Assignment 5

1. Projection in 2D

    a. Read the seed dataset using panda data frame. Convert to numpy array
    b. Plot the data points using the first two dimensions (area, perimeter). Use three different shapes (triangle, square, circle) to plot datapoints for three different classes. You should use the class information from class label and use them when you decide on shapes (see slide 6 and slide 10)
    c. Calculate the mean data point for each class and show them with similar shape with the larger size on the above plot.
    d. Calculate the **centered version** of area, perimeter. Plot them in a new plot with tile: "Centered data" . Let's $\overline{D}$ is the centered data matrix with two columns.
    e. Now calculate the projection of each data points on the line l (spanned by the vector $[-1.75\ 1.75\,]$). And plot the projected point on the line using the same shape but smaller size. So all smaller shapes would be on the line. (See slide 10)
    f. Now, in another figure, Plot them in a new plot with tile: "Centered data"
    g. Now, calculate the best basis vector u using the PCA method we have learned.
        i. Calculate cov = $\overline{D}^{T}\overline{D}$
        ii. Calculate S = n*cov
        iii. Find the two eigen vectors and their eigen values of S. Use **library**
        iv. Check whether they are orthonormal vectors. Otherwise use gram Schmidt **library** to get two orthonormal vectors from these eigen vectors.
        v. Take the eigen vector with largest eigen value and make it u.
        vi. Project all 2d points from Q1d on u.
        vii. Plot u and the projected points on u
        viii. Calculate new coordinates $[X]_U$ based on the basis U = [u]. Remember, new coordinates will be 1 dimensional

2. Projection in 3D:
    a. Plot the data points using the first three dimensions (area, perimeter, compactness) .
    b. Use three different shapes (triangle, square, circle) to plot datapoints for three different classes. You should use the class information from class label and use them when you decide on shapes (see slide 6 and slide 10)
    c. Calculate the mean data point (3D) for each class and show them with similar shape with the larger size on the above plot.
    d. Calculate the **centered version of** area, perimeter, compactness. Plot centered data points in 3D in a new plot. Let's $\overline{D}$ is the centered data matrix with three columns.
    e. Now, plot the plane (with yellow plane) spanned by two normal vector($[1 - 2,\ 1]^T$, $[2, 1, 0]^T$) on the plot done in step 4.
    f. Now, calculate the two best basis vectors u_1 and u_2 using the PCA method we have learned.
        i. Calculate cov = $\overline{D}^{T}\overline{D}$
        ii. Calculate S = n*cov
        iii. Find the three eigen vectors and their eigen values of S. Use library
        iv. Check whether they are orthonormal vectors. If not, use gram smidt library to calculate three ortho-normal vectors.

v. Take two orthonormal eigen vectors with largest eigen values and make them $\underline{u_1}$ and $\underline{u_2}$

vi. Project all 3d points from Q2d on $\underline{u_1}$ and $\underline{u_2}$.

vii. Plot the plane spanned by $\underline{u_1}$ and $\underline{u_2}$ and also the projected points on u.

viii. Calculate new coordinates $[X]_U$ based on the basis U = [u1, u2]. Remember, new coordinates will be 2 dimensional