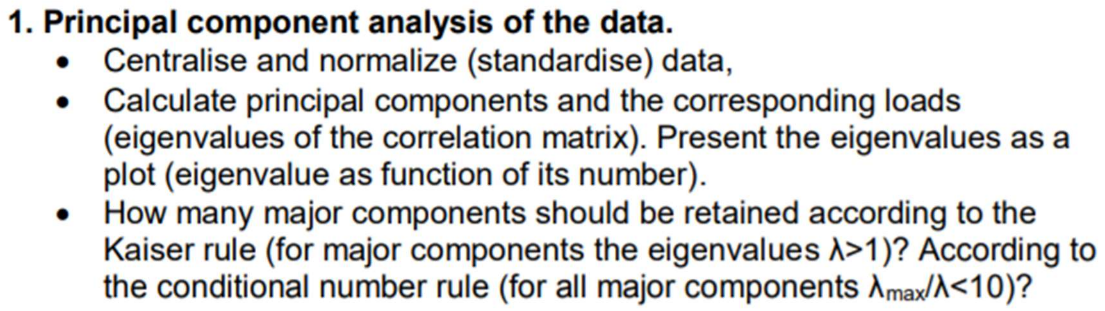
# HIRING TASK

Data Mining and Neural Networks

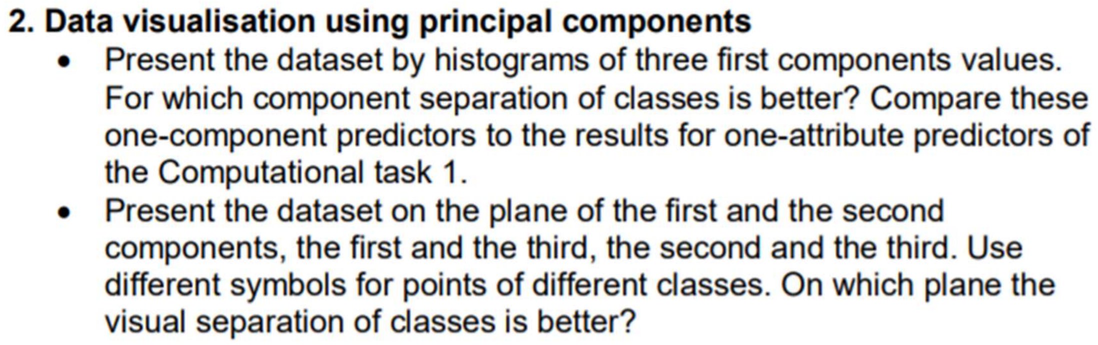
**Breast cancer diagnosis – 2**

Use Breast Cancer Wisconsin (Diagnostic) Data Set you have used in the first task and perform principal component analysis and k-means clustering.



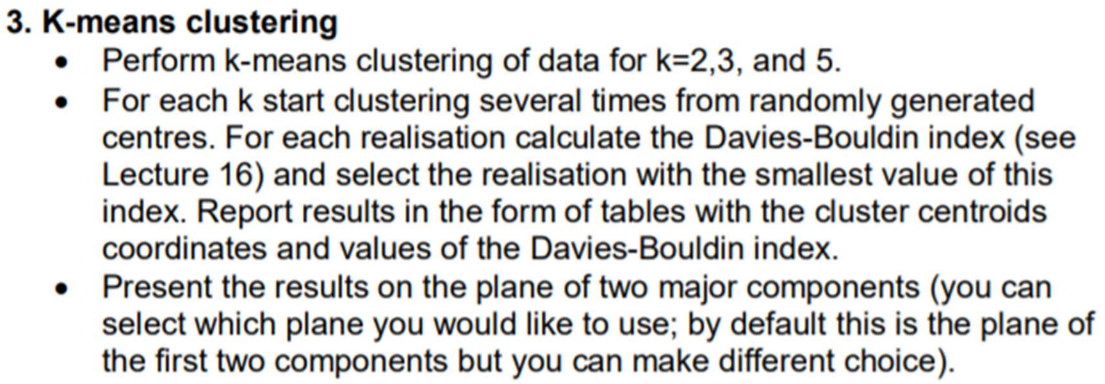
### 1. Principal component analysis of the data.

- Centralise and normalize (standardise) data,
- Calculate principal components and the corresponding loads (eigenvalues of the correlation matrix). Present the eigenvalues as a plot (eigenvalue as function of its number).
- How many major components should be retained according to the Kaiser rule (for major components the eigenvalues $\lambda>1$)? According to the conditional number rule (for all major components $\lambda_{max}/\lambda<10$)?



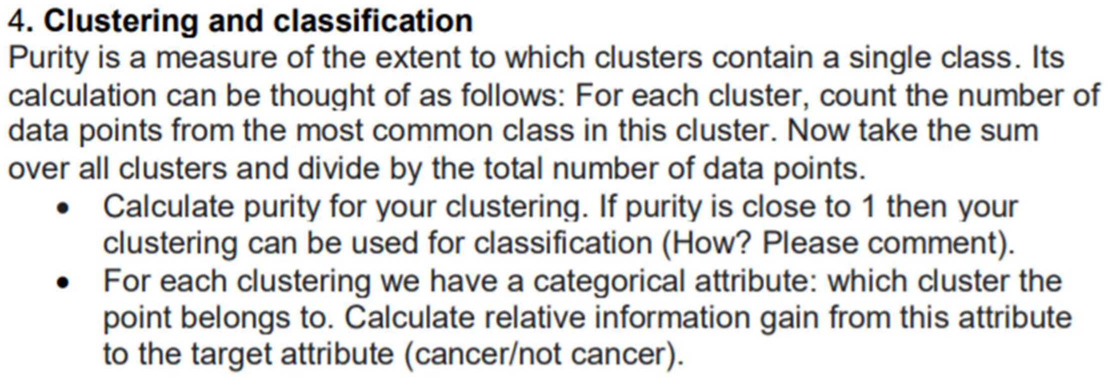
### 2. Data visualisation using principal components

- Present the dataset by histograms of three first components values. For which component separation of classes is better? Compare these one-component predictors to the results for one-attribute predictors of the Computational task 1.
- Present the dataset on the plane of the first and the second components, the first and the third, the second and the third. Use different symbols for points of different classes. On which plane the visual separation of classes is better?



### 3. K-means clustering

- Perform k-means clustering of data for k=2,3, and 5.
- For each k start clustering several times from randomly generated centres. For each realisation calculate the Davies-Bouldin index (see Lecture 16) and select the realisation with the smallest value of this index. Report results in the form of tables with the cluster centroids coordinates and values of the Davies-Bouldin index.
- Present the results on the plane of two major components (you can select which plane you would like to use; by default this is the plane of the first two components but you can make different choice).



### 4. Clustering and classification

Purity is a measure of the extent to which clusters contain a single class. Its calculation can be thought of as follows: For each cluster, count the number of data points from the most common class in this cluster. Now take the sum over all clusters and divide by the total number of data points.

- Calculate purity for your clustering. If purity is close to 1 then your clustering can be used for classification (How? Please comment).
- For each clustering we have a categorical attribute: which cluster the point belongs to. Calculate relative information gain from this attribute to the target attribute (cancer/not cancer).