We are performing Principal Component Analysis (PCA), K_means Clustering on Breast Cancer Wisconsin (Diagnostic) Data. The detailed report after performing PCA and K_means Clustering is described below with the help of diagrams.

1.Principal Component Analysis

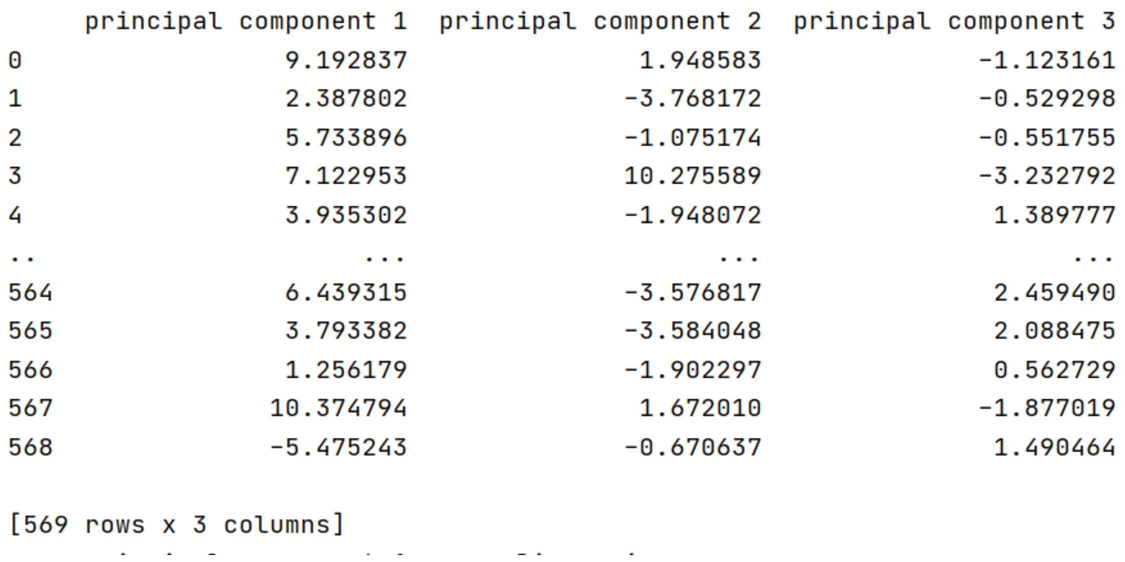
Principal Component Analysis is a technique used to reduce the dimensionality of a dataset into lower dimensions. Our dataset contains 30 real attribute, we can reduce it by 3 components using Principal Component Analysis technique.

a. Centralise and normalize (standardise) data

We are using StandardScaler to normalise the data.

b. Principal components and Eigen values of correlation matrix

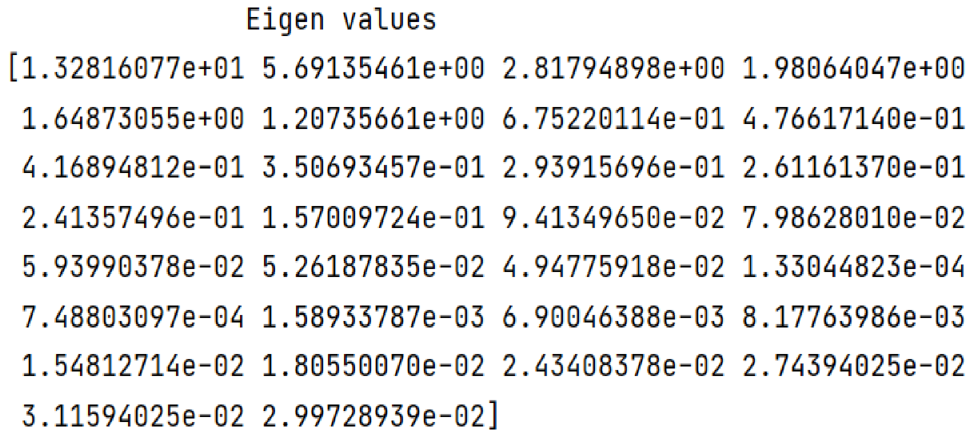
Breast Cancer Wisconsin (Diagnostic) Data with 30 attributes is reduced to 3 attributes using Principal Component Analysis. The screenshot of data we get after performing PCA is attached below:



```
     principal component 1  principal component 2  principal component 3
0                 9.192837               1.948583              -1.123161
1                 2.387802              -3.768172              -0.529298
2                 5.733896              -1.075174              -0.551755
3                 7.122953              10.275589              -3.232792
4                 3.935302              -1.948072               1.389777
..                     ...                    ...                    ...
564               6.439315              -3.576817               2.459490
565               3.793382              -3.584048               2.088475
566               1.256179              -1.902297               0.562729
567              10.374794               1.672010              -1.877019
568              -5.475243              -0.670637               1.490464

[569 rows x 3 columns]
```

The screenshot of calculated Eigen Values is given below:



```
                Eigen values
[1.32816077e+01 5.69135461e+00 2.81794898e+00 1.98064047e+00
 1.64873055e+00 1.20735661e+00 6.75220114e-01 4.76617140e-01
 4.16894812e-01 3.50693457e-01 2.93915696e-01 2.61161370e-01
 2.41357496e-01 1.57009724e-01 9.41349650e-02 7.98628010e-02
 5.93990378e-02 5.26187835e-02 4.94775918e-02 1.33044823e-04
 7.48803097e-04 1.58933787e-03 6.90046388e-03 8.17763986e-03
 1.54812714e-02 1.80550070e-02 2.43408378e-02 2.74394025e-02
 3.11594025e-02 2.99728939e-02]
```

Graph of computed Eigen value is given below (Figure.1.):

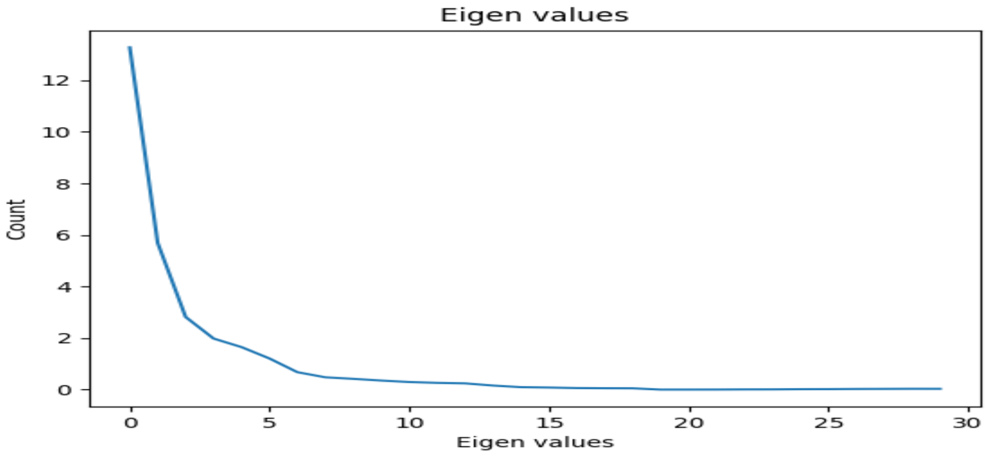


Figure.1 Eigen Values

c. Major components should be retained according to the Kaiser rule is first 6 Eigen values and first 5 Eigen values will retain according to the conditional number rule.

2. Data visualisation using principal components (0-Malignant case,1-Benign case)

a. Histogram of Principal Component 1(Figure 2):

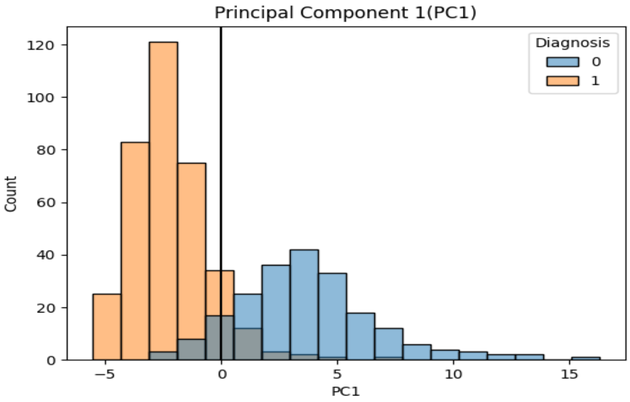


Figure2. PC1

Histogram of Principal Component 2(Figure 3):

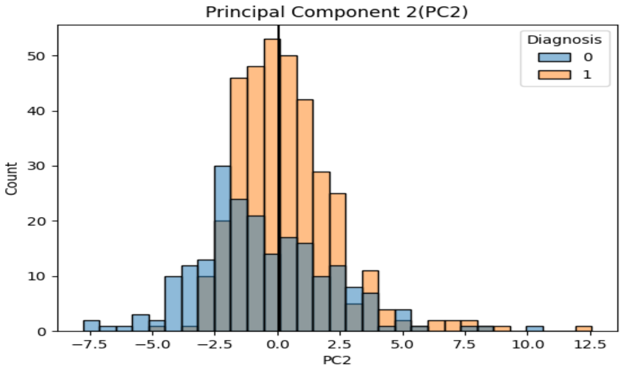


Figure 3. PC2

Histogram of Principal Component 3(Figure 4):

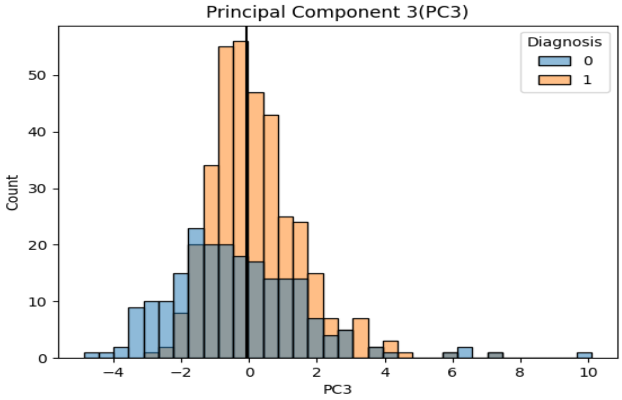


Figure4. PC3

$$\text{Classification Error (0)}= \frac{Number\ of\ wrongly\ classified\ Malignant\ case}{Total\ number\ of\ Malignant\ case}$$

$$\text{Classification Error (1)}= \frac{Number\ of\ wrongly\ classified\ Benign\ cases}{Total\ number\ of\ Benign\ case}$$

From these diagrams, we can analyse that PC1 provides better separation of classes (Malignant, Benign) than PC2 and PC3. We get Attribute 23 as better attribute to predict the classification in Computational Task 1. But PC1 predicts the classification better than Attribute 23 of Breast Cancer Wisconsin (Diagnostic) Data.

b. Dataset on plane for Principal component 1 and Principal component 2(*Figure.5*):

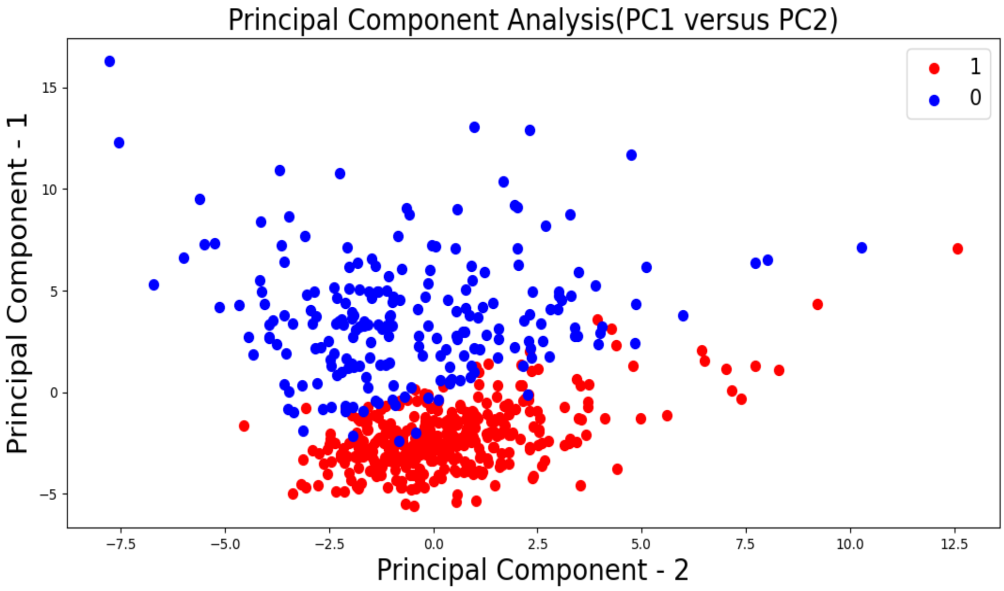


Figure.5

Dataset on plane for Principal component 2 and Principal component 3(Figure.6):

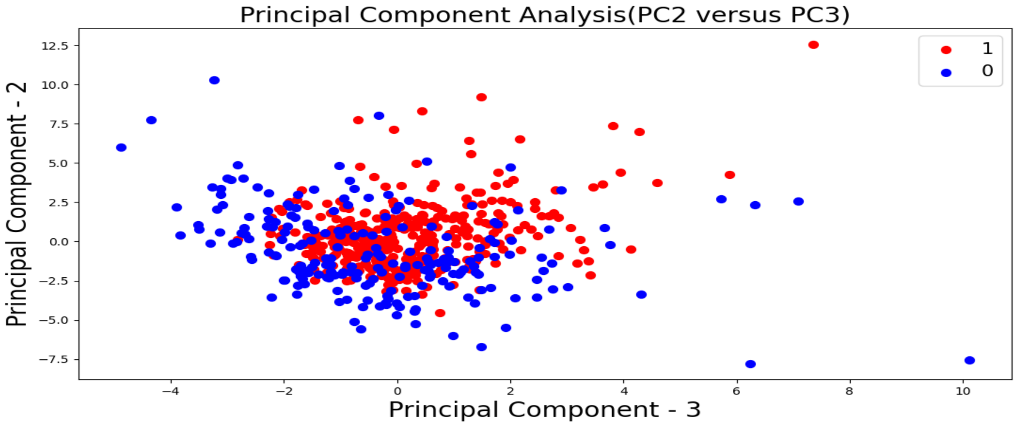


Figure.6

Dataset on plane for Principal component 1 and Principal component 3(Figure.7):

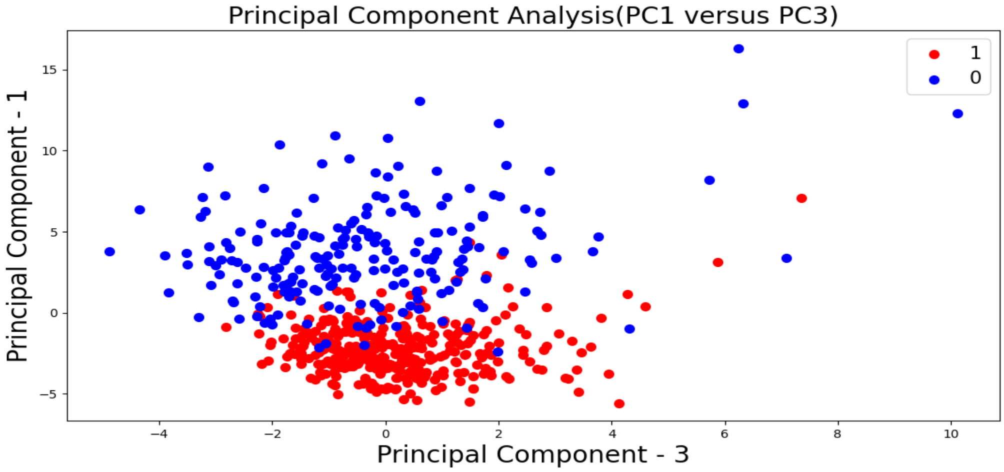


Figure.7

From the above 3 scatter plots, PC1 and PC2 gives better classification of Malignant case and Benign case than the other two scatter plots.

3. K_means Clustering

K-means clustering choose k objects as initial cluster centres and assign each object to clusters according to the cluster centroid. Cluster centroid will update by calculating the mean of all objects in that cluster until we observe there is no change in cluster centroid. But the K-means clustering is performed based on initial position of centres.

Centroid, DB Index when k=2 is shown in the below table:

Number	Centroid	DB Index
1	[[-2.20247773 -0.0217161] [4.35883027 0.04297741]]	0.8467403807096248
2	[[4.35883027 0.04297744] [-2.20247773 -0.02171612]]	0.8467403805997208

3	[[-2.20247773 -0.02171611] [4.35883027 0.04297743]]	0.8467403815006082
4	[[-2.20247773 -0.02171612] [4.35883027 0.04297744]]	0.8467403807255939
5	[[-2.20247773 -0.02171611] [4.35883027 0.04297742]]	0.8467403807712417

The best clustering for k = 2 is in second one.

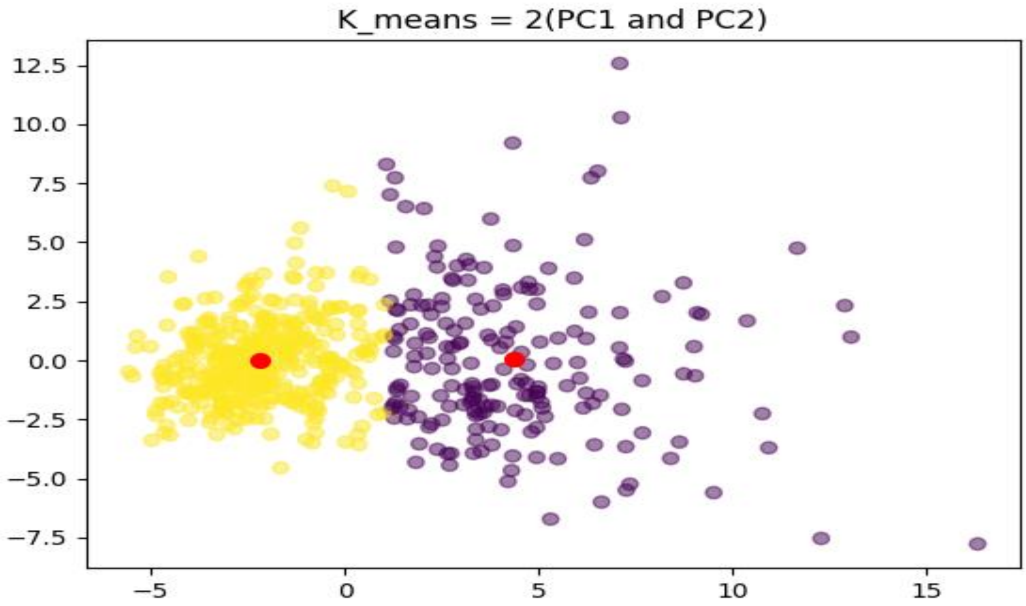


Figure 8. The best clustering of PC1 and PC2, when k=2

Figure 8. shows that there are 2 centroids with colour 'red' and clusters belonged to each centroid as shown in 'yellow' colour and 'violet' colour respectively. Here we have taken the clustering with lower DB Index = 0.8467403805997208.

<u>Centroid, DB Index when k=3 is shown in the below table:</u>

Number	Centroid	DB Index
1	[[-2.36453918 -0.25542324] [5.05903111 -1.74041619] [2.18623605 3.19710261]]	0.9105499640870622
2	[[-2.36453918 -0.25542325] [2.18623605 3.19710261] [5.05903111 -1.74041615]]	0.9105499633107857
3	[[-2.36453918 -0.25542325] [2.18623605 3.19710261] [5.05903111 -1.74041617]]	0.9105499624479153
4	[[5.05903111 -1.74041615] [-2.36453918 -0.25542325] [2.18623605 3.1971026]]	0.9105499650307233

5	[[2.18623605 3.19710262] [-2.36453918 -0.25542324] [5.05903111 -1.74041618]]	0.9105499614360678

The best clustering for k=3 is fifth one.

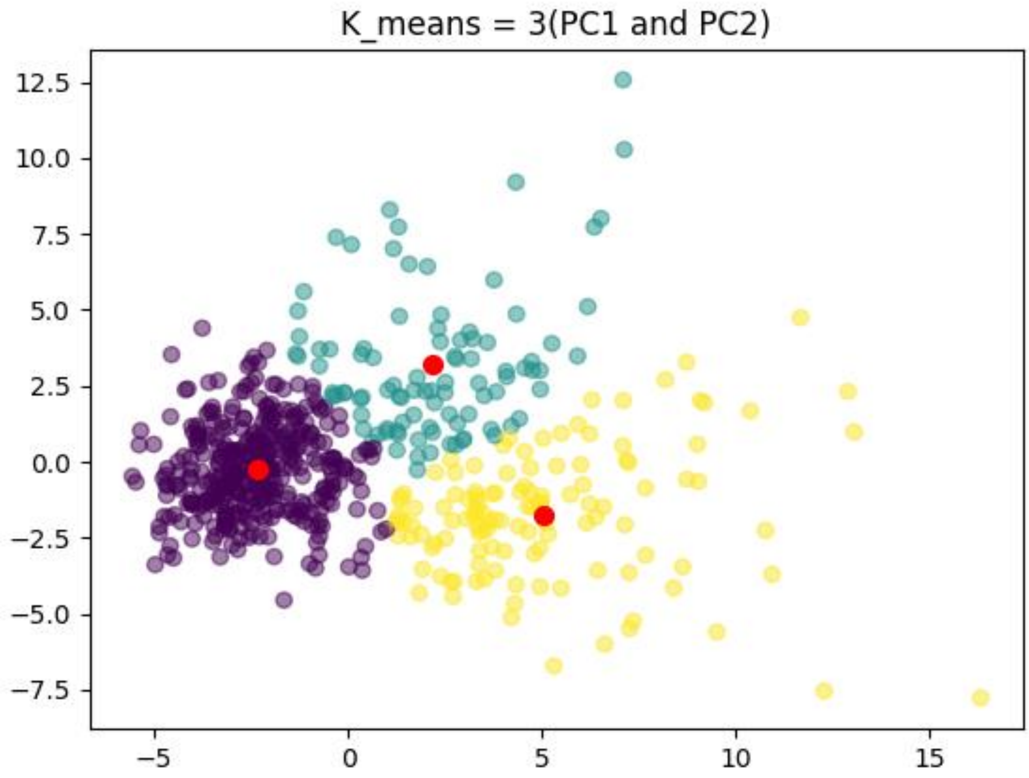


Figure 9. The best clustering of PC1 and PC2, when k=3

Figure.9. shows that there are 3 centroids with red colour and there corresponding clusters are in violet, yellow, cyan colours respectively.

Centroid, DB Index when k=5 is shown in the below table:

Number	Centroid	DB Index
1	[[3.4267872 4.42736211] [-2.87805583 -0.98406106] [-1.26405677 1.43527933] [8.34177559 -1.21122888] [3.08847132 -1.70766792]]	0.9586561878234068
2	[[3.08847132 -1.70766794] [-2.87805583 -0.98406106] [8.34177559 -1.21122872] [-1.26405678 1.43527928] [3.4267872 4.42736216]]	0.9586561998945384
3	[[3.42678719 4.42736215] [3.1621223 -1.68371417]	0.9577836892080217

	[8.48509005 -1.27137637] [-2.9068289 -0.98835296] [-1.24573133 1.3810009]]	
4	[[8.48509005 -1.2713763] [-2.87805583 -0.98406105] [3.4267872 4.42736216] [3.1621223 -1.68371418] [-1.25389106 1.42006687]]	0.9576435316607668
5	[[3.08847132 -1.70766791] [3.4267872 4.42736213] [-2.87805583 -0.98406107] [-1.26405677 1.43527932] [8.34177559 -1.21122887]]	0.9586561896622257

The best clustering for k=5 is fourth trial

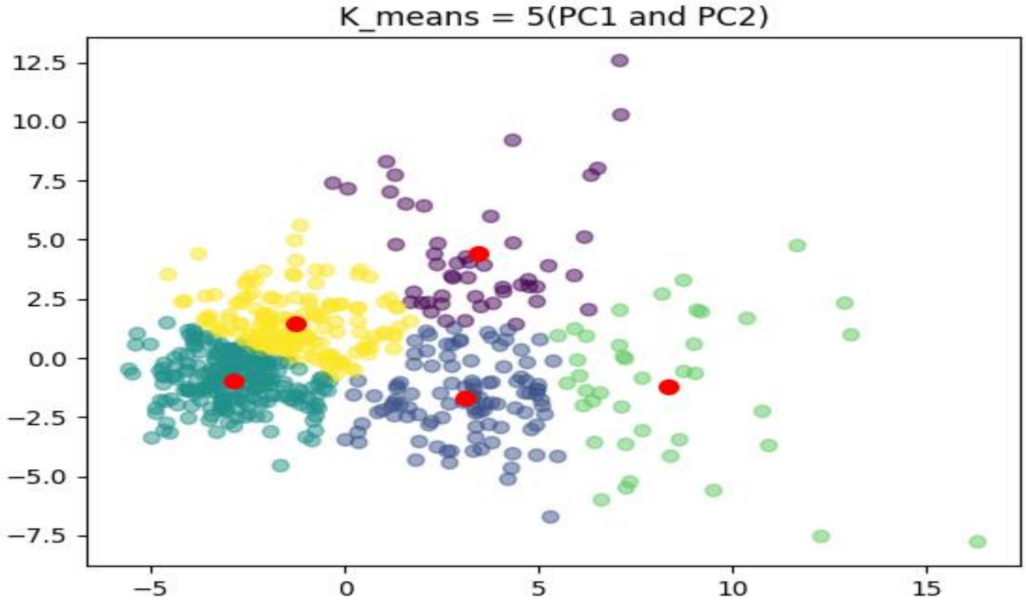


Figure 10. The best clustering of PC1 and PC2, when k = 5

Figure.10. shows there are 5 centroids with red colour and there corresponding clusters in cyan, yellow, green, violet, blue colours respectively.

4. Clustering and classification

1. K = 2

Cluster	Majority class	Number of cases of Majority classes	Total
1	0(Malignant case)	175	212
2	1(Benign case)	341	357
Total		516	569

Purity $= \frac{516}{569} \approx .90$
The purity is close to 1, so we can take k=2 clustering for classification.

Entropy of X, $\mathbf{H}(\mathbf{x}) \quad = -\sum_{j=1}^{m} \boldsymbol{p_j} \, \mathbf{log_2} \, \boldsymbol{p_j}$

$$\text{H}(class) = -\frac{212}{569}\log_2\frac{212}{569} - \frac{357}{569}\log_2\frac{357}{569}$$
$$= -0.37 * (-1.43) - 0.627 * (-0.673)$$
$$= 0.5291 + 0.4219$$
$$\approx 0.95$$

$$\text{H}(class|cluster = 1) = -(\frac{175}{212})\log_2\frac{175}{212} - \frac{37}{212}\log_2\frac{37}{212}$$
$$= -0.825 * (-0.277) - 0.174 * (-2.522)$$
$$= 1.54$$

$$\text{H}(class|cluster = 2) = -(\frac{341}{357})\log_2\frac{341}{357} - \frac{16}{357}\log_2\frac{16}{357}$$
$$= -0.955 * (-0.066) - 0.044 * (-4.50)$$
$$= 0.261$$

$$\text{H}(class|cluster) = \frac{212}{569} * (1.54) + \frac{357}{569} * 0.261$$
$$= 0.7375$$

$$Information\ Gain(class|cluster) = H(class) - H(class|cluster)$$
$$= 0.95 - 0.7375$$
$$= 0.2125$$

$$Relative\ Information\ Gain = \frac{Information\ Gain\ (class|cluster)}{H(class)}$$
$$= \frac{0.2125}{0.95}$$
$$= 0.223$$

2. K = 3

Cluster	Majority Class	Number of cases of Majority Classes	Total
1	0(Malignant case)	56	95
2	1(Benign case)	318	351
3	0(Malignant case)	123	123
Total		497	569

Purity $= \frac{497}{569} \approx 0.873$

The purity_value is close to 1, so we can take k=3 clustering for classification.

$$\text{H}(class) = -\frac{212}{569}\log_2\frac{212}{569} - \frac{357}{569}\log_2\frac{357}{569}$$
$$= -0.37 * (-1.43) - 0.627 * (-0.673)$$
$$= 0.5291 + 0.4219$$
$$\approx 0.95$$

$$\text{H}(class|cluster = 1) = -(\frac{56}{95})\log_2\frac{56}{95} - \frac{39}{95}\log_2\frac{39}{95}$$
$$= -0.58 * (-0.78) - 0.41 * (-1.28)$$
$$= 0.974$$

$$\begin{aligned} \text{H}(class|cluster = 2) &= -(\tfrac{318}{351})\log_2 \tfrac{318}{351} - \tfrac{33}{351}\log_2 \tfrac{33}{351} \\ &= -0.90 * (-0.15) - 0.09 * (-3.47) \\ &= \ 0.4473 \end{aligned}$$

$$\begin{aligned} \text{H}(class|cluster = 3) &= -(\tfrac{123}{123})\log_2 \tfrac{123}{123} \\ &= -1 * 0 \\ &= \ 0 \end{aligned}$$

$$\begin{aligned} \text{H}(class|cluster)) &= \tfrac{95}{569} * 0.974 + \tfrac{351}{569} * 0.4473 + \tfrac{123}{123} * 0 \\ &= 0.4385 \end{aligned}$$

$$\begin{aligned} Information\ Gain(class|cluster) &= H(class) - H(class|cluster) \\ &= 0.95 - 0.4385 \\ &= \ 0.5115 \end{aligned}$$

$$\begin{aligned} Relative\ Information\ Gain &= \frac{Information\ Gain\ (class|cluster)}{H(class)} \\ &= \frac{0.5115}{0.95} \\ &= 0.538 \end{aligned}$$

3. K = 5

Cluster	Majority class	Number of cases of Majority classes	Total
1	0(Malignant case)	105	105
2	0(Malignant case)	35	49
3	1(Benign case)	202	218
4	1(Benign case)	141	157
5	0(Malignant case)	40	40
Total		523	569

Purity= $\frac{523}{569} \approx 0.919$

The purity _value is close to 1. So, k=5 gives a good classification.

$$\begin{aligned} \text{H}(class) &= -\tfrac{212}{569}\log_2 \tfrac{212}{569} - \tfrac{357}{569}\log_2 \tfrac{357}{569} \\ &= -0.37 * (-1.43) - 0.627 * (-0.673) \\ &= 0.5291 + 0.4219 \\ &\approx 0.95 \end{aligned}$$

$$\begin{aligned} \text{H}(class|cluster = 1) &= -(\tfrac{105}{105})\log_2 \tfrac{105}{105} \\ &= -1 * 0 \\ &= \ 0 \end{aligned}$$

$$\begin{aligned} \text{H}(class|cluster = 2) &= -(\tfrac{35}{49})\log_2 \tfrac{35}{49} - \tfrac{14}{49}\log_2 \tfrac{14}{49} \\ &= -0.714 * (-0.486) - 0.285 * (-1.81) \\ &= \ 0.862 \end{aligned}$$

$$\begin{aligned} \text{H}(class|cluster = 3) &= -(\tfrac{202}{218})\log_2 \tfrac{202}{218} - \tfrac{16}{218}\log_2 \tfrac{16}{218} \\ &= -0.92 * (-0.12) - 0.073 * (-3.77) \end{aligned}$$

$$= \ 0.38561$$

$$\begin{aligned}\mathrm{H}(class|cluster=4) &= \ -(\tfrac{141}{157})\log_2\tfrac{141}{157} - \tfrac{16}{157}\log_2\tfrac{16}{157} \\ &= \ -0.89*(-0.16) - 0.10*(-3.32) \\ &= \ 0.4744\end{aligned}$$

$$\begin{aligned}\mathrm{H}(class|cluster=5) &= \ -(\tfrac{40}{40})\log_2\tfrac{40}{40} \\ &= \ -1*0 \\ &= \ 0\end{aligned}$$

$$\begin{aligned}\mathrm{H}(class|cluster) &= \tfrac{105}{569}*0 + \tfrac{49}{569}*0.862 + \tfrac{218}{569}*0.385 + \tfrac{157}{569}*0.4744 + \tfrac{40}{569}*0 \\ &= 0.352\end{aligned}$$

$$\begin{aligned}Information\ Gain(class|cluster) &= H(class) - H(class|cluster) \\ &= 0.95 - 0.352 \\ &= 0.598\end{aligned}$$

$$\begin{aligned}Relative\ Information\ Gain &= \frac{Information\ Gain(class|cluster)}{H(class)} \\ &= \frac{0.598}{0.95} \\ &= 0.629\end{aligned}$$