**Experiment No 10**

**Title:** To Implement PCA for given problem

**Tools Required: Anaconda Navigator**

**Concept :**

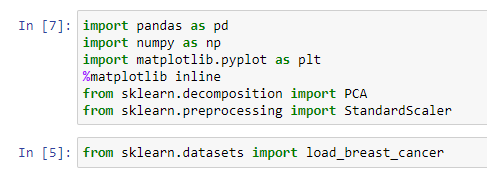
PCA is an unsupervised pre-processing task that is carried out before applying any ML algorithm. PCA is based on “orthogonal linear transformation” which is a mathematical technique to project the attributes of a data set onto a new coordinate system. The attribute which describes the most variance is called the first principal component and is placed at the first coordinate**.**Similarly, the attribute which stands second in describing variance is called a second principal component and so on. In short, the complete dataset can be expressed in terms of principal components. Usually, more than 90% of the variance is explained by two/three principal components.

**Example Problem**

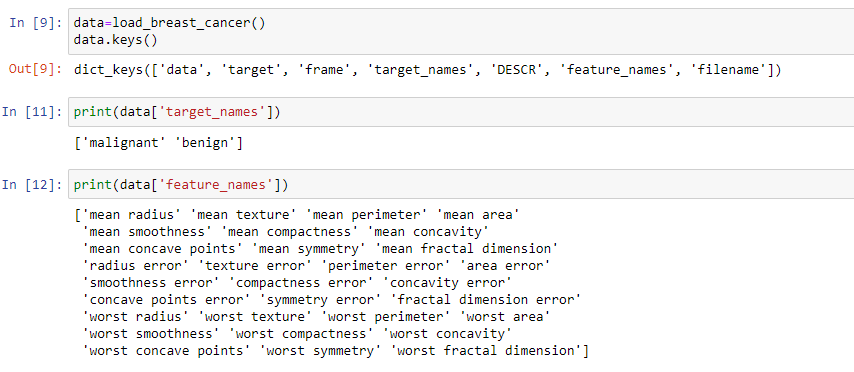
Here the breast\_cancer dataset is used. The dataset has 569 data items with 30 input attributes. There are two output classes-benign and malignant. Due to 30 input features, it is impossible to visualize this data. So we will find the principal components for the dataset given.

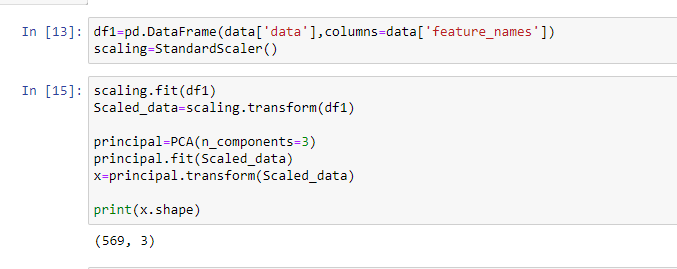
**Steps:**

1. Import the required libraries and the dataset from scikit-learn datasets.

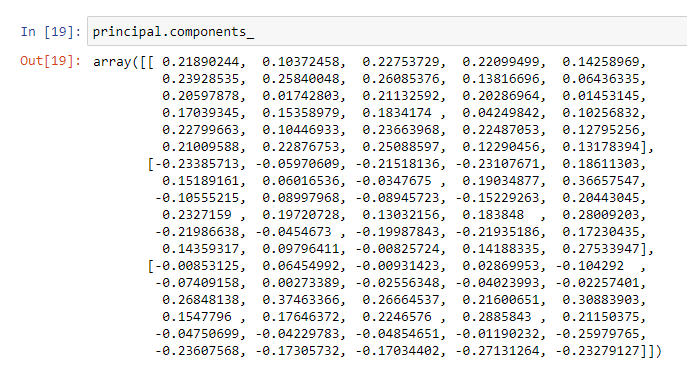


1. Apply PCA
   1. Standardize the dataset prior to PCA -it is essential to standardize/normalize the data before applying PCA.
   2. Import PCA from sklearn.decomposition -
   3. Choose the number of principal components- Usually, n\_components is chosen to be 3 for better visualization but it matters and depends on data.

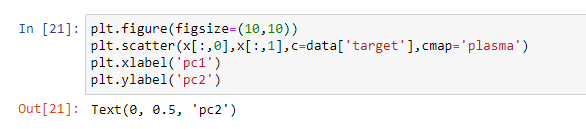


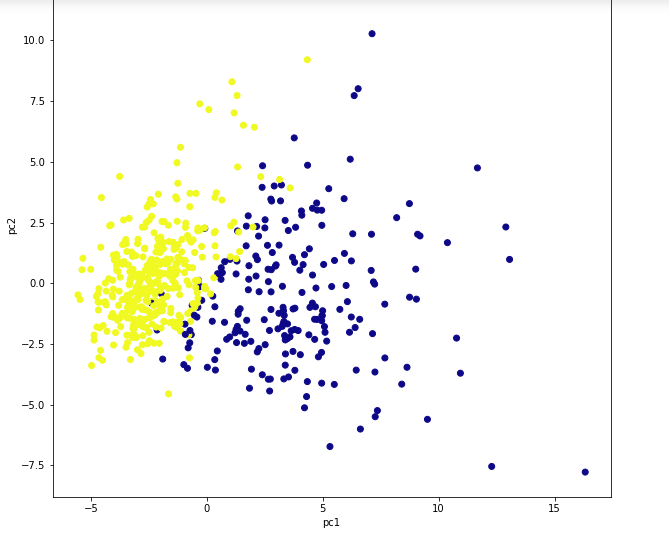


1. Check Principal Components.



1. Plot the components (Visualization)





1. Calculate variance ratio- Explained\_variance\_ratio provides an idea of how much variation is explained by principal components.

