

**Assessment Cover Sheet**

This sheet must be signed by the student and attached to the assessment being submitted.

**IX726001 Data Science and Machine Intelligence**

**Assignment 2**

**Individual Assessment**

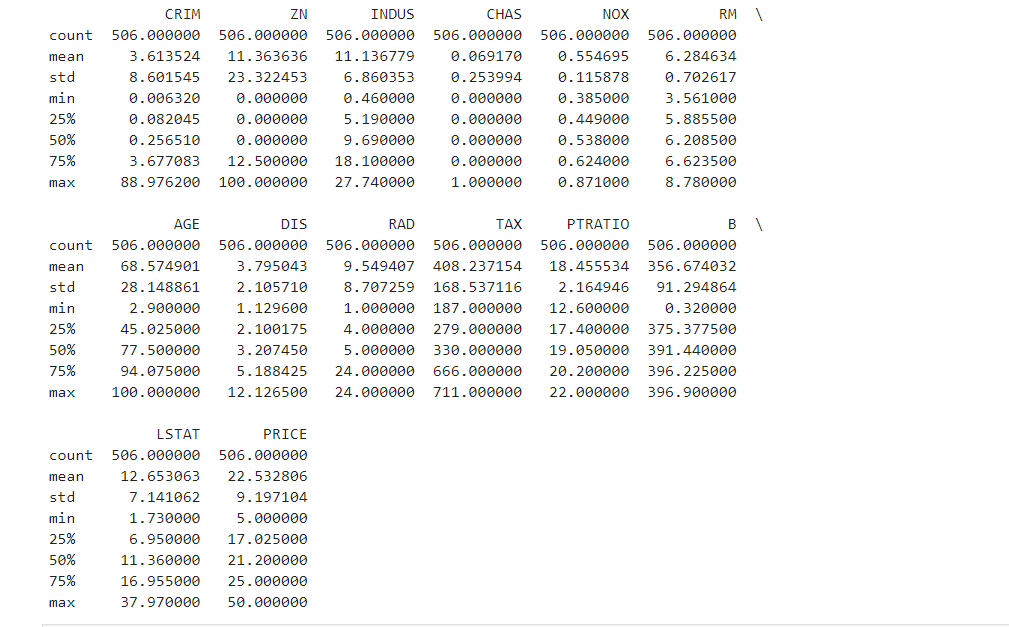
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Study block | Study Block 4 (August 2019) | | | | | |
| Date issued | 17/09/2019 | | | | | |
| Due date | 01/10/2019 | | | Time | 23.59 | |
| Delivery: | Submit via Moodle before deadline | | | | | |
| Weighting | 25% | | | | | |
| Marks out of | 30 | | | | | |
| Instructions | Complete this cover sheet and attach to your assignment.   * This assignment must be your own work. * Collusion, copying or plagiarism may result in disciplinary action. * We advise that you keep a copy of this assignment. * Refer to the following website for reference related resources:   http://www.cite.auckland.ac.nz/index.php?p=quickcite | | | | | |
| Student Name | Vimitaben Mukeshchandra Vaidya | | | | | |
| Student ID No | 1000063758 | Cohort number | | | |  |
| Lecturer | Farhad Mehdipour | Class time Tue & Thu 13.00-17.00 | | | |  |
| Student declaration: | I confirm that:  This is an original assessment and is entirely my own work.  Where ideas, tables, diagrams etc. of other writers have been used, I have acknowledged the source in every case.  This assignment has not been, nor will be, submitted as assessed work for any other academic course. | | | | | |
| Signature of student: | Vimita Vaidya | | Date : 3/10/19 | | |  |

**Task 1. Regularization**

**Action I. Separate dataset in X (attributes) and Y (target)**



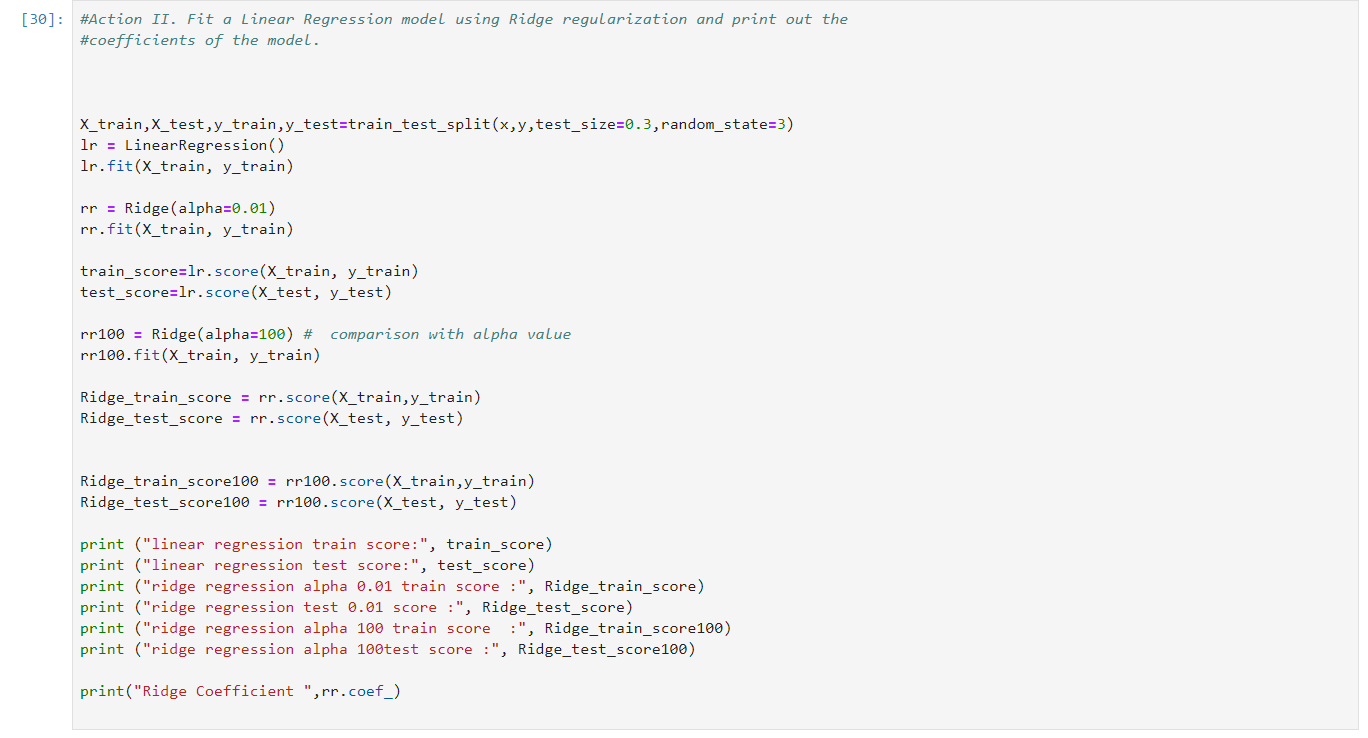
**Result**



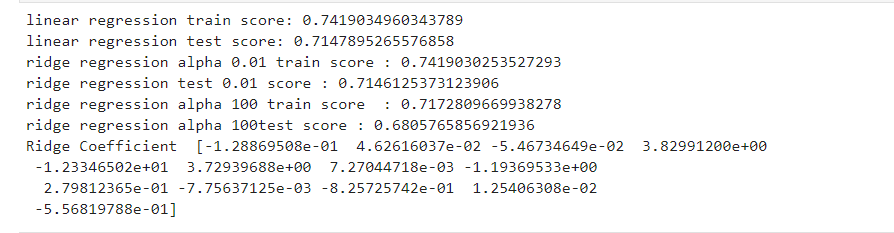
Background

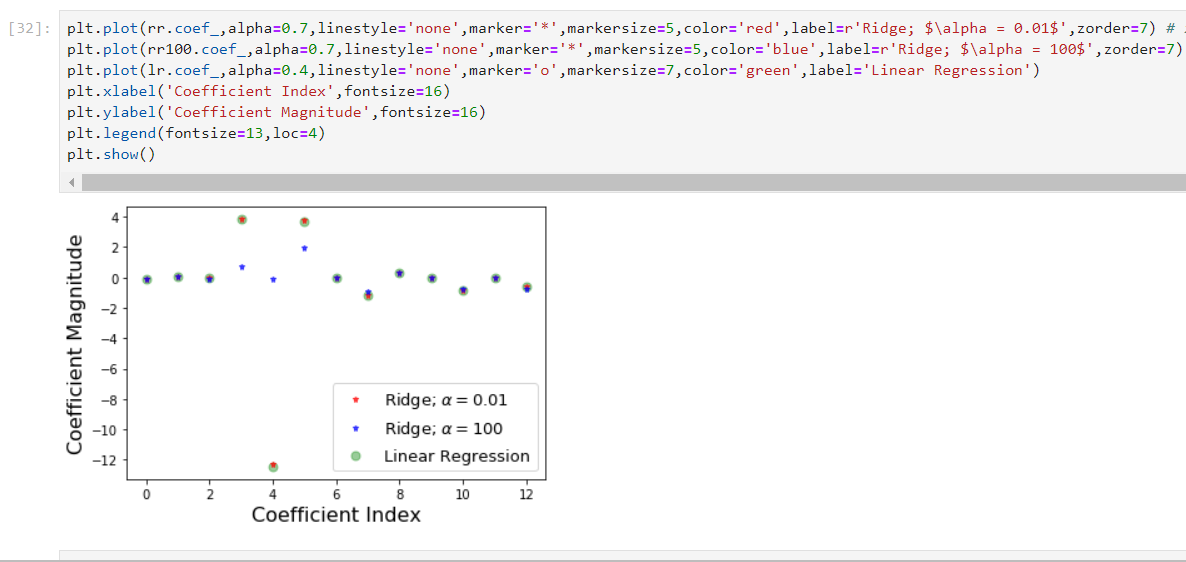
There are total 14 features of Boston data set. Where on X- axis define with different characteristics such as, CRIME, ZN, INDUS, CHAS,NOX,RM,AGE,DIS,RAD,TAX,PTRATIO,B,LSTAT, and on Y-axis I target PRICE.

**Action II. Fit a Linear Regression model using Ridge regularization and print out the coefficients of the model.**



**Result**

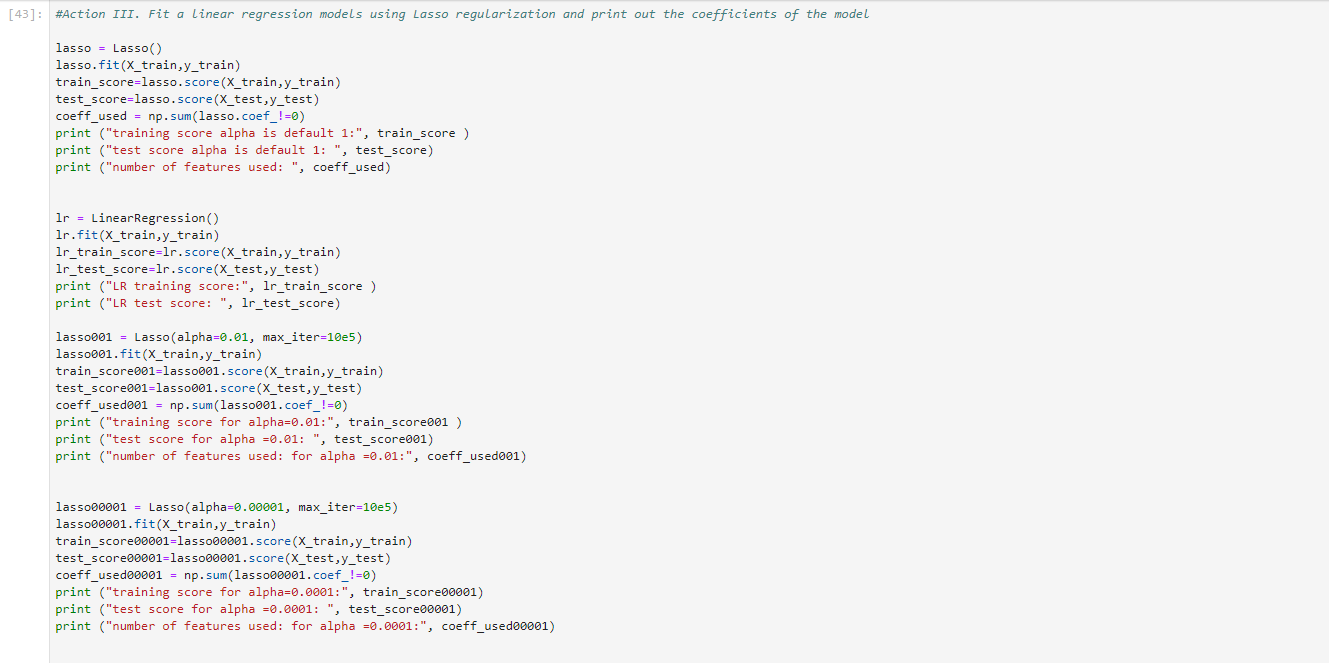




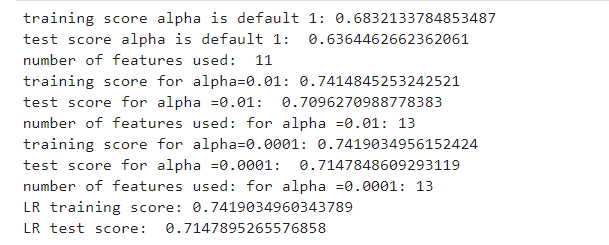
* As per the above figure, The coefficient index is plotted in X axis and there are total 13 columns of Boston dataset. For the small value of α (0.01), the value of the coefficients are nearly the same as those of linear regression when the coefficients are less constrained.
* For α (100) greater value, We see that in the above figure the magnitudes of coefficient indices 3,4,5 are significantly lower than the situation of linear regression.

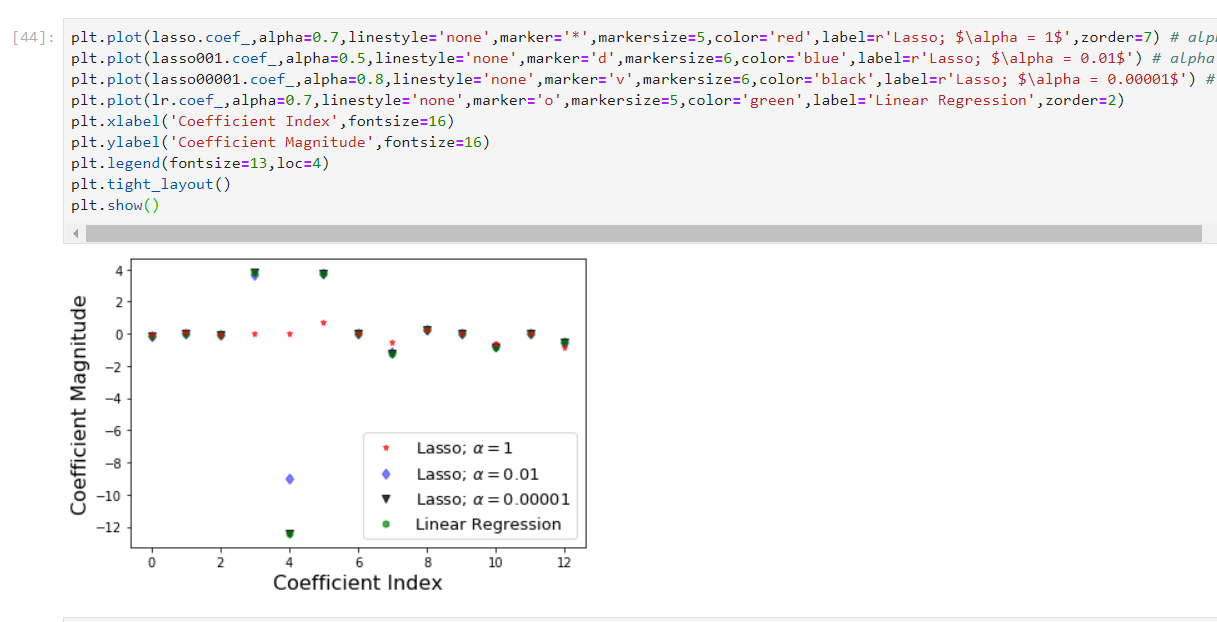
**Action III. Fit a linear regression models using Lasso regularization and print out the coefficients of the model.**

Here, I took the different alpha value 0.01 and 0.00001. which shows that the Lasso regression is not only prevents the over-fitting but it will also use in feature selection.



**Result**





* The default value for the Lasso regression regularization parameter (provided by α) is 1.

With this, only four characteristics (non-zero coefficient value) are used out of 30 attributes in the Boston data set.

* Both the training score and the test score (with only 11 characteristics) are small, conclude that the model is under fitting the data set for Boston. Reduce this issue by decreasing the amount of alpha iterations and increasing them.
* Now α = 0.01, non-zero characteristics = 13, improves the rating for practice and testing. For alpha =1, we can see most of the coefficients are zero or nearly zero, which is not the case for alpha=0.01.
* However, when α = 0.0001, has also use the same attributes as α = 0.01. But, both increases value result is the same as the Linear regression score.

**Action IV. Critically discuss the most striking difference between the coefficients of Ridge regression and Lasso regression.**

**Ridge Regularization:**

It reduces the parameters, so it is mostly used to avoid multicollinearity. It decreases the complexity of the model by reducing the coefficient.

**Lasso Regularization:**

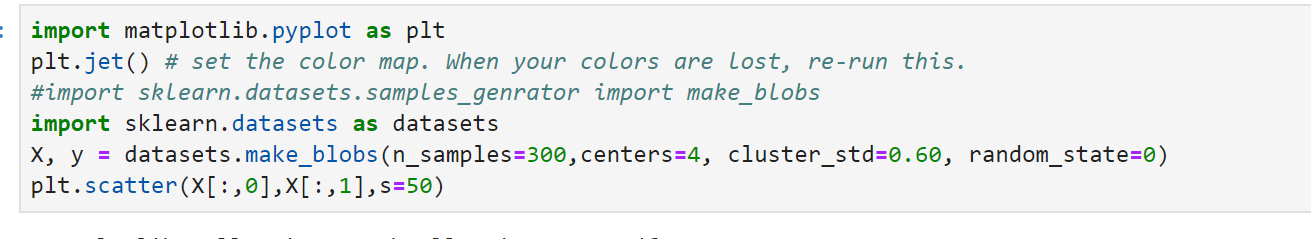
It is usually used when we have more characteristics, as it selects characteristics automatically.

Using lasso regression set the coefficient of the predictor variable age to zero, leading to a simpler model compared to the ridge regression, which include all predictor variables.

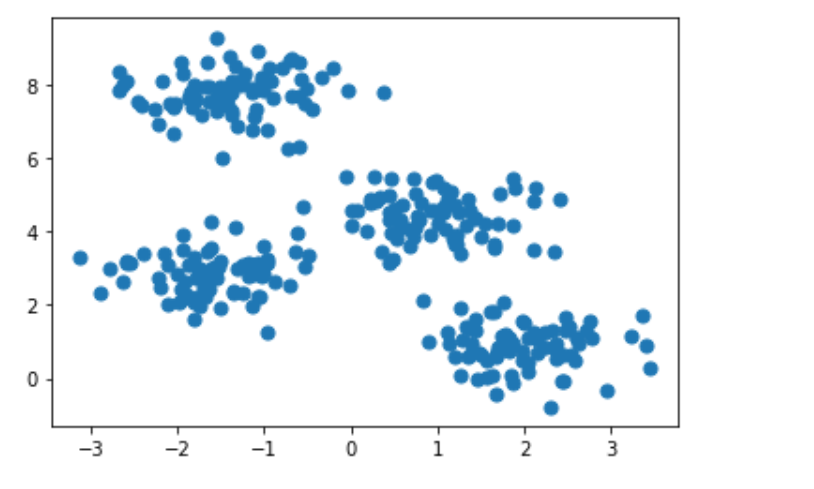
We can see that from above example, as we increased the value of alpha, coefficients were approaching towards zero, but if you see in case of lasso, even at smaller alpha’s, our coefficients are reducing to absolute zeroes. Therefore, lasso selects the only some feature while reduces the coefficients of others to zero. This property is known as feature selection and which is absent in case of ridge.

**Task 2. Data Clustering**

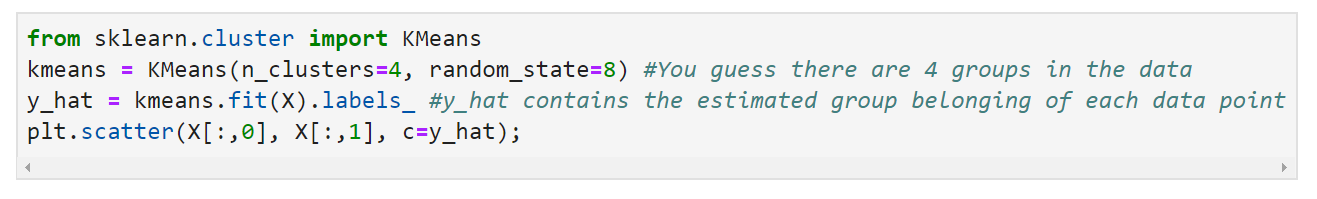
**Action I. Using the k-means clustering algorithm, perform the clustering using the K-means algorithm.**



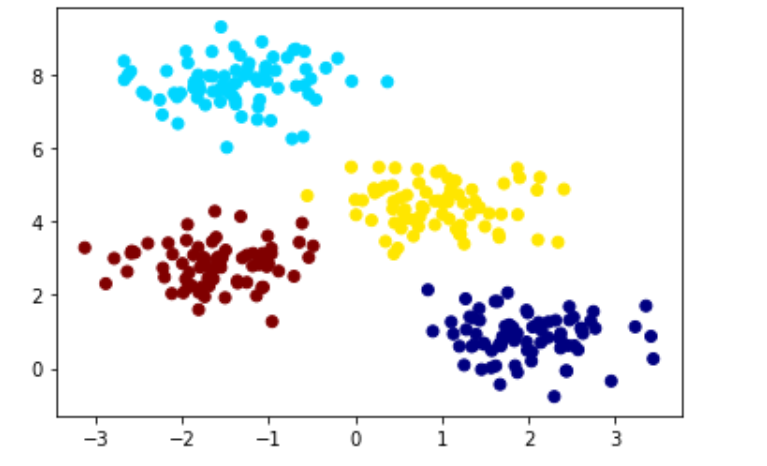
**Result**



**Action II. Plot the data for estimated clusters with different colour-coding.**



**Result**



**Task 3. Principal Component Analysis**

**Action I. Understand the dataset by using the information from the above link and briefly describe it.**

These data are 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.

Attributes:

1) Alcohol

2) Malic acid

3) Ash

4) Alcalinity of ash

5) Magnesium

6) Total phenols

7) Flavanoids

8) Nonflavanoid phenols

9) Proanthocyanins

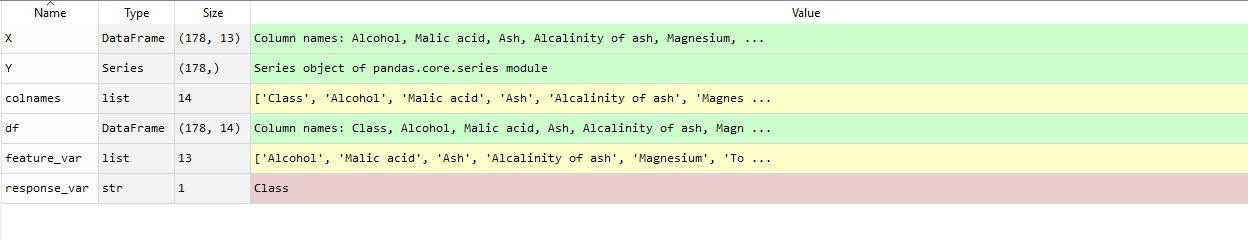
10)Color intensity

11)Hue

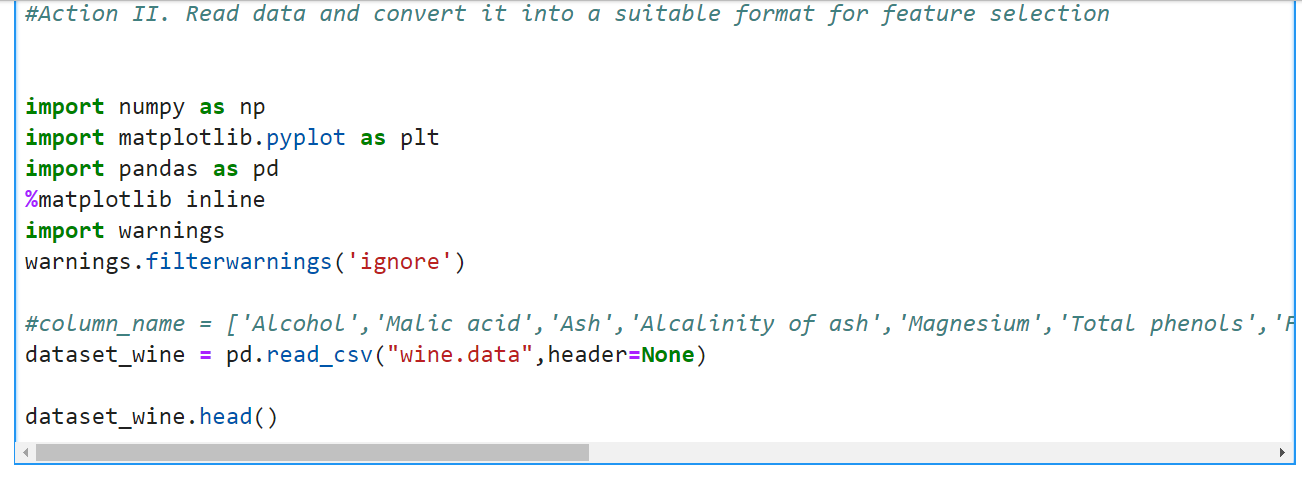
12)OD280/OD315 of diluted wines

13)Proline

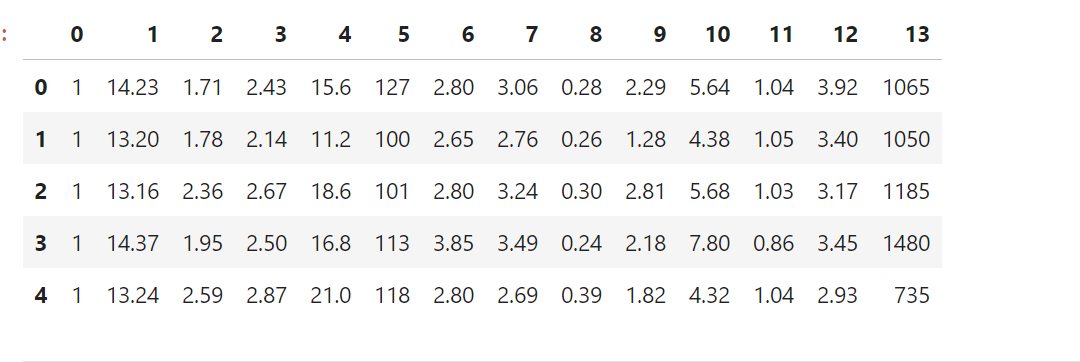
All attributes are continuous



**Action II. Read data and convert it into a suitable format for feature selection**



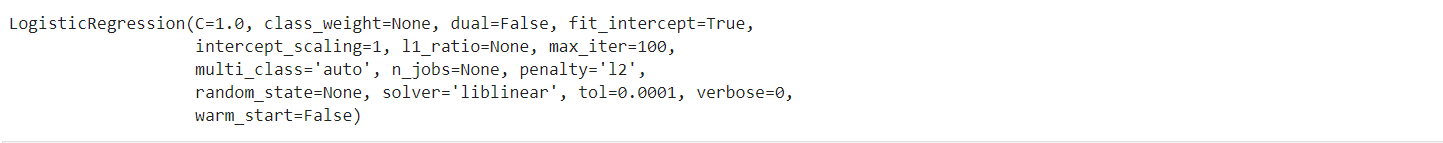
**Result**



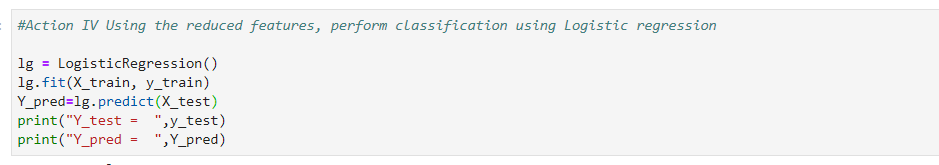
**Action III. Perform PCA-based feature selection analysis using *n\_components=2***



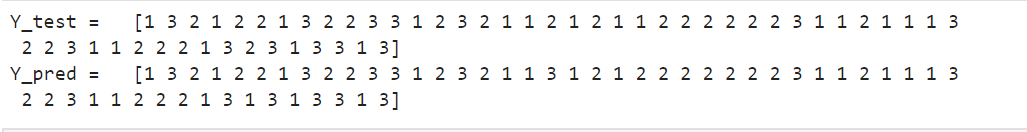
**Result**



**Action IV. Using the reduced features, perform classification using Logistic regression**



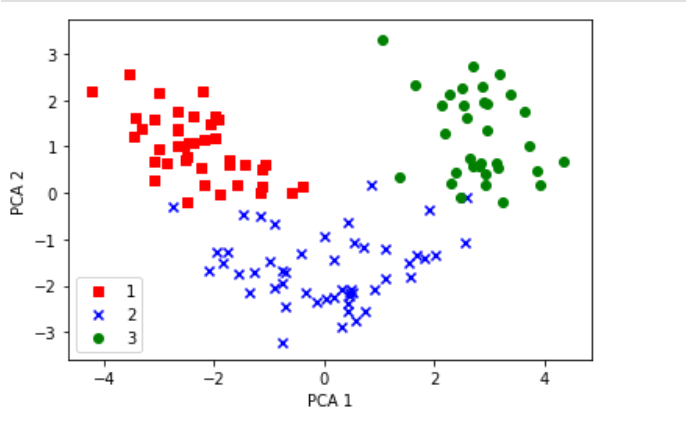
**Result**



**Action V. Plot the results in a scatter plot using different colour codes and discuss the results.**



**Result**



As we can see in below scatter plot, it shows that data are more spread on x- axis which is first principal component than second principal on y – axis. Which defined variance ratio.