# Topic: Dimension Reduction With PCA

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: DHEERAJ MISHRA Batch ID:** DS\_ 01072021

**Topic: Principal Component Analysis**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline

(OR)

* + Less than 30% of problems in the assignments are submitted before deadline
* **Grade F: (< 50):** No submission (or) malpractice.

**Hints:**

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary.**
   2. **Univariate analysis.**
   3. **Bivariate analysis.**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options).**

**5.2 Perform PCA analysis and get the maximum variance between components.**

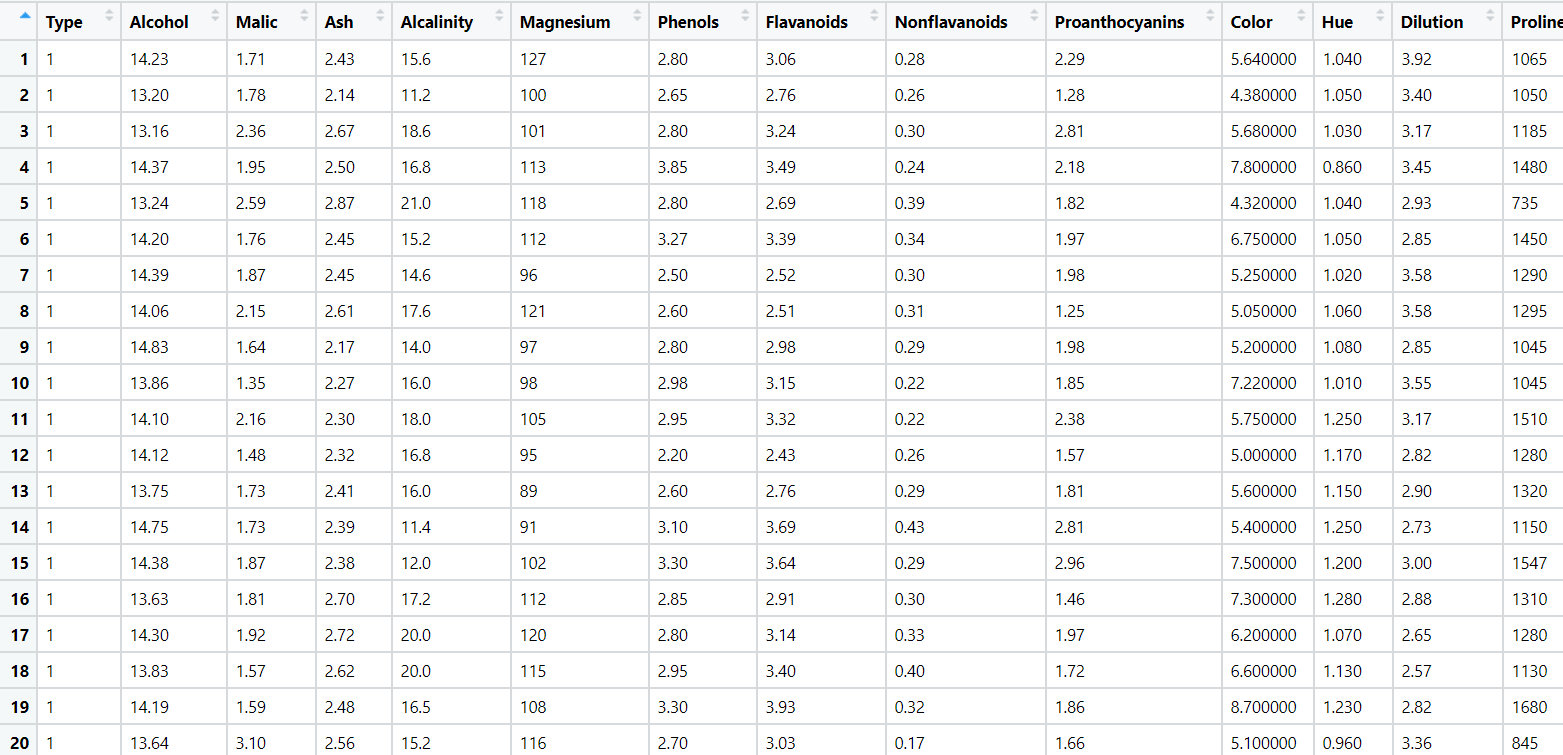
**5.3 Perform clustering before and after applying PCA to cross the number of clusters formed.**

**5.4 Briefly explain the model output in the documentation.**

1. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Problem Statement: -**

Perform hierarchical and K-means clustering on the dataset. After that, perform PCA on the dataset and extract the first 3 principal components and make a new dataset with these 3 principal components as the columns. Now, on this new dataset, perform hierarchical and K-means clustering. Compare the results of clustering on the original dataset and clustering on the principal components dataset (use the scree plot technique to obtain the optimum number of clusters in K-means clustering and check if you’re getting similar results with and without PCA).



1. **Business objective :**

**Max:- Information in minimum columns**

**Constraints:- Lack of data**

1. Data understanding :

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Types | Relevant |
| Type | Type of wine | Discrete , nominal | Not relevant |
| Alcohol | Amount of alcohol present | Continuous | Relevant |
| Malic | Amount of malic acid present | Continuous | Relevant |
| Ash | Amount of ash present | Continuous | Relevant |
| Alcalinity | Property to neutralize acid | Continuous | Relevant |
| Magnesium | Amount of magnesium present | Continuous | Relevant |
| Phenols | Amount of phenols present | Continuous | Relevant |
| Flavonoids | Amount of flavonoids present | Continuous | Relevant |
| Nonflavonoids | Amount of nonflavonoids present | Continuous | Relevant |
| Proanthocyanis | Amount of proanthocyanis | Continuous | Relevant |
| Color | Color amount | Continuous | Relevant |
| Hue | Amount of hue present | Continuous | Relevant |
| Dilution | Amount of water added | Continuous | Relevant |
| Proline | Amount of protein present | Discrete | Relevant |

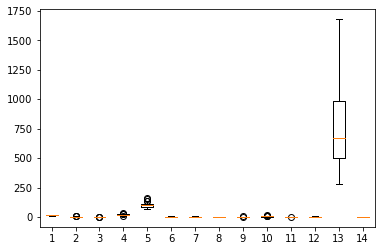
1. EDA

Summary:-

1. Calculated mean, median , sd, quartiles ,min and max values for each feature
2. No null values found in each feature
3. Duplicate row does not exists
4. All features are of type int64 and flot64
5. Dataset consists of 178 rows and 14 columns

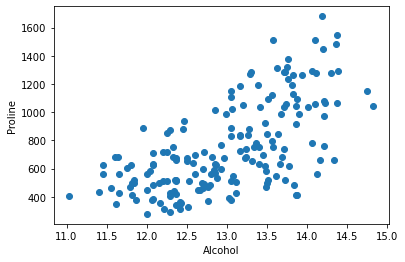
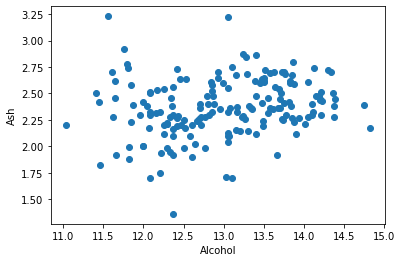
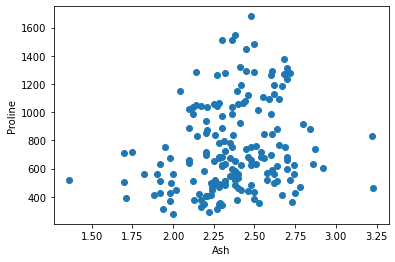
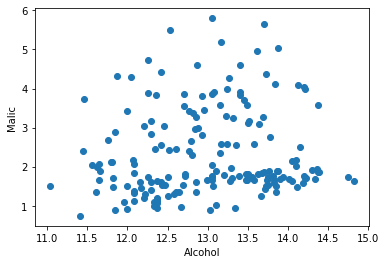
Univariate analysis :-

1. From boxplot most features does not have normal distribution
2. From box plot most of features are right skewed
3. From box plot outliers detected

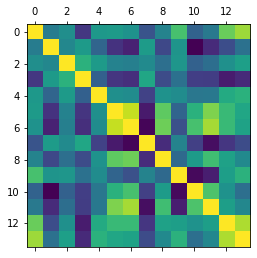


Bivariate analysis :-

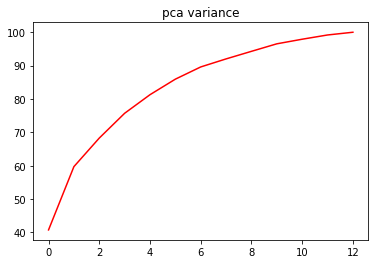
1. From scatter plots observed that there is no linear relationship between any features .
2. From scatter plot some clusters are recognized for different features

1. From mat plot weak correlations for different features

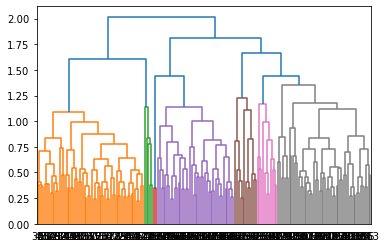
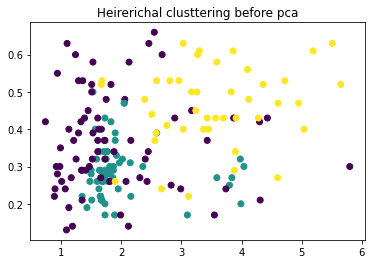


1. Data preprocessing:-
2. Type feature does not provide useful information so dropped it
3. Outliers detected but they are retain for better result
4. Normalization technique used for scaling dataset for better performance
5. Model building :-
6. Model builded on scaled data set using normalization technique
7. From plot we can observe pca variance

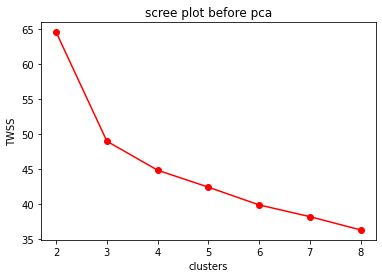
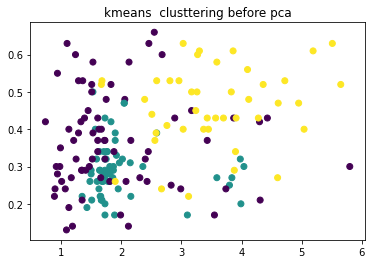


Clustering before PCA

1. Hierarchal clustering

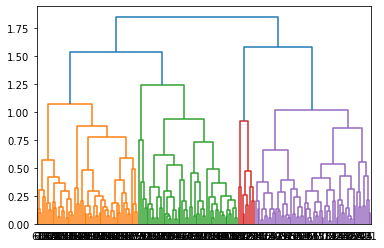
 

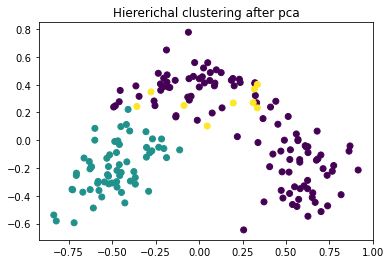
1. Kmeans clustering

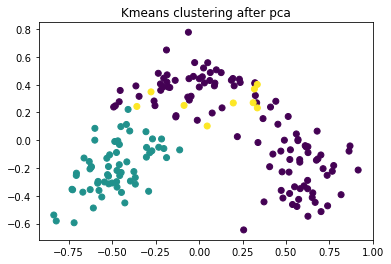
Clustering after PCA

1. Hierarchal clustering





1. Kmeans clustering

1. Output :- from var value first three columns contains 69% of information . clustering values are same for Kmeans and Hierarchal clustering
2. Impact :-

From above information we can conclude that from 14 columns 69% of information are there in three columns . Also we are getting clusters without correlation after PCA.

**Problem Statement: -**

A pharmaceuticals manufacturing company is conducting a study on a new medicine to treat heart diseases. The company has gathered data from its secondary sources and would like you to provide high level analytical insights on the data. Its aim is to segregate patients depending on their age group and other factors given in the data. Perform PCA and clustering algorithms on the dataset and check if the clusters formed before and after PCA are the same and provide a brief report on your model. You can also explore more ways to improve your model.

Note: This is just a snapshot of the data. The datasets can be downloaded from AiSpry LMS in the Hands-On Material section. A screenshot of a cell phone

Description automatically generated

1. **Business objective :**

**Max:- Information in minimum columns**

**Constraints:- Lack of data**

1. Data understanding :

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Types | Relevant |
| Age | Age of the individual | Discrete | Relevant |
| Sex | Gender of the individual | Discrete | Relevant |
| Cp | Type of chest-pain | Discrete | Relevant |
| trestbps | Resting blood pressure value | Discrete | Relevant |
| Chol | Serum cholesterol | Discrete | Relevant |
| Fbs | Fasting blood sugar value | Discrete | Relevant |
| restecg | Resting electrocardiographic results | Discrete | Relevant |
| thalach | The max heart rate achieved | Discrete | Relevant |
| exang | **Exercise induced angina** | Discrete | Relevant |
| Oldpeak | **Depression induced by exercise** | Continuous | Relevant |
| slope | **Peak exercise ST segment** | Discrete | Relevant |
| Ca | **Number of major vessels** | Discrete | Relevant |
| Thal | Displays the thalassemia | Discrete | Relevant |
| Target | Suffering from heart disease or not : | Discrete | Relevant |

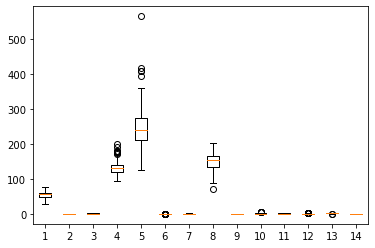
1. EDA

Summary:-

1. Calculated mean, median , sd, quartiles ,min and max values for each feature
2. No null values found in each feature
3. Duplicate row exists
4. All features are of type int64 and flot64
5. Dataset consists of 303 rows and 14 columns

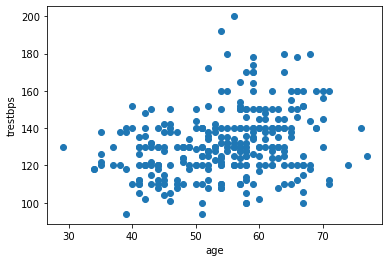
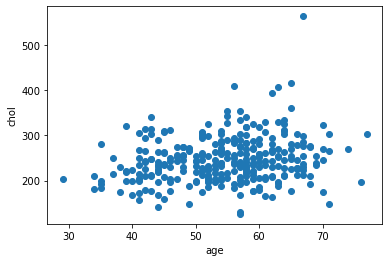
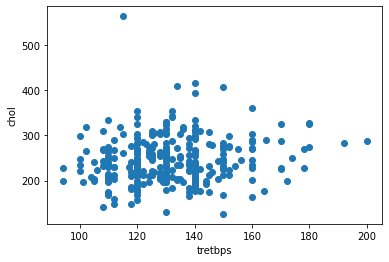
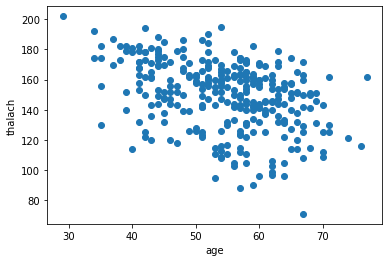
Univariate analysis :-

1. From boxplot most features does not have normal distribution
2. From box plot most of features are right skewed
3. From box plot outliers detected

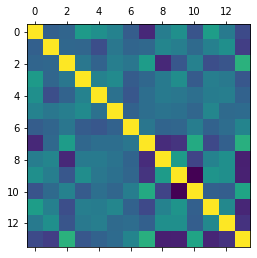


Bivariate analysis :-

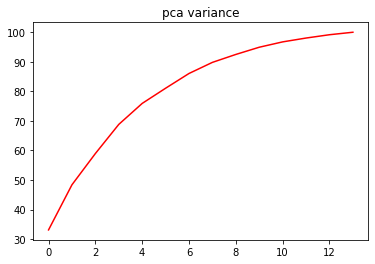
1. From scatter plots observed that there is no linear relationship between any features .
2. From scatter plot some clusters are recognized for different features



1. From mat plot weak correlations for different features

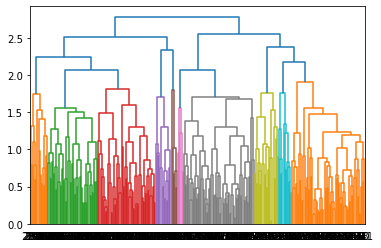


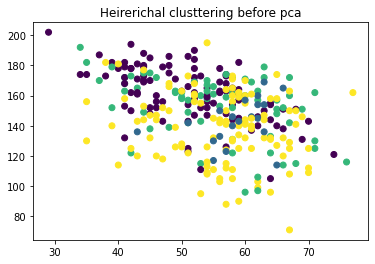
1. Data preprocessing:-
2. Duplicate row found and removed
3. Outliers detected but they are retain for better result
4. Normalization technique used for scaling dataset for better performance
5. Model building :-
6. Model builded on scaled data set using normalization technique
7. From plot we can observe pca variance



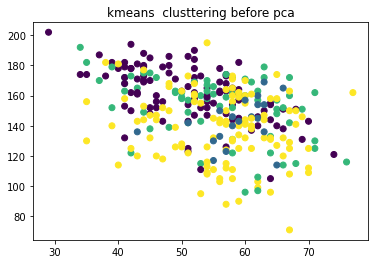
Clustering before PCA

1. Hierarchal clustering



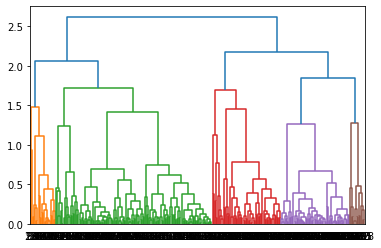


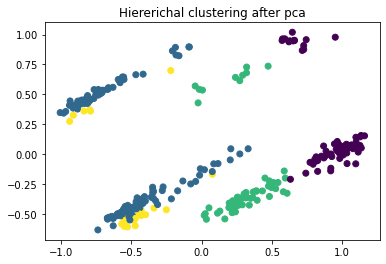
1. Kmeans clustering

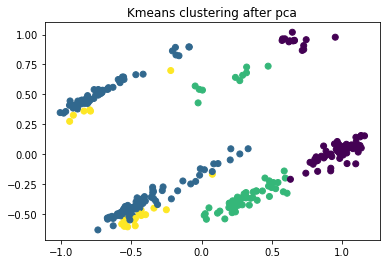
Clustering after PCA

1. Hierarchal clustering





1. Kmeans clustering

1. Output :- from var value first four columns contains 69% of information . clustering values are same for Kmeans and Hierarchal clustering
2. Impact :-

From above information we can conclude that from 14 columns 69% of information are there in four columns . Also we are getting clusters without correlation after PCA.