This document provides a brief sketch on how to interpret and explain clusters derived from datasets, focusing on identifying defining features of each cluster through various data treatment approaches and, most importantly, visualization techniques.

- **Summarizing and testing features:** Features within clusters should be summarized using means, standard deviations, and medians for continuous variables, and counts and proportions for binary variables.

  All these statistics may be visualized (boxplots for continuous variables, bar plots for binary variables)

  Statistical tests such as ANOVA or Kruskal-Wallis for continuous data and Chi-square or Fisher tests for binary data, along with effect sizes and p-value adjustments, help determine significant differences between clusters.

- **Multivariate analysis and interpretation:** Training explainable machine learning models like Random Forests or logistic regression to predict cluster labels allows assessment of feature importance through techniques such as permutation importance (for Random Forest) or analysis (for Logistic Regression).

  Other explanations may be produced by SHAP values.

  Combining univariate and multivariate evidence supports meaningful interpretation of clusters, exemplified by characterizing clusters with specific clinical traits.

  Feature importance, odds ratios, and SHAP values may be visualized (bar plots for feature importance, forest plots for logistic regression, waterfall plots for SHAP).

- **Visualization and outputs:**

  Effective visualization methods for the identified partitions include heatmaps of feature means or reordered sample features and dimensionality reduction scatter plots colored by cluster assignment.

Toy datasets

Name	Description	Key features / advantages	Limitations / considerations
Heart Failure Clinical Records (UCI)	~300 patients with features like age, ejection fraction, serum creatinine, etc., plus whether they died/survived. (archive-beta.ics.uci.edu)	Relatively small & clean; has both continuous and categorical / binary-like features; good for teaching.	Small sample size; not many binary "drug/disease" features; outcome label might bias interpretation; missing data possible.
ILPD (Indian Liver Patient Dataset)	~584 patients; features are biochemical / continuous + a binary "disease/no-disease" target. (uci-ics-mlr-prod.aws.uci.edu)	Good mix of continuous; plenty of variation; UCI dataset (easy to load).	Fewer binary features; doesn't include many comorbidity / drug features.
Icentia11k ECG dataset (PhysioNet)	Long ECG recordings from ~11,000 patients, annotated beats etc. (PhysioNet)	Large; some strong signal annotations; suitable if you want to include signal-processing or extract features (heart rate, beat type) and cluster by those.	Raw time series; more work needed to preprocess / extract "features"; less immediate in terms of many "binary drug/disease" features.
PTB-XL ECG dataset (PhysioNet)	Clinical ECG records (12-lead, etc.) plus diagnostic labels. (PhysioNet)	Large; can extract features; has diagnostic / pathology categorical labels; suitable for clustering if you derive features.	Similar to Icentia: signal data → need feature extraction; not many "standard" binary comorbidity or drug columns.
SUPPORT2	From UCI: ~9,000 critically ill patients across multiple US centres; many physiological and demographic variables. (UCI Machine Learning Repository)	Larger; multiple continuous + categorical features; clinical severity etc.; good for clustering / survival modelling.	More complex; may have missing data; fewer "drug usage" features; also ethical/legal constraints depending on access/licensing.
UCI Heart Disease / Cleveland database	Classic dataset with mixed features: age, sex, chest pain type, blood sugar, ECG, etc. (UCI Machine Learning Repository)	Many people are familiar with it; has categorical, binary, continuous; small enough to teach.	The number of features is limited; drug/disease binary features not rich; target variable is disease/no disease.

Handout: Explaining Clusters - example

1. Goal

You clustered your dataset (e.g. patient samples). Now: *which features define each cluster?*

2. Workflow

1. **Summarize features by cluster**
    - Continuous: mean, std, median.
    - Binary: counts, proportions.
2. **Test differences**
    - Continuous → ANOVA/Kruskal-Wallis + Cohen's d effect size.
    - Binary → Chi-square/Fisher + difference in proportions.
    - Adjust p-values (Benjamini–Hochberg).
3. **Multivariate check**
    - Train a Random Forest (or logistic regression) to predict cluster labels.
    - Look at feature importance (permutation / SHAP).
4. **Interpret clusters**
    - Combine univariate + multivariate evidence.
    - Example: *Cluster 2 = high BMI, high diabetes prevalence, older patients.*
5. **Visualize**
    - Heatmap of feature means per cluster (or Heatmap of sample features, after reordering samples by cluster).
    - Boxplots for key continuous vars
    - Bar plots for binary vars.
    - UMAP/t-SNE scatter colored by features.

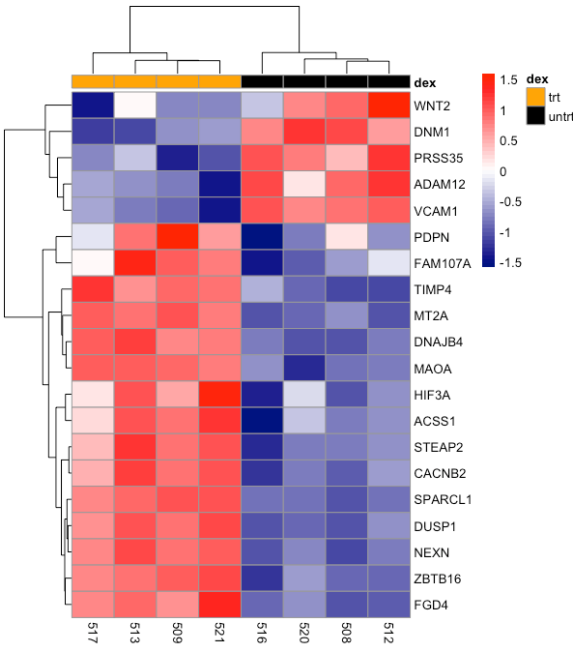

Figure 1: example of heatmap with sample features. Samples have been sorted by cluster (trt - yellow, untr1 - black)

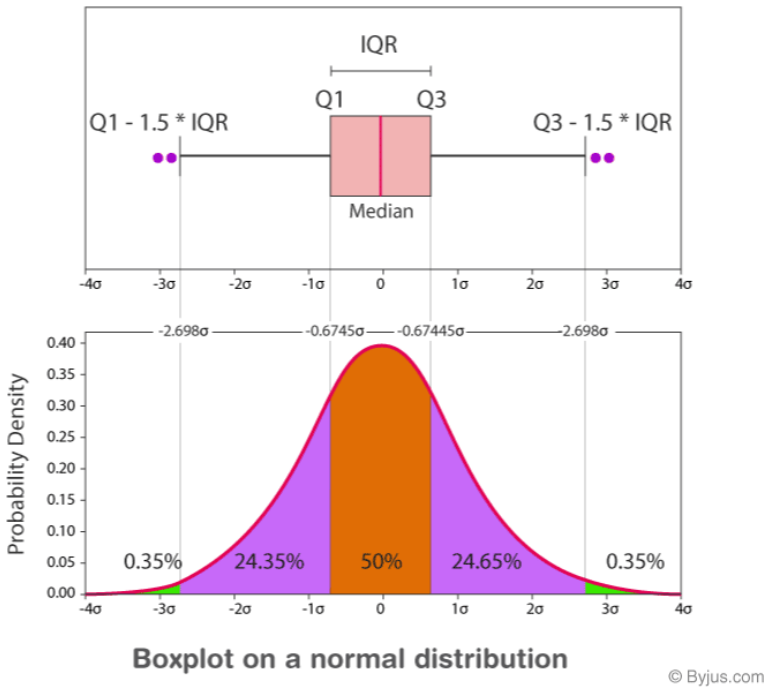

Figure 2: boxplots

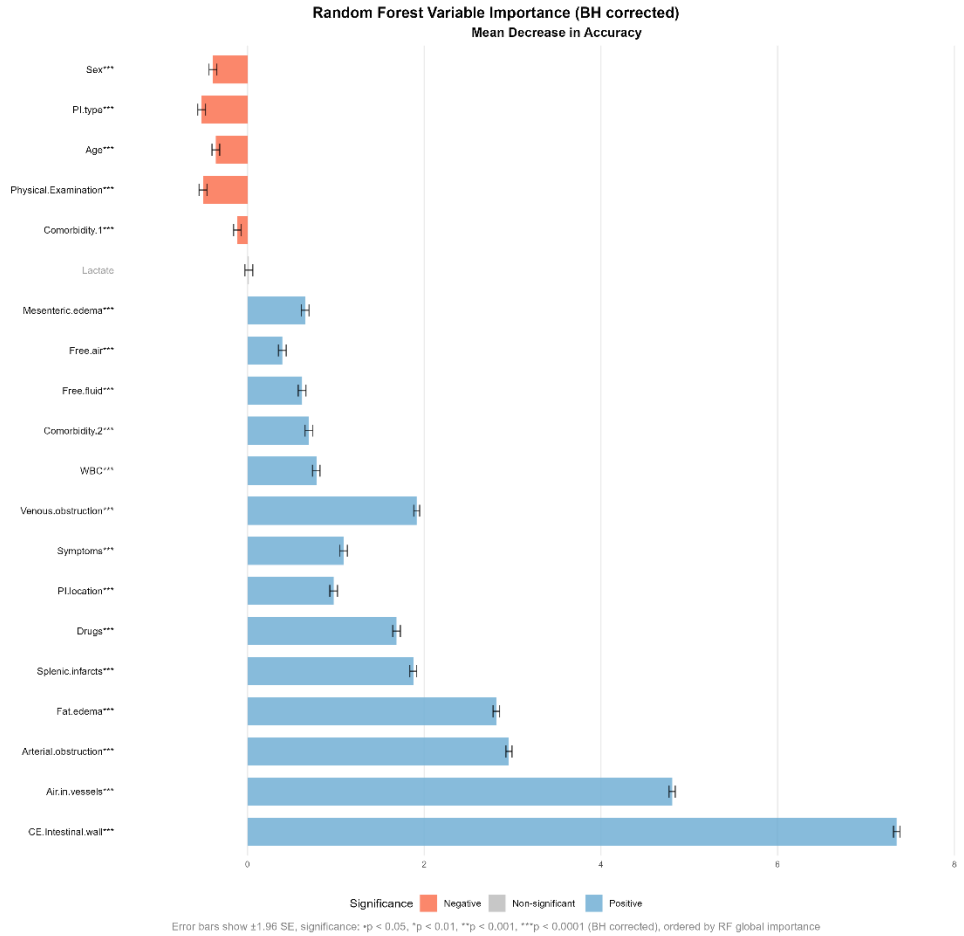

Figure 3: feature permutation importance as computed by RF

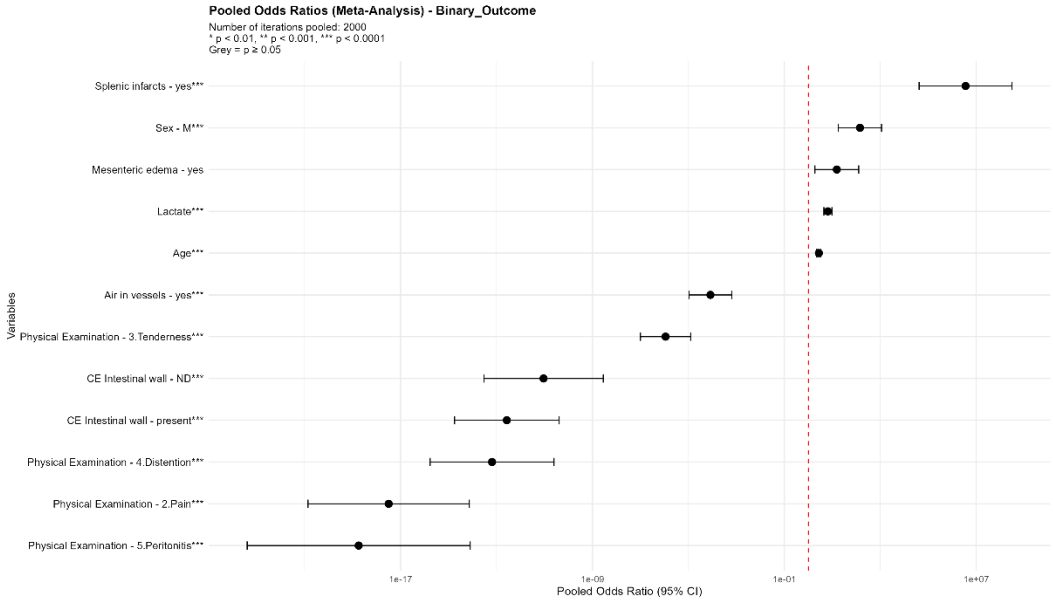

Figure 4: forest plots visualizing odds ratios of variables

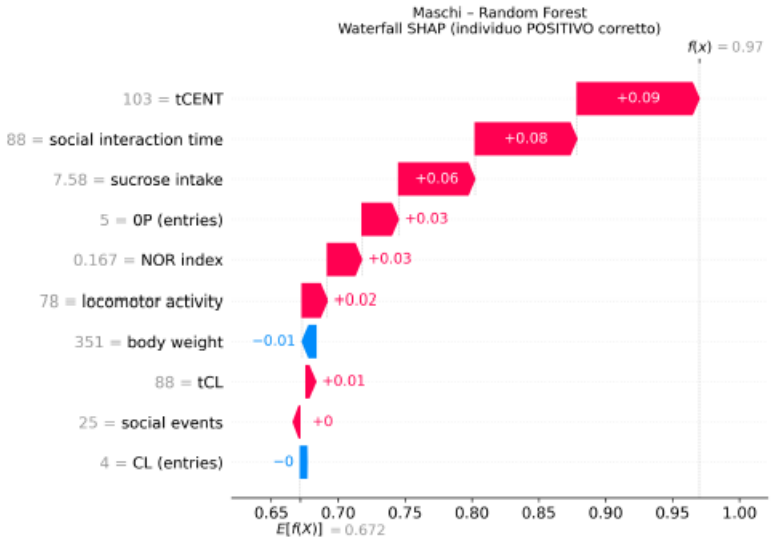

Figure 5: waterfall plots for visualizing SHAP values

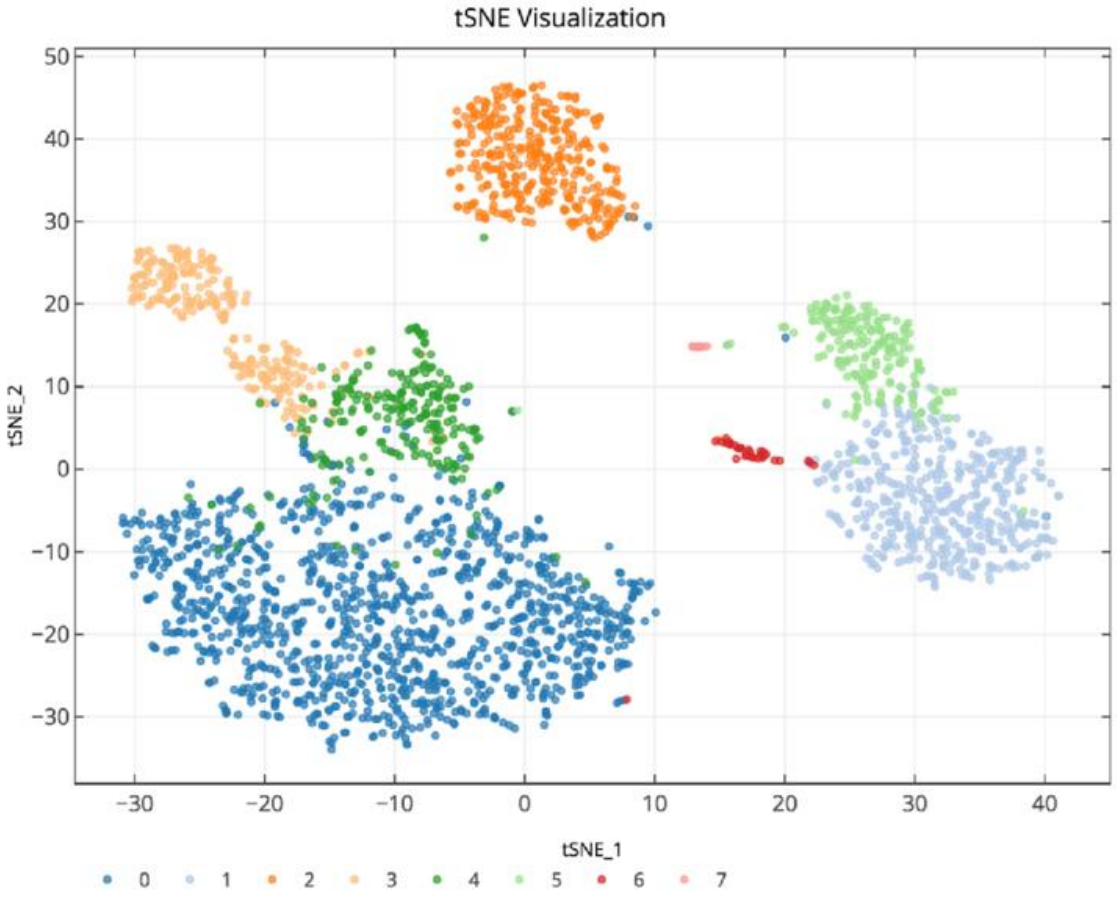

Figure 6: tsne visualization of clusters