LAB 2(A)

Name: Swarad Ganesh Gat
SBU ID: 114833402

Dataset: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

Description: The dataset contains numerical data regarding breast cancer. It contains 31 attributes, which describe the various properties of the tumour. The label of this dataset is ‘diagnosis’, which has two labels. The tumour is either malignant or benign, malignant meaning that it is cancerous and benign meaning that it is non-cancerous or harmless. There are 31 attributes which describe the tumour, including ‘radius_mean’, ‘texture_mean’, ‘area_mean’, etc.

Snapshot of data:

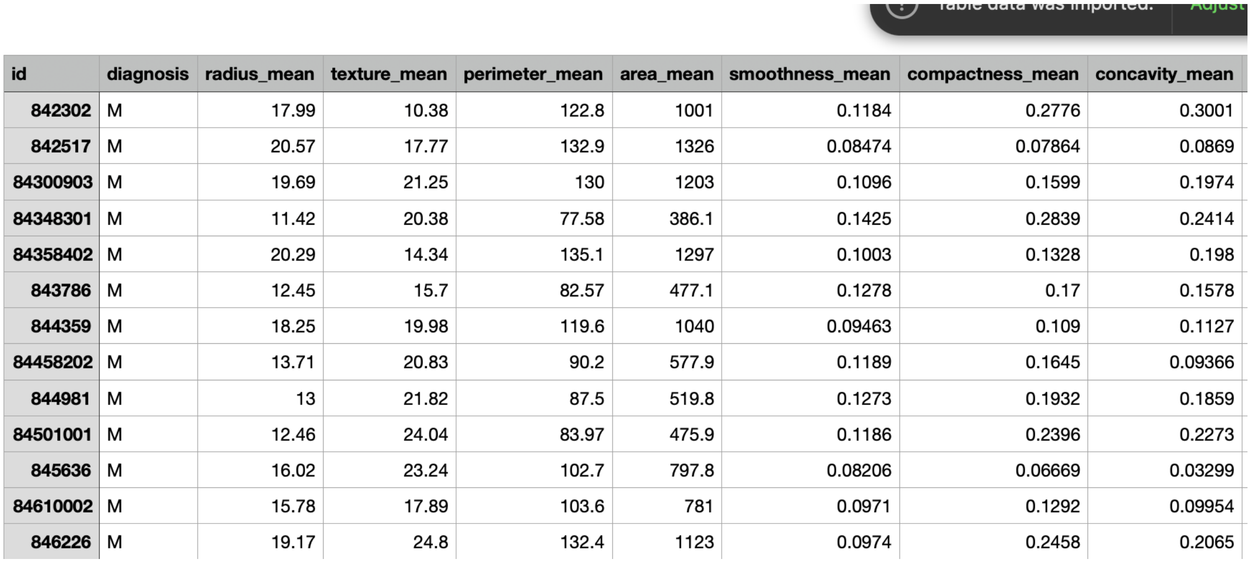


Table data was imported. Adjust

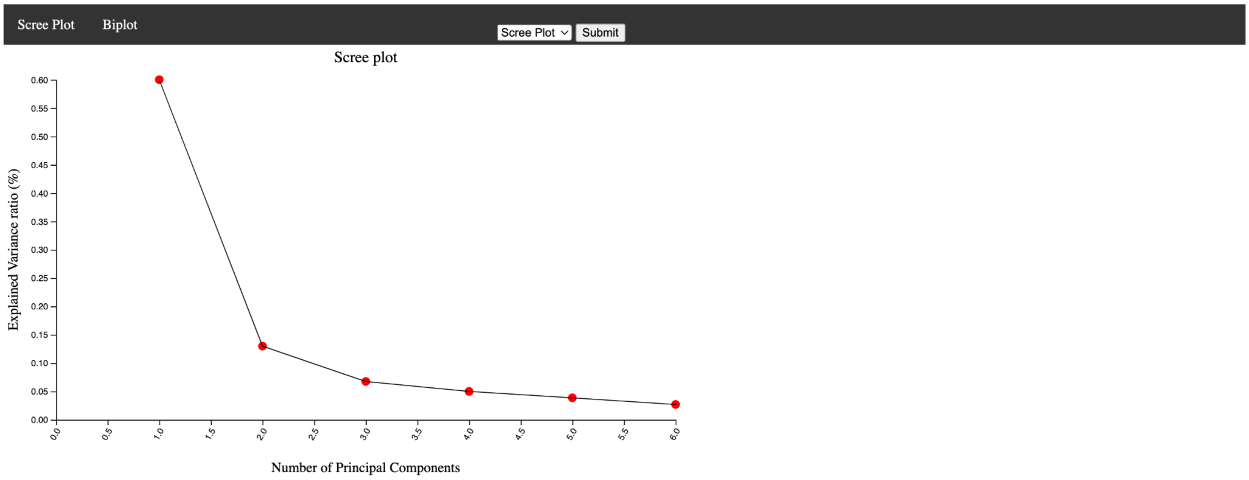
id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299
84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954
846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065

I have performed Principal Component Analysis on this dataset and visualised the results. Visualisations I have performed include:

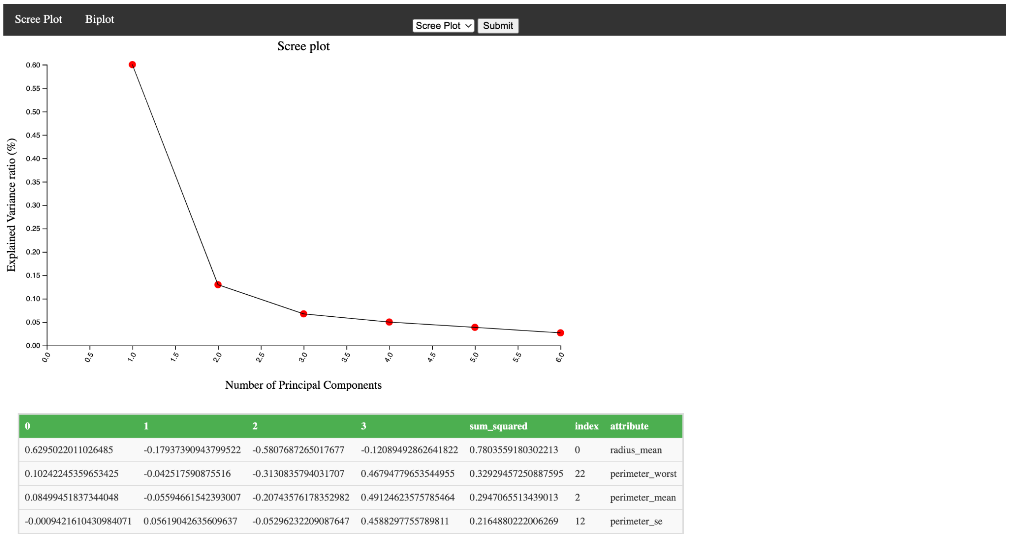
- Scree Plot
- Biplot
- Scatterplot

Visualizations:

1) Scree Plot:



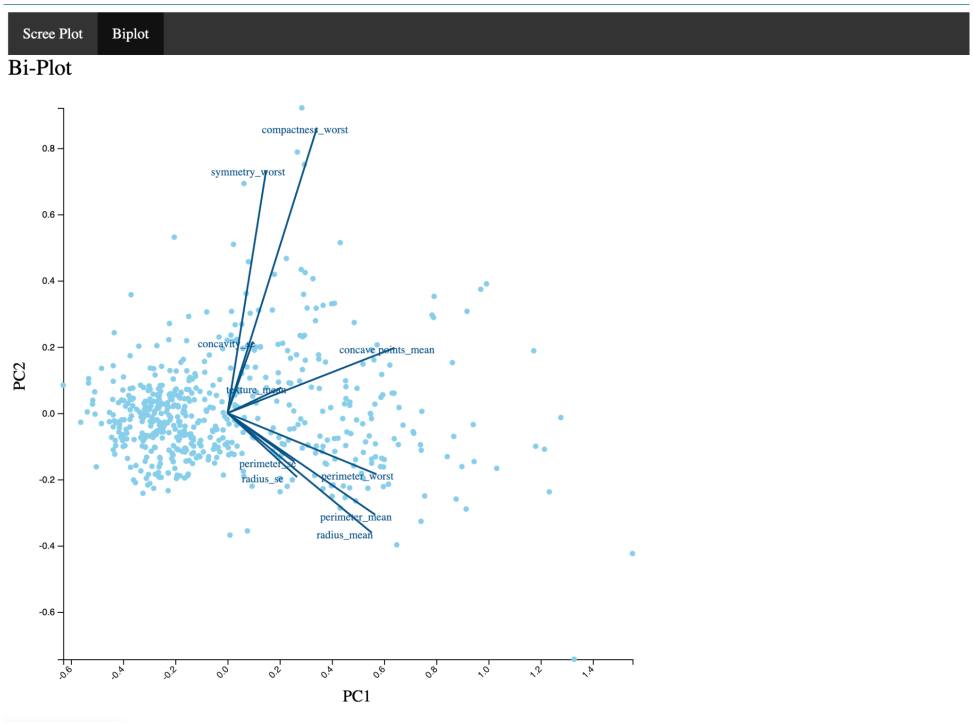
→ I have taken 6 principal components. The Y-Axis contains '% explained ration variance', which shows the percentage variance ratio along each component.



0	1	2	3	sum_squared	index	attribute
0.6295022011026485	-0.17937390943799522	-0.5807687265017677	-0.12089492862641822	0.7803559180302213	0	radius_mean
0.10242245359653425	-0.042517590875516	-0.3130835794031707	0.46794779653544955	0.32929457250887595	22	perimeter_worst
0.08499451837344048	-0.05594661542393007	-0.20743576178352982	0.49124623575785464	0.2947065513439013	2	perimeter_mean
-0.0009421610430984071	0.05619042635609637	-0.05296232209087647	0.4588297755789811	0.2164880222006269	12	perimeter_se

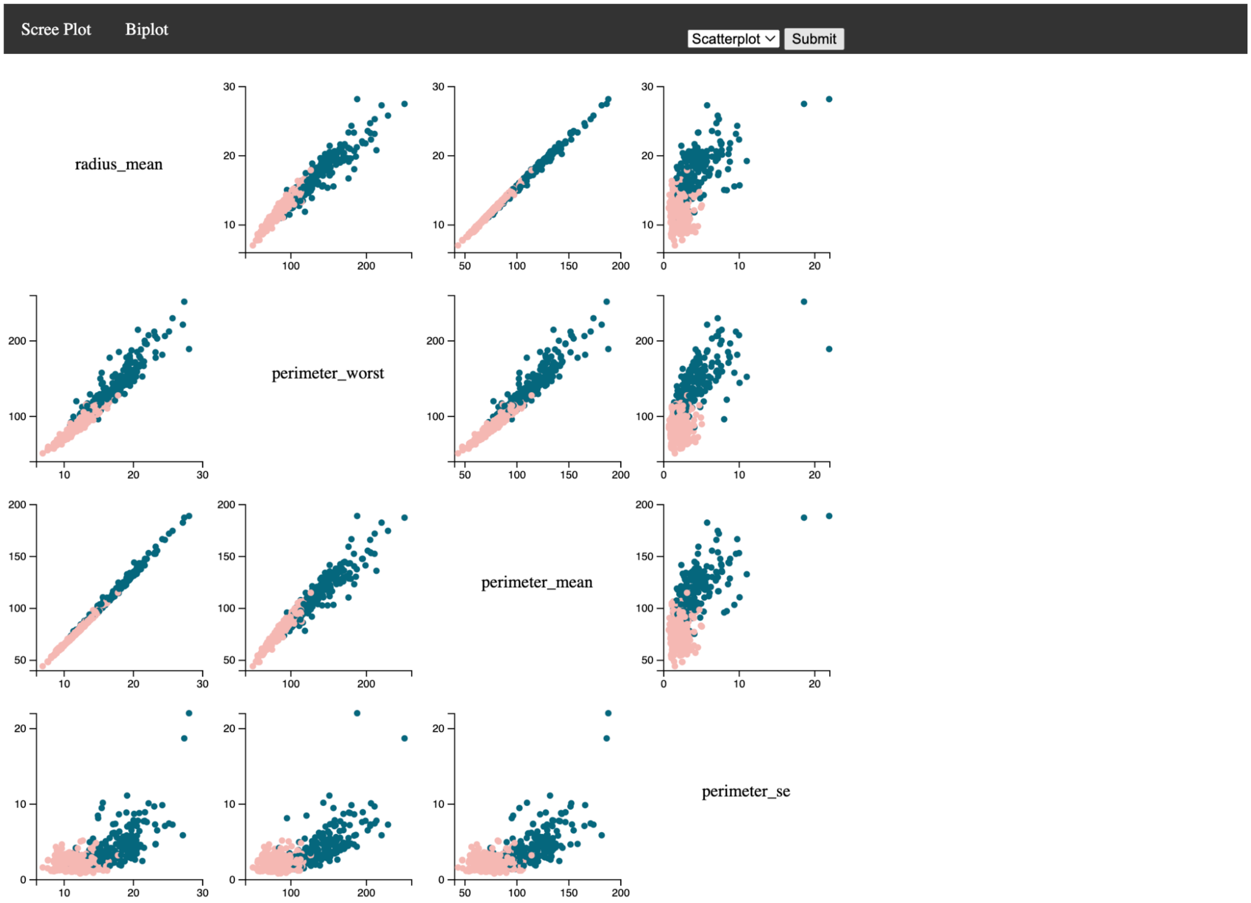
→ On clicking the fourth principal component, i get the following values. We can see the top four attributes for this component. On clicking on each component, we see different attributes.

2) Biplot:



→ This is the biplot. On clicking on the navbar, we can see the top ten attributes, and the datapoints. We can see the attribute names on the lines.

3) Scatterplot:



→ On selecting the intrinsic dimensionality index(di), we get the top 4 attributes for that principal component. The four attributes are taken and a dynamic scatterplot is plotted onto the html file. The four attributes are on the diagonal. The two labels are “Malignant” and “Benign”. Accordingly, it is color coded and we can see the two classes.

→ **Concluding, I have successfully visualized the "Breast Cancer" dataset. I have implemented Screeplot, Biplot and Scatterplot matrix.**