**Code for homework 3**

# Bhargavi Moram

#999992733

#data mining assignment 3

**Import pandas and numpy**

#import the package pandas

import pandas

#import numpy

import numpy

**2 Read data set.**

# read the arrhythimia dataset using method read\_csv

data = pandas.read\_csv('arrhythmia.data',header=None)

**3 Clean data set.**

#replace the data using replace method

data = data.replace('?',numpy.NaN)

#import Imputer method

from sklearn.preprocessing import Imputer

#replace the all NaN missing values with column ean using Imputer method

imp = Imputer(missing\_values= 'NaN' , strategy='mean', axis = 0)

imp.fit(data)

data\_clean = imp.transform(data)

**4 Apply PCA to the cleaned data set.**

# import PCA

from sklearn.decomposition import PCA

#apply PCA to cleaned data set

pca = PCA(n\_components = 100)

pca.fit(data\_clean)

data\_reduced = pca.transform(data\_clean)

print(pca.explained\_variance\_ )

print(pca.components\_)

print(pca.explained\_variance\_ratio\_.sum())

**Output:** 







