# Assignment 3- Scatterplot, dimension reduction and clustering

You are strongly encouraged to work on this assignment with one or two partners. Please ensure only one submission per group: upload the PDF report to Gradescope and the code (or Jupyter notebook) to D2L. The dataset CSV file for Assignment 3 can be found on D2L.

This dataset is based on real-world data, not synthetic, so don't anticipate very clear-cut clusters or trends. There's no single correct answer to any question. Aim to thoroughly explain your analysis and the reasoning behind your choices as effectively as possible.

In [3]:

```
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
```

In this assignement, we focus on `heart failure clinical records dataset` ,from UCI Machine Learning repository. It includes the medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features and you can access more information about this dataset here

In [116…

```
heart=pd.read_csv('heart_failure_clinical_records_dataset.csv')
```

In [117…

```
heart.head()
```

Out[117]:

| | age | anaemia | creatinine_phosphokinase | diabetes | ejection_fraction | high_blood_pressure | platelets | serum_creatinine | serum_sodium | sex | sm |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 75.0 | 0 | 582 | 0 | 20 | 1 | 265000.00 | 1.9 | 130 | 1 | |
| **1** | 55.0 | 0 | 7861 | 0 | 38 | 0 | 263358.03 | 1.1 | 136 | 1 | |
| **2** | 65.0 | 0 | 146 | 0 | 20 | 0 | 162000.00 | 1.3 | 129 | 1 | |
| **3** | 50.0 | 1 | 111 | 0 | 20 | 0 | 210000.00 | 1.9 | 137 | 1 | |
| **4** | 65.0 | 1 | 160 | 1 | 20 | 0 | 327000.00 | 2.7 | 116 | 0 | |

## Problem 1- Scatterplots (50 points)

### (a) 10 points-

Get to know the dataset by generating summary statistics for the numerical variables, such as mean, median, and standard deviation. Address any missing values and outliers. Which variables do you think might be crucial in determining patient survival? Which variables seem to be most closely correlated.

### (b) 20 points

Create pair plot and correlation matrix among the different numerical variables. Comment on your findings.

### (c) 20 points

Explore different pairs of variables. Create two separate scatterplots and color points based on Death variable, `DEATH_EVENT` or other categorical variables. Do you see any ditinctive groups.

## Problem 2- Dimension Reduction (50 points)

### (a) 15 points

Perform PCA on the **standardized numerical variables**. Plot cumulative explained variance. How many principal components are needed to capture 80% of the variance? Comment on your results.

### (b) 20 points

Construct a biplot that displays the data points and the loadings of each original feature in the PC1-PC2 space. Label the loading vectors (arrows). Share your observations. Discuss which type of patient is best represented in each quadrant of the plot.

### (c) 15 points

Color-code the data points based on DEATH_EVENT or other categorical variables within the dataset. Evaluate whether dimension reduction has facilitated a clearer distinction between the various patient groups.

# Problem 3- Clustering (100 points)

## (a) 15 points

Apply K-means clustering to the PCA-transformed data (referencing the number of PCA components selected in 2a). Experiment with a variety of K values and graph the resulting Inertias and Average Silhouette Coefficients for each K. Analyze the graphs to determine the optimal number of clusters. Which cluster count do you prefer when considering both metrics?

## (b) 20 points

Interpret the characteristics of typical patients in each cluster for the chosen K value from the previous question. This involves reversing the PCA and standard scaler transformations applied to the centroid vectors. Share your insights on the findings. Are there any notable differences between the clusters?

## (c) 20 points

Using only the first two principal components (PCs), generate two scatterplot subplots. In the first subplot, color-code the data points according to the predicted cluster assignments. In the second subplot, use color-coding to represent 'death' or other categorical variables. Examine both plots and share any significant findings or patterns that become apparent from this comparative analysis.

## (d) 15 points

Construct dendrograms for various linkage methods (complete, single, ward). Assess which linkage method appears most suitable. Determine the most appropriate number of clusters based on the dendrograms.

## (e) 15 points

Redo part a of problem 3, but use hierarchical clustering (AgglomerativeClustering) in place of k-means. Use the the linkage type you chose in previous question. Test different numbers of clusters and plot the resulting Inertias and Average Silhouette Coefficients for each cluster count. Examine these plots to identify the ideal number of clusters. Based on both metrics, which number of clusters seems most appropriate? Compare the results with those obtained from part a; did the choice of clustering method lead to a different outcome?

## (f) 15 points

Redo part c and compare the results. Any significant difference?

```
In [ ]:
```