COMP809 – K-means

Lab 5

World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardiovascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. The dataset framingham.csv is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. Each variable considered in this studied is a potential risk factor. There are both demographic, behavioral and medical risk factors. The variables are the following:

Demographic:

- sex: male or female.
- age: age of the patient.

Behavioural current

- Education: education level, being 0 the lowest level.
- Smoker: whether or not the patient is a current smoker.
- cigsPerDay: the number of cigarettes that the person smoked on average in one day.

Medical (history):

- BPMeds: whether or not the patient was on blood pressure medication.
- prevalentStroke: whether or not the patient had previously had a stroke.
- prevalentHyp: whether or not the patient was hypertensive.
- diabetes: whether or not the patient had diabetes.

Medical(current):

- totChol: total cholesterol level.
- sysBP: systolic blood pressure.
- diaBP: diastolic blood pressure.
- BMI: Body Mass Index.
- heartRate: heart rate.
- glucose: glucose level.

Predict variable (desired target):

- 10 year risk of coronary heart disease CHD (“1”, means “Yes”, “0” means “No”).

Work in the following:

1. Perform a cluster analysis via K-means.
    a. What is the optimal number of clusters according to the Elbow method? Justify your answer.
    b. What is the optimal number of clusters according to the Silhouette score? Justify your answer.
    c. What is the number of clusters that you propose? Justify your answer.
    d. Plot the cluster using the first two principal components. Comment on it.
2. Now we will evaluate the K-means method, an unsupervised machine learning technique,

taking advantage of the information we have in the data set. We will evaluate how good this method is at predicting the 10 year risk of coronary heart disease CHD . Note that the K-mean method does not require this variable.

a. Equate the number of 0s and 1s through the oversampling technique.
b. Perform a cluster analysis using K=2.
c. Plot the clusters using the first 2 principal components. Comment on it.
d. Identify the cluster that represents TenYearCHD= 0 and 1.
e. Using the clusters defined above as a prediction method. Calculate accuracy, sensitivity, and specificity. Comment on it.