Experiment 9: Principle Component Analysis

Table of Contents

Name: Ventrapragada Sai Shravani

PRN:17070123120

Batch:Entc(2017-21) G-3

Aim:

Implementation of Principle Component Analysis (PCA) for any dataset

Theory:

PCA stands for Principle component Analysis and is a dimensionality reduction method that is often used to reduce dimensionality of large datasets by transforming large sets of variabes into a smaller one that still contains most of the information in the large set.

Steps to perform PCA

1. Standardisation: To get to same level

2. Covariance Matrix computation: To understand how the variables of the input data are varying from the mean with respect to each other.

3. Take the sum of variables and derive the percentage of variance for each principle compotent.

The larger dispersion indicates more variance

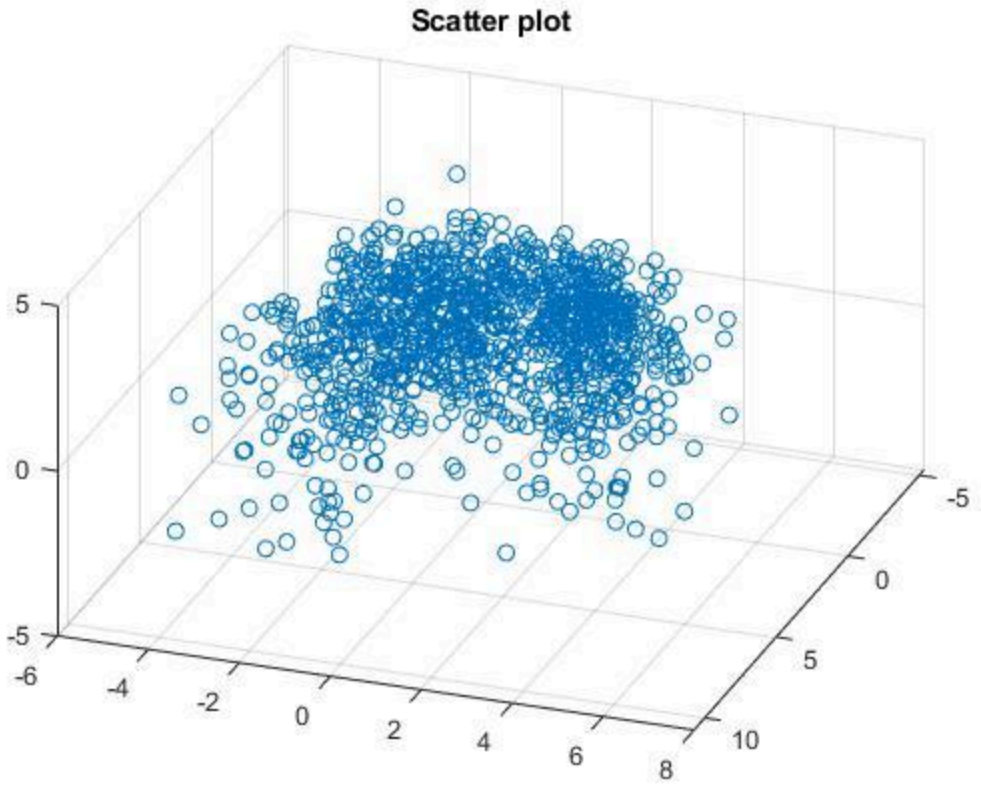
A heat map (or heatmap) is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.

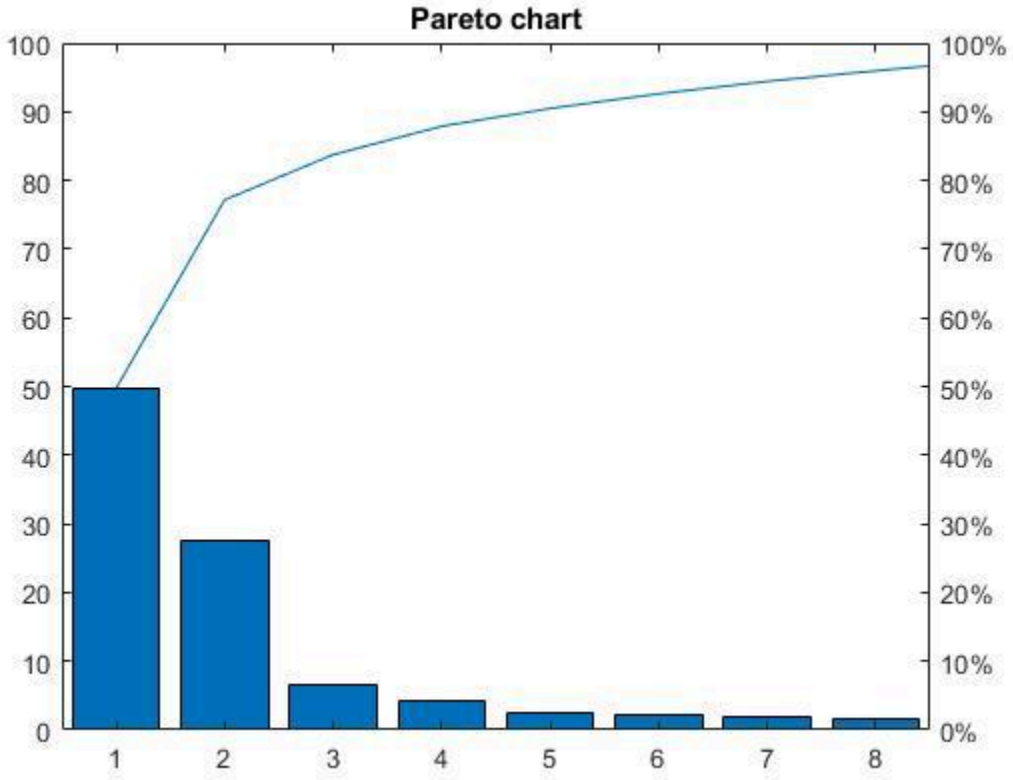
Principle Component Analysis

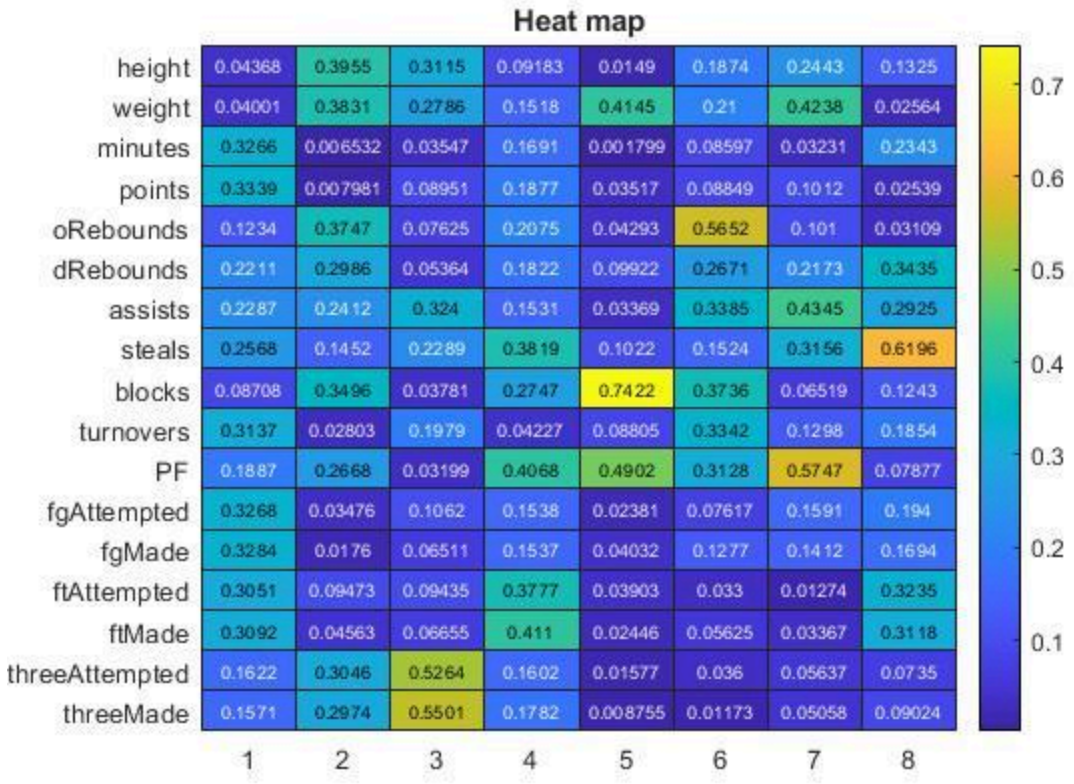
```
clc;
clear all;
close all;
```

```
data = readtable("basketballDataProcessed.csv");
posnames = ["G","G-F","F-G","F","F-C","C-F","C"];
data.pos = categorical(data.pos,posnames);
labels = data.Properties.VariableNames(4:end);
stats = data{:,4:end};
statsNorm = normalize(stats);
rng(0)
[pcs,scrs,latent,~,pexp] = pca(statsNorm,'Algorithm','eig');
fprintf("Percentage of total variance of each component is\n");
disp(pexp);
% PCA scatter
figure();
scatter3(scrs(:,1),scrs(:,2),scrs(:,3))
view(110,40)
title("Scatter plot")
% PCA pareto
figure();
pareto(pexp)
title("Pareto chart")
% Visualise PCA through heat map
varNames = data.Properties.VariableNames(4:end);
figure();
heatmap(abs(pcs(:,1:8)),"YDisplayLabels",varNames,"Colormap",parula);
title("Heat map")
```

```
Percentage of total variance of each component is
   49.7745
   27.4075
    6.5706
    4.1332
    2.6006
    2.1326
    1.8005
    1.5383
    1.4464
    0.8912
    0.7011
    0.5070
    0.3330
    0.0871
    0.0587
    0.0167
    0.0010
```


Scatter plot
5
0
-5
-6
-4
-2
0
2
4
6
8
-5
0
5
10


Pareto chart
100
90
80
70
60
50
40
30
20
10
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
8



Conclusion

In the above experiment we got reid of redundant data as shown in pareto char. The pareto chart has % of variance on the y axis and principle components on the x axis. Plotting the basketball data we can see that it becomes irrelavant as we move down the x- axis and can neglect those components for some reduction in accuracy, but also creates a better model.

The heat map plotted is a standalone visualisation. A standalone visualization is a chart designed for a special purpose that works independently from other charts. a standalone visualization has a preconfigured axes object built into it.

Published with MATLAB® R2020a