**PRACTICE #03 – Scikit-learn with PCA & LDA**

Multivariate Statistical Analysis – 19TGMT



**Teacher:**

* PhD. Lý Quốc Ngọc
* MS. Nguyễn Mạnh Hùng
* MS. Phạm Thanh Tùng

**Student:**

* Chung Kim Khánh (19127644)

**Table of content**

[I. INFORMATION 2](#_Toc101802691)

[II. DESCRIBE THE DATA INFORMATION 2](#_Toc101802692)

[III. SOME BASIC MULTIVARIATE ANALYSIS WITH VISUALIZATION 9](#_Toc101802693)

[IV. PCA & LDA 11](#_Toc101802694)

[V. REFERENCE 18](#_Toc101802695)

# INFORMATION

|  |  |  |
| --- | --- | --- |
| student id | fullname | MAil |
| 19127644 | Chung Kim Khánh | ckkhanh19@clc.fitus.edu.vn |

# DESCRIBE THE DATA INFORMATION

The data is the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

The attributes are:

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

All attributes are continuous. The 1st attribute is the class identifier (1-3).

## Read data from the given sample CSV file

Text

Description automatically generated

The output:

Calendar

Description automatically generated

Text

Description automatically generated

## Describe the data information

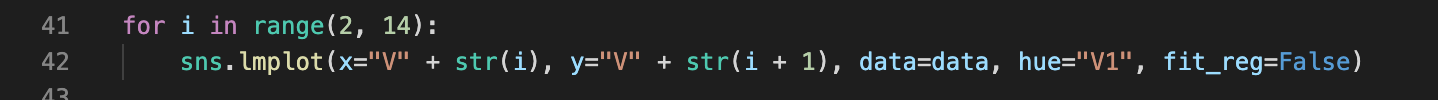
Plot data for visualization



Diagram

Description automatically generated

Scatterplot with the data points labelled by their Group



Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

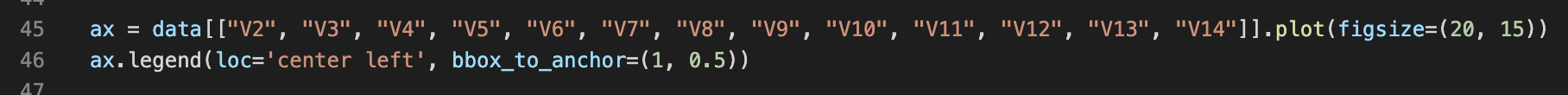
Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Profile plot, used to shows the variation in each of the variables, by plotting the value of each of the variables for each of the samples



Chart, line chart

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

# SOME BASIC MULTIVARIATE ANALYSIS WITH VISUALIZATION

Calculating summary statistics for multivariate data

Text

Description automatically generated

The output:

* Mean

Text

Description automatically generated

* Standard deviation

Text

Description automatically generated

* Max

Shape

Description automatically generated with medium confidence

* Min

Shape

Description automatically generated with medium confidence

Means and variances per group

Text

Description automatically generated

The output:

Text

Description automatically generated

# PCA & LDA

Standardising variables

Text

Description automatically generated

The output:

Text

Description automatically generated

## PCA

This method is called Principal Component Analysis (PCA). This method is based on the observation that the data are not normally distributed randomly in space but are often distributed near certain special lines/surfaces. PCA considers a special case where such special faces have linear form as subspaces. PCA is an unsupervised learning method, i.e. it only uses vectors describing the data and not the labels, if any, of the data.

Apply PCA using sklearn



Check the summary of PCA results

Text

Description automatically generated

The output:

Text

Description automatically generated

Check how many principal components to retain

Text

Description automatically generated

The output:

Chart, line chart

Description automatically generated

Calculate the values of the first principal component

Text

Description automatically generated

The output:

Graphical user interface

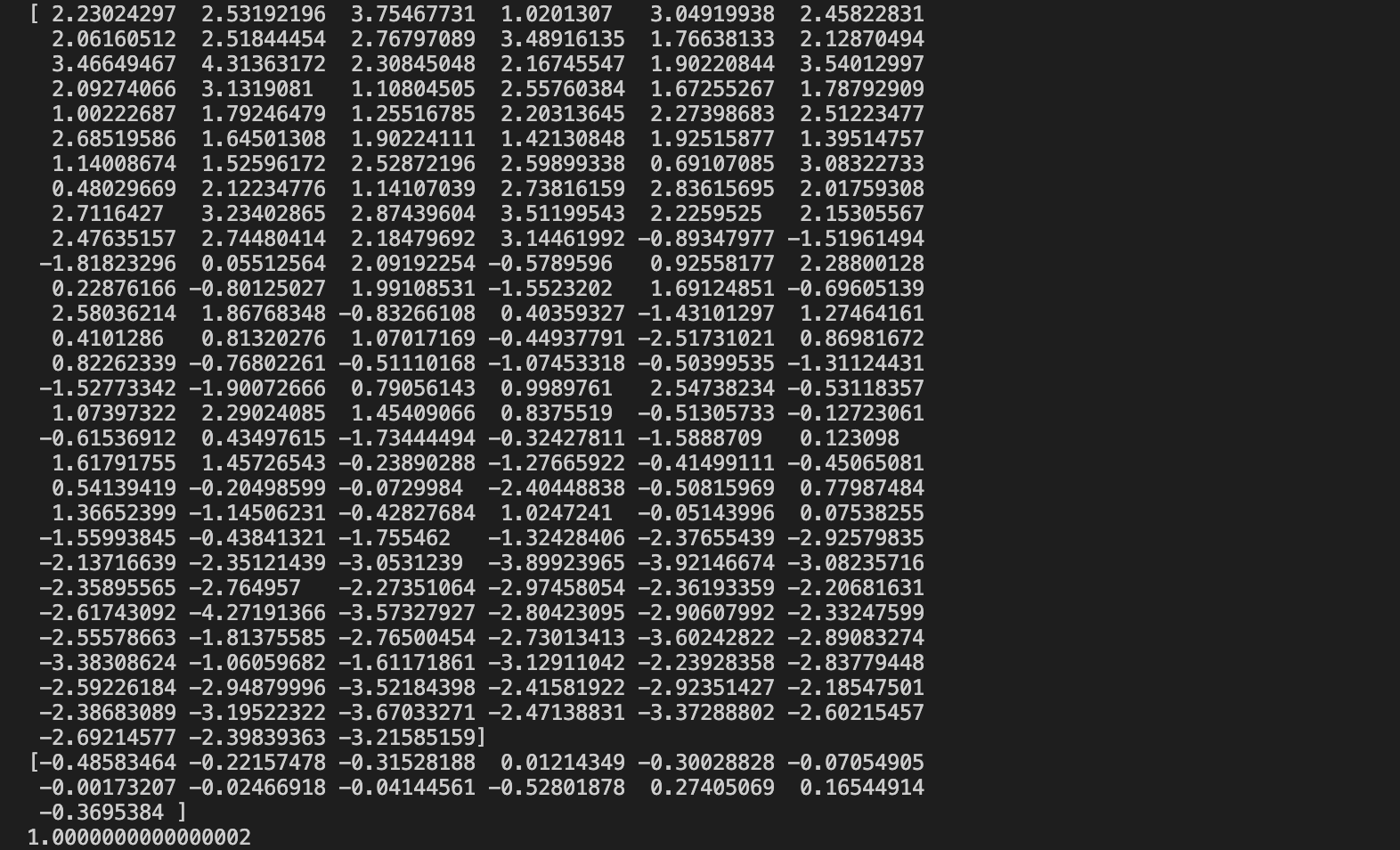
Description automatically generatedGraphical user interface, text

Description automatically generated

Obtain the loadings for the second principal component



The output:



Visualize scatterplots of the principal components

Text

Description automatically generated

The output:

Chart, scatter chart

Description automatically generated

Get mean, standard deviations, and sample sizes



The output:

Graphical user interface

Description automatically generated with medium confidence

## LDA

LDA is a data dimensionality reduction method for classification problems. LDA can be thought of as a method of dimensionality reduction, or it can be thought of as a classification method, or it can be applied simultaneously to both (i.e., reducing the data dimensionality so that the most efficient classification).

Apply LDA using sklearn (“main\_LDA\_2.py”)

Graphical user interface, text

Description automatically generated

The output:

Chart, scatter chart

Description automatically generated

Evaluation with Confusion matrix “main\_LDA\_1.py”

The output:

Graphical user interface

Description automatically generated with medium confidence

# REFERENCE

1. <https://machinelearningcoban.com/2017/06/15/pca/>
2. <https://machinelearningcoban.com/2017/06/30/lda/>
3. <https://github.com/Ayantika22/Linear-discriminant-Analysis-LDA-for-Wine-Dataset>
4. <https://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_vs_lda.html?fbclid=IwAR2YF0CQwa3DKx0GtolPwulVjoZwmcnxyRBuvTIR2dNNa7kFN6o5CtprhFg>