# Topic: Dimension Reduction (PCA)

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Shahina Athar**

**Batch Id:** 10122020

**Topic: Principal Component Analysis**

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
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, if not relevant provide reasons and provide description of the feature.**

**Using R and Python codes perform:**

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)**
   2. **Perform PCA analysis and get the maximum variance between components**
   3. **Perform clustering before and after applying PCA to cross the number of clusters formed.**
   4. **Briefly explain the model output in the documentation.**

1. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

# Note:

The assignment should be submitted in the following format:

* R code
* Python code
* Code Modularization should be maintained
* Documentation of the model building (elaborating on steps mentioned above)

**Problem Statement: -**

Perform Principal component analysis and perform clustering using first 3 principal component scores (both Hierarchical & K-Mean clustering). Use Scree plot or elbow curve and obtain optimum number of clusters and check whether we have obtained same number of clusters with the original data

**1.Business Problem:**

* 1. **Objective:** Using PCA we will find the Quality of wine

**2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Types** | **Relevance** |
| Type | Type of wine  (1,2,3) | Quantitative, Discrete, Ordinal | Irrelevant |
| Alcohol | Alcohol Percentage | Quantitative, Continuous,  Ratio | Relevant |
| Malic | Malic Percentage | Quantitative, Continuous,  Ratio | Relevant |
| Ash | Ash Percentage | Quantitative, Continuous,  Ratio | Relevant |
| Alcalinity | Alcalinity Percentage | Quantitative, Continuous,  Ratio | Relevant |
| Magnesium | Number of Amount Magnesium | Quantitative, Discrete,  Count | Relevant |
| Phenols | Phenols Percentage | Quantitative, Continuous,  Ratio | Relevant |
| Flavanoids | Flavanoids Percentage | Quantitative, Continuous,  Ratio | Relevant |
| Nonflavanoids | Nonflavanoids Percentage | Quantitative, Continuous,  Ratio | Relevant |
| Proanthocyanins | Proanthocyanins Percentage | Quantitative, Continuous,  Ratio | Relevant |
| Color | Color Percentage | Quantitative, Continuous,  Ratio | Relevant |
| Hue | Hue Percentage | Quantitative, Continuous,  Ratio | Relevant |
| Dilution | Dilution Percentage | Quantitative, Continuous,  Ratio | Relevant |
| Proline | Number of Amount  Proline | Quantitative, Discrete,  Count | Relevant |

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

**Answer: When we apply Hirarchical Clustering and K-Means Clustering(before PCA) we see that we can easily get clusterings and we get 3 clusters in both the methods and there is a lot of time required for computation, after applying PCA we select 3 PC’s and with these 3 columns, again we apply Hirarchical and K-Means Clustering and it gives same number of 3 clusters like before & reduces the time required for computation, what we observe that there some negative values also exist in this PC’s and that we obtain 3 clusters easily with these 3 clusters we can easily observe and conclude the quality of wine and also we reduce dimentions (columns) and It improves the model performance.**

**4.Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

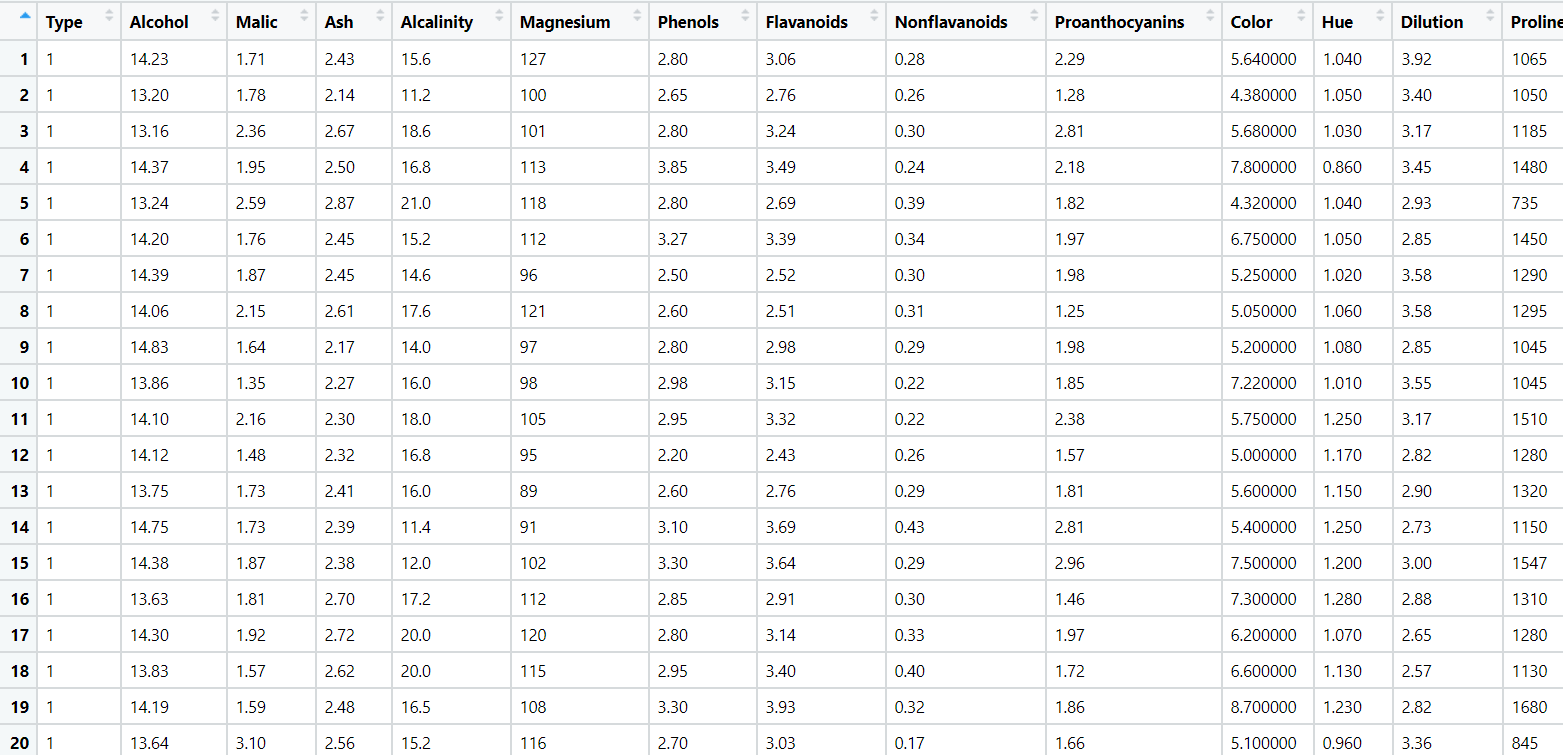
**Answer:**

**1) The first benefit is we reduce the dimension (columns) from all these 14 columns we extract 3 columns means we extract maximum information and after that it’s easy to perform Hirarchical and K-Means Clustering and we can see 3 clusters, with these 3 clusters we can easily observe and conclude the quality of wine.**

**2) Second benefits it reduces the time required for computation since less dimensions require less computation.**

**3) It eliminates the redundant features.**

**4) It improves the model performance.**



**Problem Statement: -**

A Pharmaceutical drug manufacturing company is studying on a new medicine to treat Heart diseases, it has gathered data from its secondary sources, and it 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 as given in the data, perform PCA and Clustering Machine learning Algorithm on the dataset given, and check if the clusters formed before and after PCA are same and provide a brief report on your model. You can also explore more on ways to improve your model.

Note: - This is just a snap shot 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 Problem:**

* 1. **Objective:** A Pharmaceutical drug manufacturing company is studying on a new medicine to treat Heart diseases, it has gathered data from its secondary sources our aim is to segregate patients depending on their age group and other factors as given in the data, perform PCA and Clustering Machine learning Algorithm on the dataset.

**2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Types** | **Relevance** |
| age | displays the age of the individual. | Quantitative, Continuous, Ratio | Relevant |
| sex | displays the gender of the individual | Quantitative, Discrete,  Binary | Relevant |
| cp | displays the type of chest-pain experienced by the individual | Quantitative, Discrete,  Ordinal | Relevant |
| trestbps | resting blood pressure (in mm Hg on admission to the hospital) | Quantitative, Discrete,  Count | Relevant |
| chol | displays the serum cholesterol in mg/dl (unit) | Quantitative, Discrete,  Count | Relevant |
| fbs | compares the fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then : 1 (true) else : 0 (false) | Quantitative, Discrete,  Binary | Relevant |
| restecg | Displays resting electrocardiographic results 0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hyperthrophy | Quantitative, Discrete,  Ordinal | Relevant |
| thalach | Max heart rate achieved : displays the max heart rate achieved by an individual. | Quantitative, Discrete,  Count | Relevant |
| exang | exercise induced angina (1 = yes; 0 = no) | Quantitative, Discrete,  Binary | Relevant |
| oldpeak | ST depression induced by exercise relative to rest | Quantitative, Continuous,  Ratio | Relevant |
| slope | Peak exercise ST segment:  0 = upsloping, 1 = flat, 2 = downsloping | Quantitative, Discrete,  Ordinal | Relevant |
| ca | Number of major vessels (0–4) colored by flourosopy : displays the value as integer or float. | Quantitative, Discrete,  Ordinal | Relevant |
| thal | **Thal** : displays the thalassemia | Quantitative, Discrete,  Ordinal | Relevant |
| target | **Diagnosis of heart disease** : Displays whether the individual is suffering from heart disease or not : 0 = absence 1 = present. | Quantitative, Discrete,  Binary | Relevant |

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

**Answer: When we apply Hirarchical Clustering and K-Means Clustering(before PCA) we see that we can easily get clusterings and we get 3 clusters in both the methods and there is a lot of time required for computation, after applying PCA we select 3 PC’s and with these 3 columns, again we apply Hirarchical and K-Means Clustering and it gives same number of 3 clusters like before & reduces the time required for computation, what we observe that there some negative values also exist in this PC’s and that we obtain 3 clusters easily with these 3 clusters we can easily observe and conclude the Patient Pattern who is more suffering from the heart disease and also we reduce dimentions (columns) and It improves the model performance.**

**4.Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

**Answer:**

**1) The first benefit is we reduce the dimension (columns) from all these 14 columns we extract 3 columns means we extract maximum information and after that it’s easy to perform Hirarchical and K-Means Clustering and we can see 3 clusters, with these 3 clusters we can easily observe and conclude the Patient Pattern who is more suffering from the heart disease.**

**2) Second benefits it reduces the time required for computation since less dimensions require less computation.**

**3) It eliminates the redundant features.**

**4) It improves the model performance.**