**Data Set- Heart disease**

The dataset used in this assignment is the **heart disease** dataset available in ***heart-c.csv*** from the **Blackboard**. This dataset describes 13 risk factors for heart disease. The attribute ***num*** represents the (binary) class attribute: class <50 means no disease; class >50\_1 indicates increased level of heart disease. The following questions will allow you to demonstrate your knowledge of Unsupervised Learning and data exploration skills.

Data Dictionary:

age = age in years   
sex = sex (1 = male; 0 = female)   
cp = cp: chest pain type   
Value 1: typical angina   
Value 2: atypical angina   
Value 3: non-anginal pain   
Value 4: asymptomatic   
trestbps = resting blood pressure (in mm Hg on admission to the hospital)   
chol = serum cholestoral in mg/dl   
fbs = (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)   
restecg = resting electrocardiographic results :   
Value 0: normal   
Value 1: having ST-T wave abnormality (T wave inversions and/or ST   
elevation or depression of > 0.05 mV)   
Value 2: showing probable or definite left ventricular hypertrophy by   
Estes' criteria   
thalach = maximum heart rate achieved   
exang = exercise induced angina (1 = yes; 0 = no)   
oldpeak = ST depression induced by exercise relative to rest   
slope = the slope of the peak exercise ST segment :   
Value 1: upsloping   
Value 2: flat   
Value 3: downsloping   
ca = number of major vessels (0-3) colored by flourosopy   
thal = 3 = normal; 6 = fixed defect; 7 = reversable defect   
num = diagnosis of heart disease (angiographic disease status) :   
Value 0: < 50% no disease   
Value 1: > 50% increased level of heart disease

**1. Run K-means clustering on the above heart disease dataset and answer the following questions**

1. Why should the attribute “*class”* in ***heart-c.csv*** (“***num***”) **not** be included for clustering?

*As we are using an unsupervised algorithm when we use K-means, we want for the algorithm to make predictions without access to any labelled data. As “class” is effectively the output we wish to predict, it is wise not to include this as we want our model to be able to predict this based on patterns in the rest of the data and not this label, as it is not a label the model will have access to in a real-case example.*

1. Run the K-means algorithm and provide reasoning for the optimum value of K.
2. Which features would you expect to be less useful when using K-means and why?

*Various features within this dataset are categorical – for example sex, cp, fbs. Alongside many others. This presents us with the following issues:*

* 1. *It is not possible to plot this data in a dimensional space in a meaningful way.*
  2. *We cannot generate a meaningful mean value.*
  3. *As a result of (2) we cannot generate an error of sum squares*
  4. *We cannot effectively measure the Euclidian distance between data points and cluster centres when using such data.*

*As a result of the above factors, I would expect that the categorical features in this data set would be less useful.*

**2. Run the hierarchical clustering on above heart disease dataset, and answer the following questions**

1. Show the clustering results in a tree structure and provide reasoning for the optimal number of clusters
2. Describe the link method you used.
3. What are the strengths and limitations of this link method in hierarchical clustering?