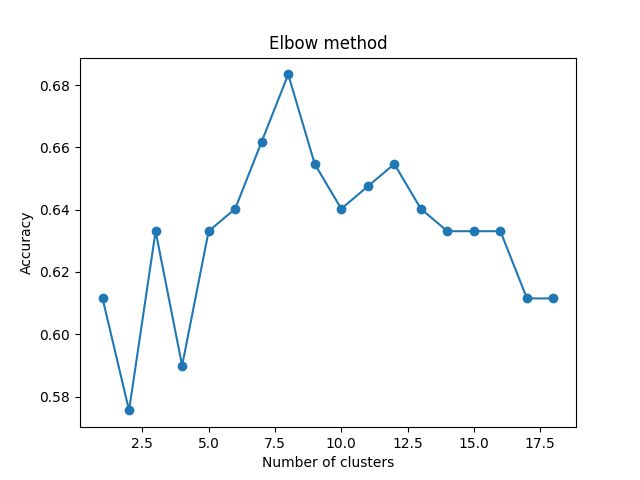
# OutPut:

R2 = 0.16492353199637777

Accuracy= [0.60431654676259, 0.5899280575539568, 0.6258992805755396, 0.5899280575539568, 0.6187050359712231, 0.6258992805755396, 0.6546762589928058, 0.6762589928057554, 0.6546762589928058, 0.6546762589928058, 0.6546762589928058, 0.6402877697841727, 0.6474820143884892, 0.6402877697841727, 0.6402877697841727, 0.6402877697841727, 0.6187050359712231, 0.6187050359712231]



K=8

# The approaches for hw4

1: Analyze data

SkinThickness: a lot of 0. which are missing.

Insulin: a lot of 0. which are missing

Pregnancies: from 0-15, seem ok

Glucose: range from 44-199, seem ok.

BloodPressure: To express the Blood pressure measurement(seems diastolic -- lower)

range from 24-122. low than 40 or higher than 120 seems impossible.but acceptable.

BMI: To express the Body mass index. 18.5-40 seems reasonable

The other columns seems acceptable.

2: remove invalid data in SkinThickness and Insulin column

3：impute data with MICE, but we need to limit the min and max for each column in case of invalid data(e.g. negative value)

4: train the model using LinearRegression and caculate the r2

5: the impute the miss values in columns “SkinThickness” and “Insulin”

6: using the model we just trained to predict BloodPressure

7:save to csv file

8:using the original data of hw4\_test with predicted BloodPressure to fit KNN model and calculate accuracy

9:using elbow method to find the k. 8 has a high accuracy, and the number is reasonable, so I choose 8.