Report : Applying PCA on Breast Cancer Data

Given : The dataset has samples labelled as M or B for Malignant or Benign and the Variables from radius_mean .. to .. fractal_dimension_worst

Goal: So we are trying to assess which Variables play a pivotal role in classifying whether a tumor is M or B; for this we use PCA

1.

PCA: Principal Component Analysis is used to reduce dimensionality of datasets:

How does it work: We project the sample data onto a line that passes through the origin and then try to consider one such line which “either minimizes the distance of the actual point and its projection” or “maximizes the distance of the projection from the origin”.

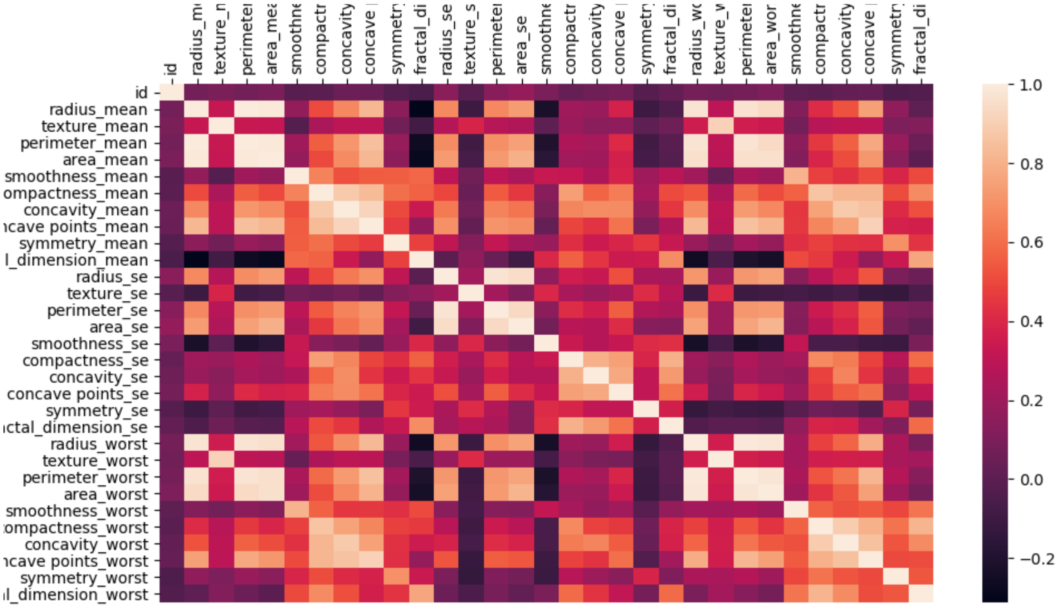
2.

The Dataset is (569, 33) and it shows that there is a “Unnamed Column : 32 ” which is extra and needs to weeded out.

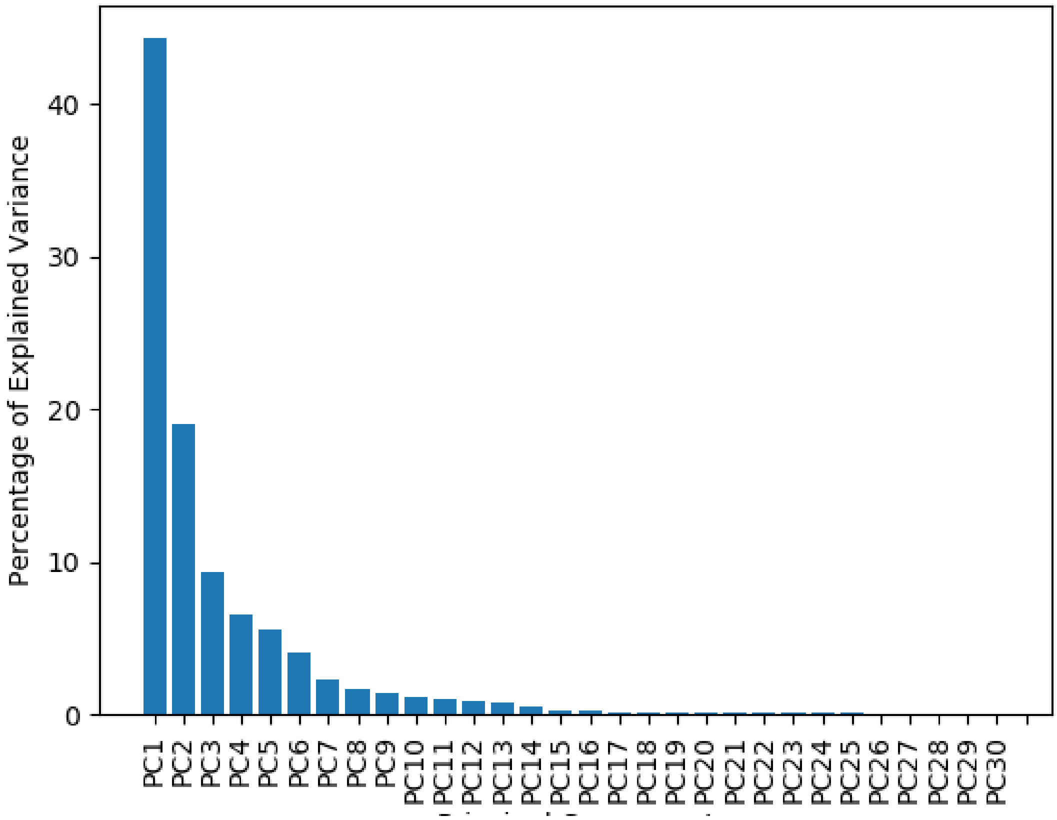
```
<bound method NDFrame.head of          id diagnosis  ...  fractal_dimension_worst  Unnamed: 32
0     842302         M  ...                  0.11890          NaN
1     842517         M  ...                  0.08902          NaN
2   84300903         M  ...                  0.08758          NaN
3   84348301         M  ...                  0.17300          NaN
4   84358402         M  ...                  0.07678          NaN
```

```
Code: empty_cols = [col for col in data.columns if data[col].isnull().all()]
data.drop(empty_cols,
        axis=1,
        inplace=True)
```

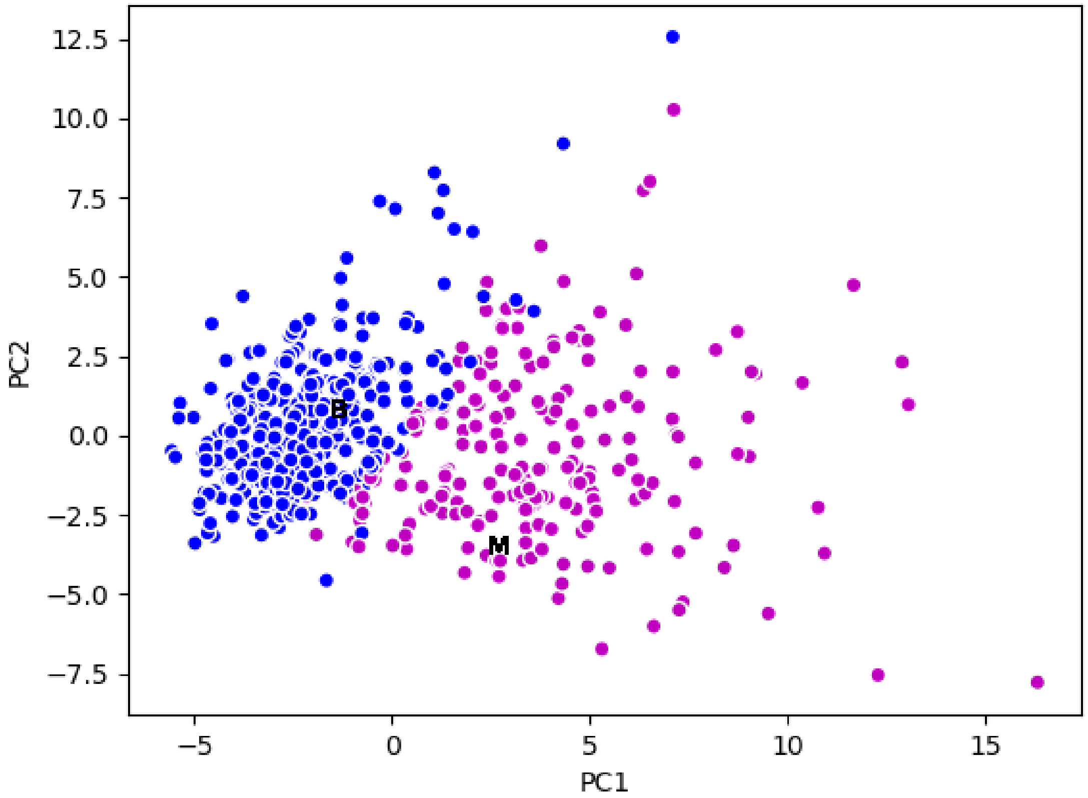
3.The following Correlation matrix highlights that there are many variables which are related.



4. So lets apply PCA to see which Principal Component should we consider. First Scale the Data so that we center the dataset. The below fig shows that PC1 and PC2 constitutes the majority of Variation.



5. Plot the points in scatter plot for PC1 and PC2:



Conclusion :. As we know from the bar graph PC1 has the maximum variance so lets see the loading_score for PC1 i.e which attributes constitute the PC1 and its weight:

Attribute	Weight
smoothness_se	0.014531
texture_se	0.017428
symmetry_se	0.042498
fractal_dimension_mean	0.064363
fractal_dimension_se	0.102568
texture_mean	0.103725
texture_worst	0.104469
symmetry_worst	0.122905
smoothness_worst	0.127953
fractal_dimension_worst	0.131784
symmetry_mean	0.138167
smoothness_mean	0.14259
concavity_se	0.15359
compactness_se	0.170393

concavepoints_se	0.183417
area_se	0.20287
radius_se	0.205979
compactness_worst	0.210096
perimeter_se	0.211326
radius_mean	0.218902
area_mean	0.220995
area_worst	0.224871
perimeter_mean	0.227537
radius_worst	0.227997
concavity_worst	0.228768
perimeter_worst	0.23664
compactness_mean	0.239285
concavepoints_worst	0.250886
concavity_mean	0.2584
concavepoints_mean	0.260854