

**DISCRIMINANT ANALYSIS OF THE SKULL DATA**

**Course No:** BAZG 524

**Student Name:** Anup Kumar Ray

**Student ID:** 2018MC21545

Table of contents

[[1] Introduction - 3 -](#_Toc24592080)

[[2] Data Description - 3 -](#_Toc24592081)

[[3] Data Analysis - 4 -](#_Toc24592082)

[[3.1] Assumptions - 4 -](#_Toc24592083)

[[3.2] Exploratory data analysis - 4 -](#_Toc24592084)

[[3.2.1] Descriptive Statistics - 4 -](#_Toc24592085)

[[3.2.2] Testing for multicollinearity - 5 -](#_Toc24592086)

[[3.2.3] Scatterplot Matrix and Histogram Distributions - 6 -](#_Toc24592087)

[[3.3] Eigen decomposition and projection of data into lower dimensional space - 8 -](#_Toc24592088)

[[3.3.1] Eigen Decomposition - 8 -](#_Toc24592089)

[[3.3.2] Transform data into lower dimensional space - 10 -](#_Toc24592090)

[[3.3.3] Testing the transformed data - 10 -](#_Toc24592091)

[[3.4] Discriminant Function Estimation - 11 -](#_Toc24592092)

[[4] Python Code - 12 -](#_Toc24592093)

# [1] Introduction

This report aims to present the analysis done for the given skull dataset. The dataset has 5 features and the entire data has two groups. This report presents the Linear Discriminant Analysis (LDA) details which is performed on this dataset. As part of this analysis, following two objectives are achieved:

1. Project the data with 5 features to a lower dimensional space (Dimensionality Reduction) without losing much of the information present in original dataset.
2. Define a discriminant function and a cut off score based on which a new data with the given 5 feature values could be classified into either of the groups.

The idea behind opting for LDA is from the fact that there are only two unique groups which the data falls into.

This analysis is performed using **Python** programming language with help of the packages like scikit-learn, matplotlib, seaborn, pandas, numpy etc. Analysis result along with the code is attached in the document in form of a PDF.

# [2] Data Description

|  |  |
| --- | --- |
| **Type of Data** | Multivariate |
| **Number of Attributes/Features** | 5 |
| **Number of groups/classes** | 2 |
| **Total Number of Observations** | 32 |

Following are the attribute details:

|  |  |
| --- | --- |
| **Attribute Name in Data** | **Description** |
| X1 | Greatest length of the skull |
| X2 | Greatest horizontal breadth of the skull |
| X3 | Height of the skull |
| X4 | Upper face height |
| X5 | Face breadth between outermost points of cheek bones |

Following are the group details:

|  |  |
| --- | --- |
| **Group Value** | **Group Name** |
| 1 | Group A |
| 2 | Group B |

# [3] Data Analysis

## [3.1] Assumptions

1. Unit of measurement for the attributes are same.
2. Data for each attribute comes from a Normal distribution.
3. Covariance matrices for the two groups are equivalent.

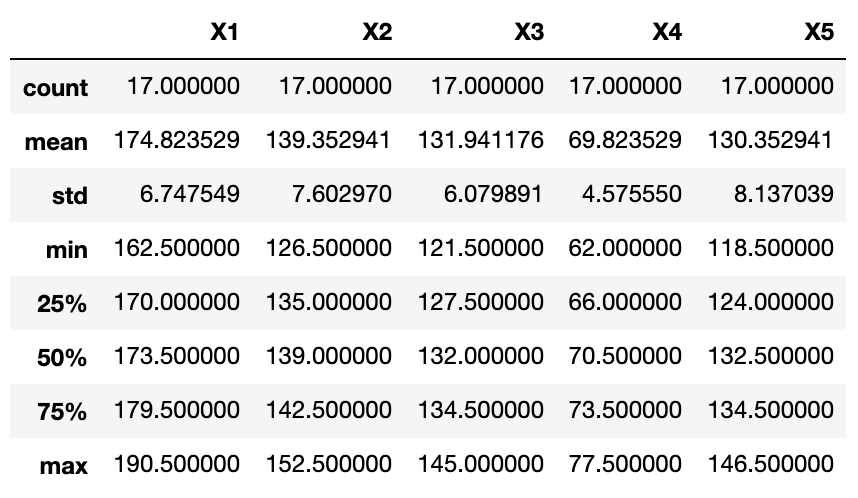
Assumptions ‘b’ & ‘c’ are simply made on the basis of the robustness of Fisher’s approach to deal with any minor departure from the needed assumptions.

## [3.2] Exploratory data analysis

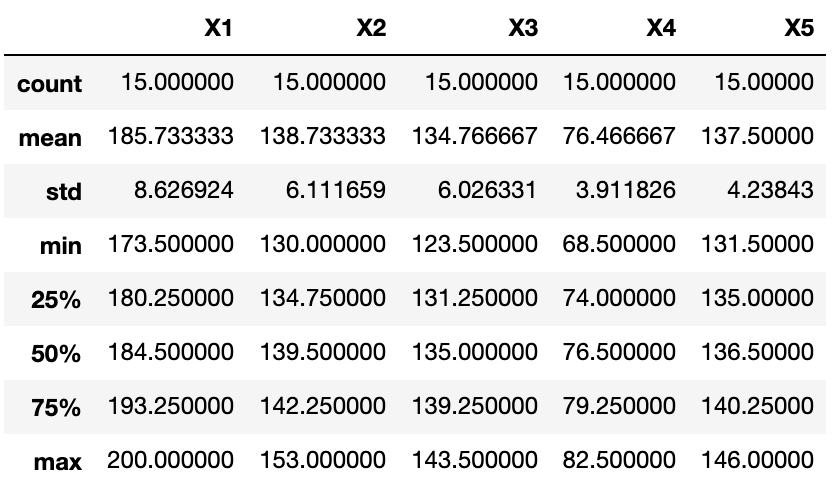
### [3.2.1] Descriptive Statistics

Following shows the descriptive statistics (mean, variations etc.) for each feature for each group.

**Group-1 Statistics:**



**Group-2 Statistics:**

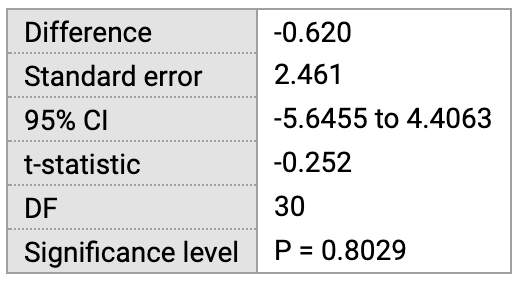


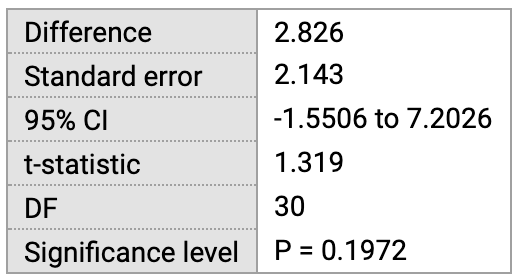
As observed from here, there seems to be good difference in means of the attributes for the two groups except for X2 & X3 attribute2 where the difference isn’t very large. To test whether the difference is significant we perform a **two-sample t-test for the means**, the result of which is given below:

Null Hypothesis (H0) 🡺 Two means are equal

Alternate Hypothesis (Ha) 🡺 Two means differ

Following is the result of the test for X2 & X3 respectively:

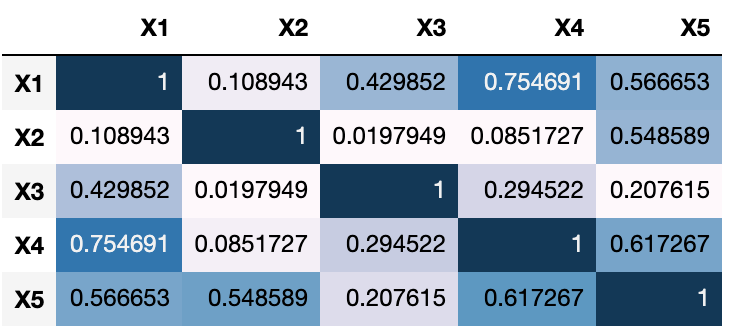




As observed here, since the *p-value > 0.5*, we can’t reject the null hypothesis and hence the difference in means between the two groups for these attributes don’t seem to be significant. Hence these two attributes will have very less impact on discrimination process implying a low weight/coefficient in the discriminant function.

### [3.2.2] Testing for multicollinearity

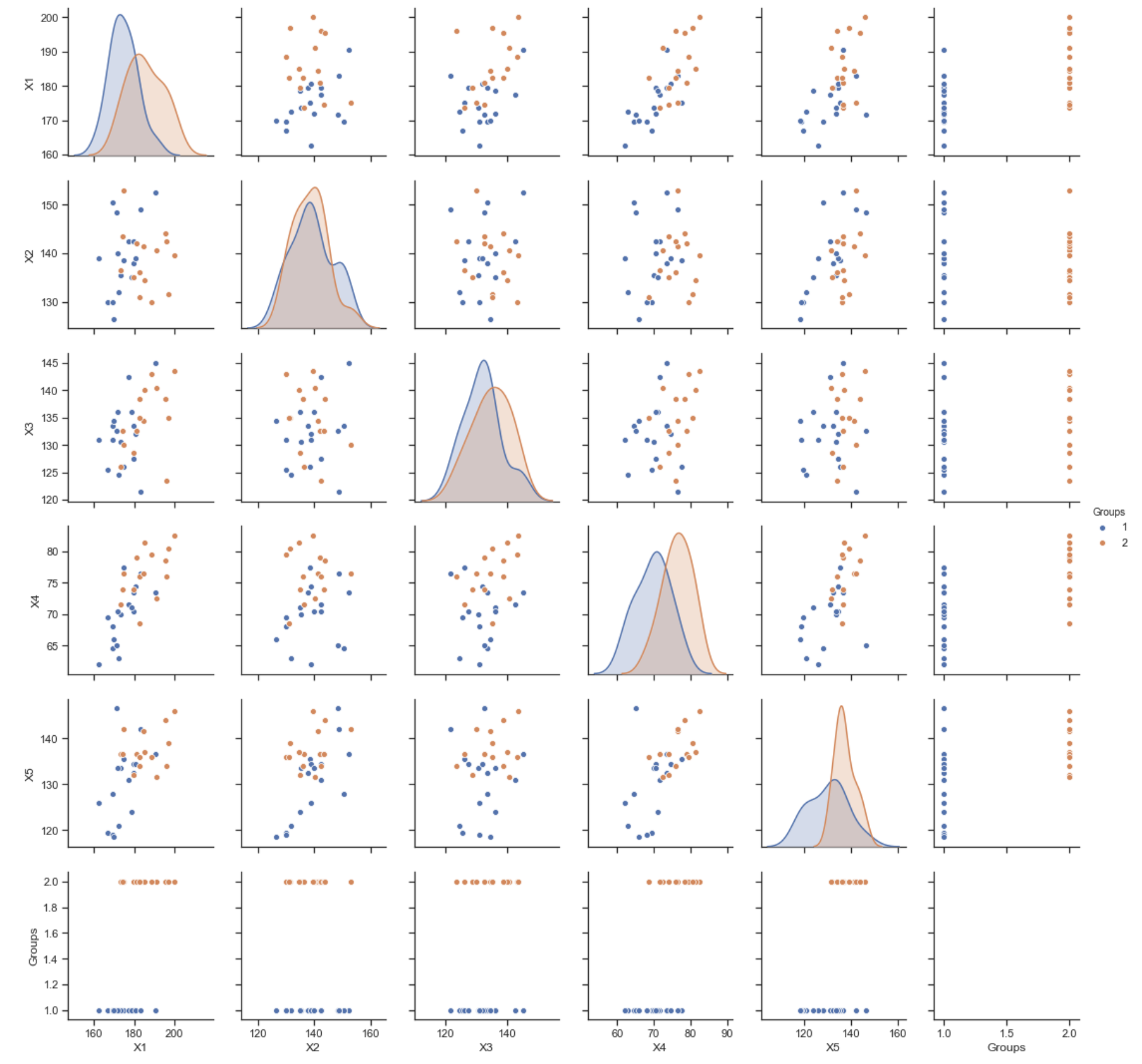
Following heatmap shows the correlation between the attributes for the complete dataset:



Though we see a high correlation of 0.754 between X1 & X4 which could lead to the issue of multi-collinearity, going by the general assumption that this will be a problem if the correlation exceeds 0.80, we can proceed with the analysis assuming no issue of multi-collinearity.

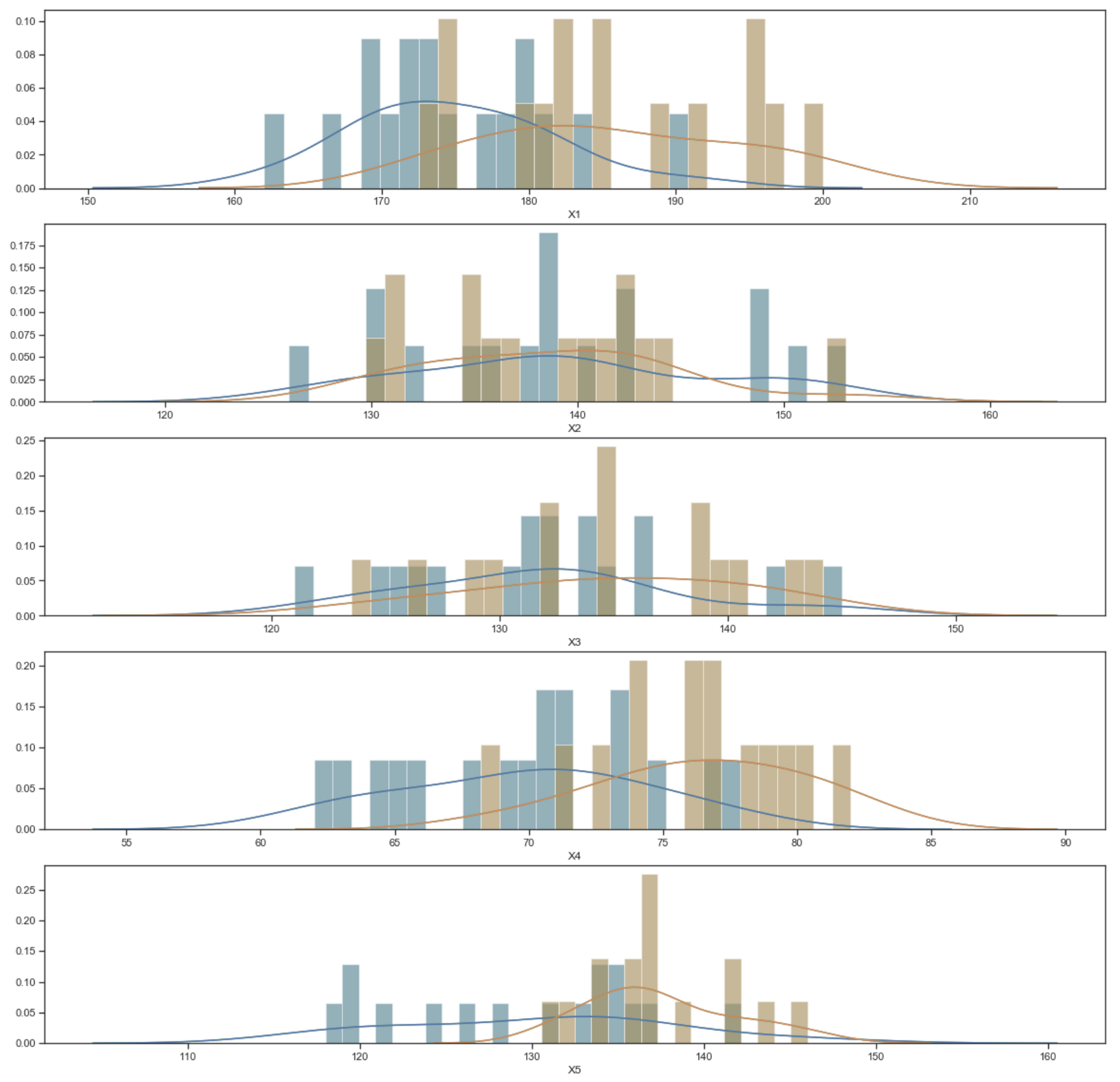
### [3.2.3] Scatterplot Matrix and Histogram Distributions

Following is the scatterplot matrix for the entire dataset:



As observed here, attributes ‘X1’,’X4’,X5’ seems to better discriminate between the groups and hence we are expecting better coefficients for these attributes in the discriminant function as compared to X2 & X3.

Following is the histogram plot for the entire data by attributes to have a rough idea of which attribute discriminate better between the groups:



As observed here, among the attributes, X5 seems to discriminate better between the groups compared to other attributes and hence this is expected to have a higher coefficient compared to other attributes.

Also, with the given approximate KDE plots on the histogram presented here, we can assume the attributes follows an approximately normal distribution.

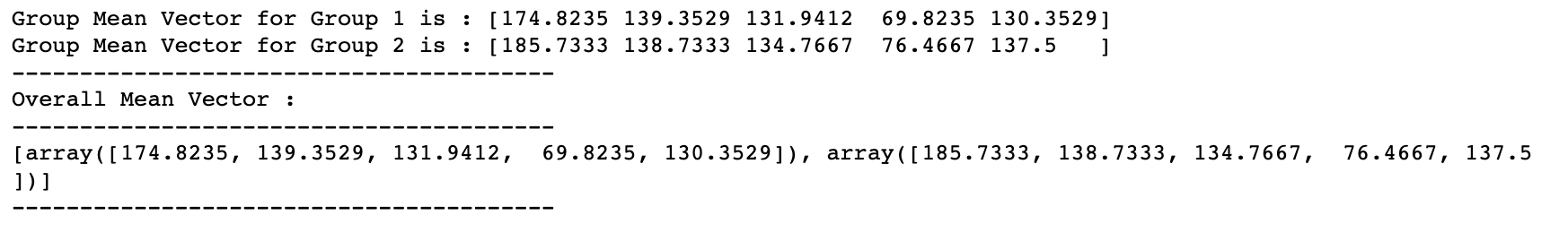
## [3.3] Eigen decomposition and projection of data into lower dimensional space

### [3.3.1] Eigen Decomposition

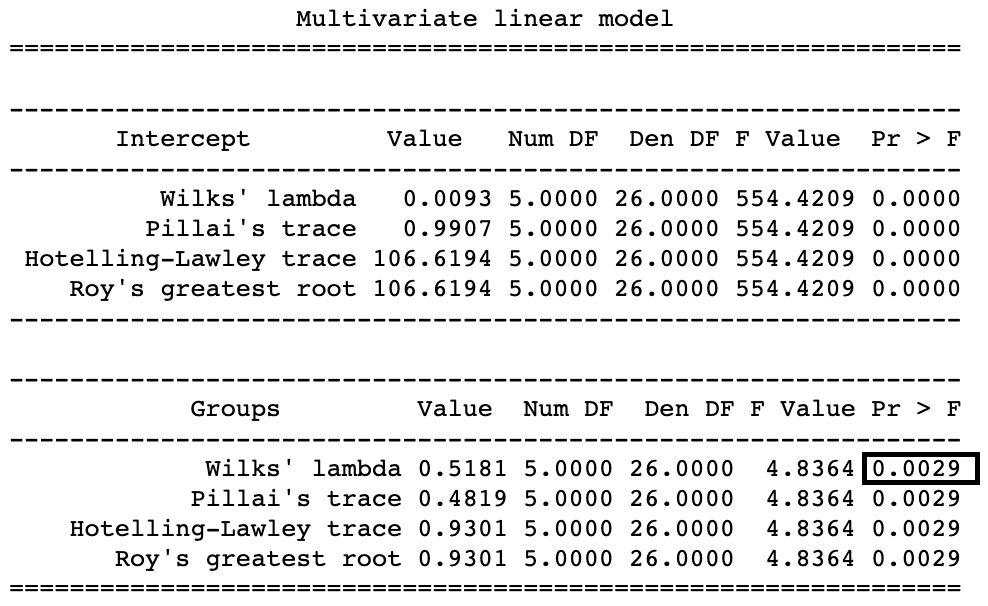
Eigen decomposition is performed to identify the eigen vector corresponding to the largest eigen value which will explain most of the variance in the data and thus reducing the dimensions to a lower dimensional space.

Following process is followed for eigen decomposition:

[**Step-1**] Get the group/class wise mean vectors for the attributes:

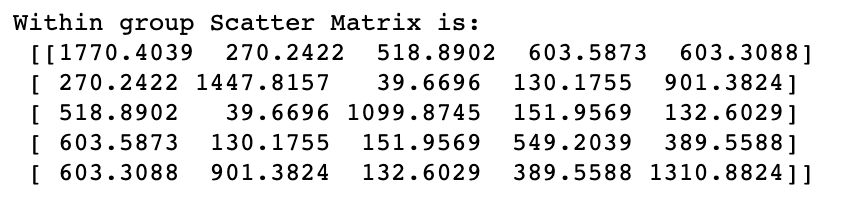


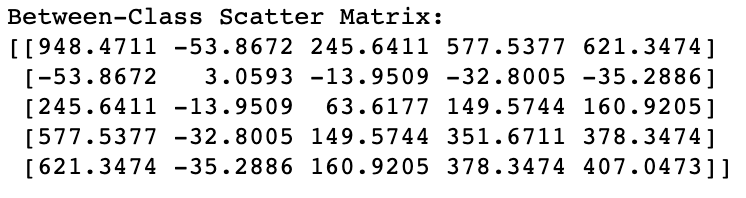
We apply Wilk’s Lambda test / MANOVA to test the null hypothesis that in the population, vectors of means of the 5 measurements for given 5 attributes don’t differ between the groups. Following is the test result:



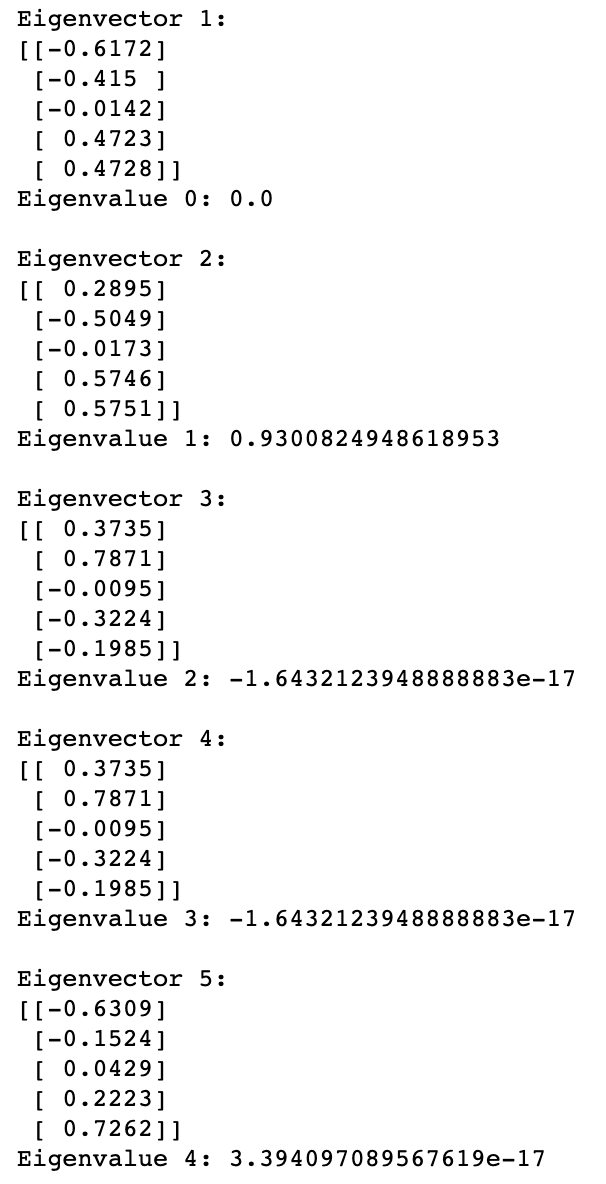
As is observed from the test, *p-value < 0.05*and hence we can reject null hypothesis. This means, we have significant difference between the mean vectors of the attributes between the groups. This indicates that LDA for this dataset makes sense.

[**Step-2**] Get the within group and between group scatter matrices:

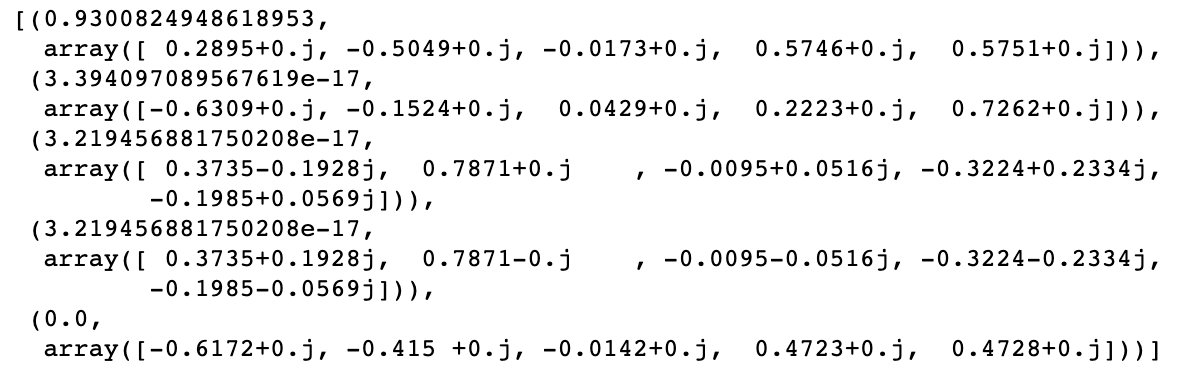


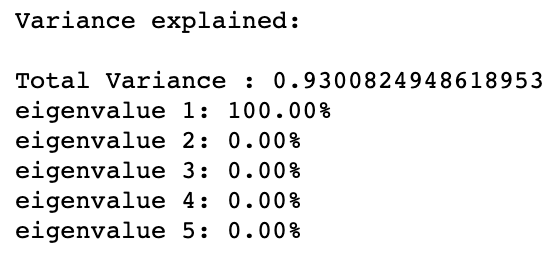


[**Step-3**] Solve for eigen vector and eigen value:



**[Step-4]** Sort the eigen values in descending order and calculate the explained variances:



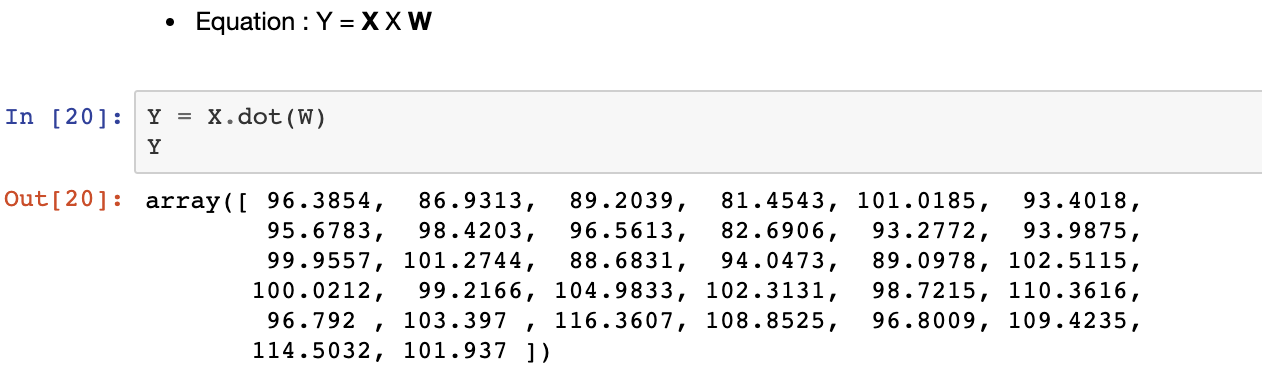


Here we see that only one eigen vector explains 100% variance which is an ideal case for data with 2 groups.

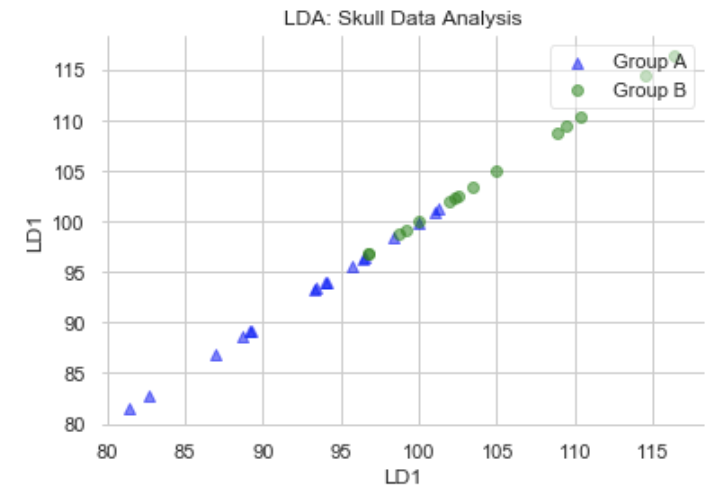
**[Step-5]** Choose the eigen vector for the largest eigen value:



### [3.3.2] Transform data into lower dimensional space



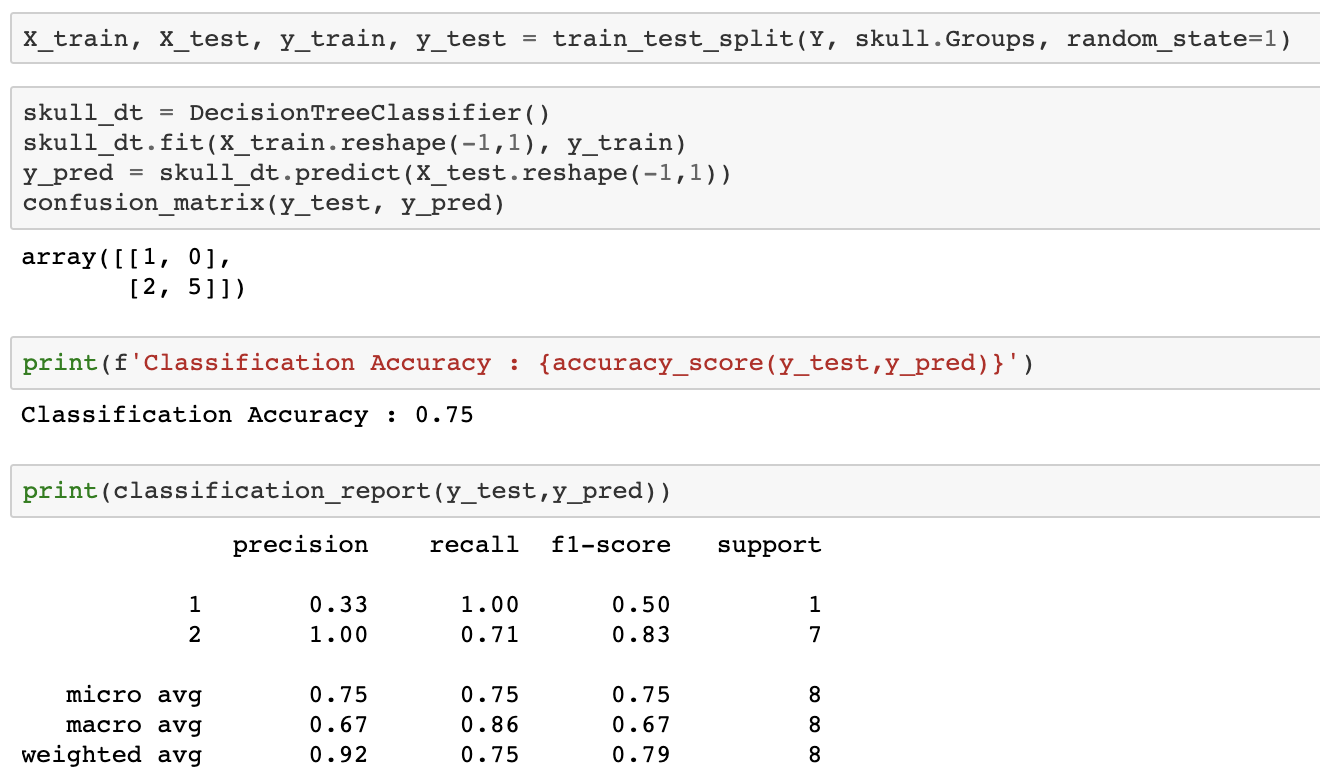
Above array is the transformed data into lower dimensional space along the direction of the determined eigen vector. Using this to plot for discrimination between groups yields following plot:



As observed here, by transforming data into lower dimensional space with eigen decomposition, we are able to nicely discriminate between the groups.

### [3.3.3] Testing the transformed data

I have used a Decision tree classifier from scikit-learn package from Python to test the accuracy of classification with the new transformed data above.

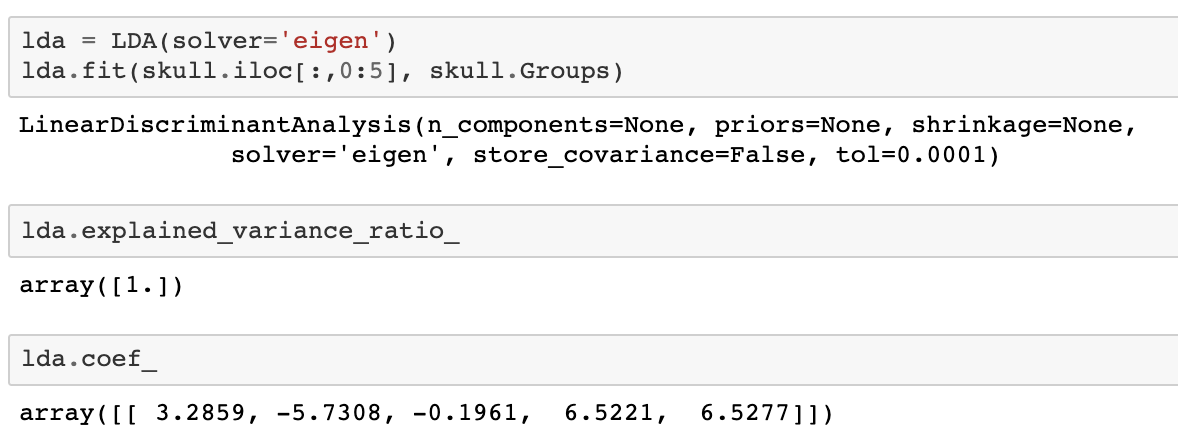


Here we get a 75% accuracy which is not very bad.

## [3.4] Discriminant Function Estimation

Here we use the linear discriminate function from scikit-learn package to estimate the discriminant function coefficients using which we will develop the discriminant function and estimate the cutoff score.

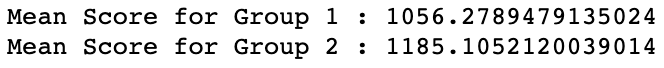
Following is the code and output coefficients:

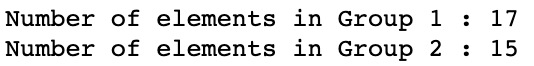


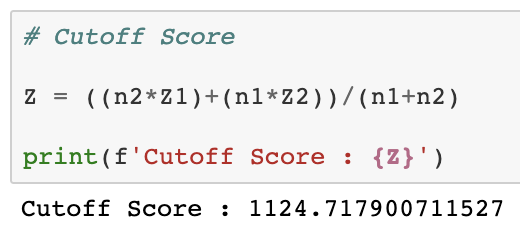
As was assumed earlier, we observe the coefficients for X2 & X3 to be negative which is less than the other coefficients since the difference in means for these attributes was not significant between the two groups. Also, we observe that the coefficient for X5 is highest as was guessed from the histogram plots.

Using these coefficients, following is the discriminant function:

Using this function, we estimate the average score for group 1 and group 2 and then the pooled cutoff score as below:







Hence for the new measurements, a score < 1124.7179 would result in the measurement to be classified as Group 1 and score > 1124.7179 would result in measurement to be classified in Group 2.

Coefficients determined above indicates how much the function value will change for one unit change in the respective attribute and hence also are considered to be weights for the attributes in the discriminant function.

Following is the Confusion matrix with the given score:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Predicted Group** | | | |
| **Actual Group** |  | **1** | **2** | **Total** |
| **1** | 14 | 3 | **17** |
| **2** | 3 | 12 | **15** |
| **Total** | **17** | **15** | **32** |

Here we see 82.3% accuracy for group 1 and 80% accuracy for group 2 based on the cut-off score. Hence this cut-off score can be used to discriminate the new observations between group-1 and group-2.

# [4] Python Code

The developed Python code can be referred from my Github links below:

|  |  |
| --- | --- |
| **Document** | **Link** |
| Jupyter Notebook | <https://github.com/anupray/BAZG524-Skull-Data-Analysis/blob/master/2018MC21545_Skull%20Data%20Analysis.ipynb> |
| PDF Document | <https://github.com/anupray/BAZG524-Skull-Data-Analysis/blob/master/2018MC21545_Skull%20Data%20Analysis.pdf> |