Introduction to Data Science - Assignment 2

1 Introduction

1.1 Practice goals

- General: Medical data sources, Data processing and Visualization
- Technical: usage of DS libraries (Pandas, NumPy, scikit-learn, SciPy, etc)
- Algorithmic: Dimensionality reduction, Clustering

1.2 Background:

You have seen in class different methods for data visualization and data clustering . In this assignment, you will use clustering to improve visualization and reveal new insights. You will use clinical data from the MIMIC-II publicly-accessible critical care database for a case study on indwelling arterial catheters (https://physionet.org/content/mimic2-iaccd/1.0/). The dataset was created to investigate the effectiveness of indwelling arterial catheters in hemodynamically stable patients with respiratory failure for mortality outcomes for 1,776 patients. Our main outcome of interest in the dataset is whether or not the patient died within 28 days from the first day of ICU stay ('day_28_flg') and your goal is to differentiate between patients in a way hopefully predictive of that target.

2 Instructions and Deliverables

- Load data from "full_cohort_data.csv", explore it and make sure you understand every feature, both technically (type, values span, missing values, etc.) and the actual meaning of it - domain awareness drives better design decisions.
- Pre-process the data:
  - Change dtypes (e.g. object-¿category) if needed.
  - Fill missing/corrupt values (explain how do you avoid introducing bias).
  - Remove/mask columns - redundant/non-informative/noisy and (!) the target..
  - Try various normalizations of numeric features (by mean&std, scale to [0,1], etc.)
- Do the following for each one of the **dimensionality reduction** methods: (a) PCA (b) TSNE
  - Transform the features table, reducing the dimension to 2.
  - Create a scatter plot of your samples after the dimensionality reduction, coloring the samples according to the target value.
  - Repeat with modifications (to preprocessing, method hyperparameters) to achieve and improve target-predictive mapping of the sample space.
  - **Clustering:** use KMeans algorithm on the 2D projections and find the best partition of the dataset (what's the optimal number of clusters?).
  - Create a scatter plot with the projections, coloring the points according to the cluster labels received from the KMeans clustering.
- Discussion: comment on the 4 final graphs you obtained (and the previous attempts if diff is insightful). What worked better (PCA/TSNE? which preproc?) and why.
- Bonus: compute potential prediction accuracy of your best clustering.