**Aim:**

To implement PCA for dimensionality reduction using scikit learn

**Theory:**

Principal component analysis, or PCA, is a statistical technique to convert high dimensional data to low dimensional data by selecting the most important features that capture maximum information about the dataset. The features are selected on the basis of variance that they cause in the output. The feature that causes highest variance is the first principal component. The feature that is responsible for second highest variance is considered the second principal component, and so on. It is important to mention that principal components do not have any correlation with each other.

Advantages of PCA

* The training time of the algorithms reduces significantly with less number of features.
* It allows to analyze data in high dimensions.

**Experiment:**

**PCA applied on Naïve Bayes Classification:**

1. Without PCA

Data Analysis Report :

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
| 14 | 0 | 0 |
| 0 | 17 | 0 |
| 0 | 1 | 14 |

Accuracy = 98 %

1. With PCA (n = 2)

Data Analysis Report :

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
| 18 | 0 | 0 |
| 0 | 16 | 0 |
| 0 | 0 | 12 |

Accuracy = 100 %

1. With PCA (n = 1)

Data Analysis Report :

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
| 17 | 0 | 0 |
| 0 | 14 | 3 |
| 0 | 3 | 9 |

Accuracy = 87 %

NB : Number of Components (Attribute Size / 2 and Attibute Size/4)

**Conclusion:** The data analysis was performed on IRIS dataset with and without applying PCA. The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent. This way PCA model was studied.