data set.
- **Introduction** In this section you should discuss the "What's" and "Why's" of your data set: (1) explanation of the choice of the data set; (2) Information of the data set: a) description of features (feature names, units of measurement, etc.); b) number of entities; c) source address; d) examples of problems that can be addressed using the data. Please be aware that your data set should include no missing values, as the current exercise involves only methods applicable to data with no missing entries.
  Finalise the Introduction making reference to the data analysis methods and algorithms that you are going to explore in the report.
- **Experimental study** You may answer question by question. When adequate, you should state the setting of the experiment. A discussion with interpretation of the results is mandatory using graphs, tables, and summary statistics.
- **Conclusion** Briefly summarize and critically interpret the obtained results of the analysis, pointing out successful aspects, and hypothesize causes for the eventual limited efficacy of the applied methods, and identify opportunities to improve the data analysis and mining process.
  List opportunities for future research that seem promising to you but for which you did not find the time applying during the course.
- **Appendix**: here you have to list the Heads of the developed functions, properly commented. Provide a Script to run your program as well as a 'readme' file, if needed.
- **Bibliography**

Note: You should produce several figures in your report. Remember that although you generate a graph in colour, the lines may not be identifiable when you print the document in greyscale. Make sure to use line markers or different types of lines to distinguish the data or, in alternative, print your report in colour. Discuss each figure and table within the text of your paper and explicitly identify the item you are discussing: "Figure 1 displays …".

Individual REPORT (supplement to be submitted by each student)

You must also submit a brief individual report (no more than one page), containing the following:

- Describe the parts of the project you worked on (which machine learning methods you applied, which pre-processing steps you performed on the data, which parts of the term paper you wrote, who you worked with on what parts, etc.) and what parts of the project your teammates worked on.
- What you learned from the project.

Good work!