Deakin University

SIG720- OnTrack Submission

Task P1

Submitted by

[Guruprasadh J P]
[223939745]
[Attempt # 1]
Date 23 July 2023

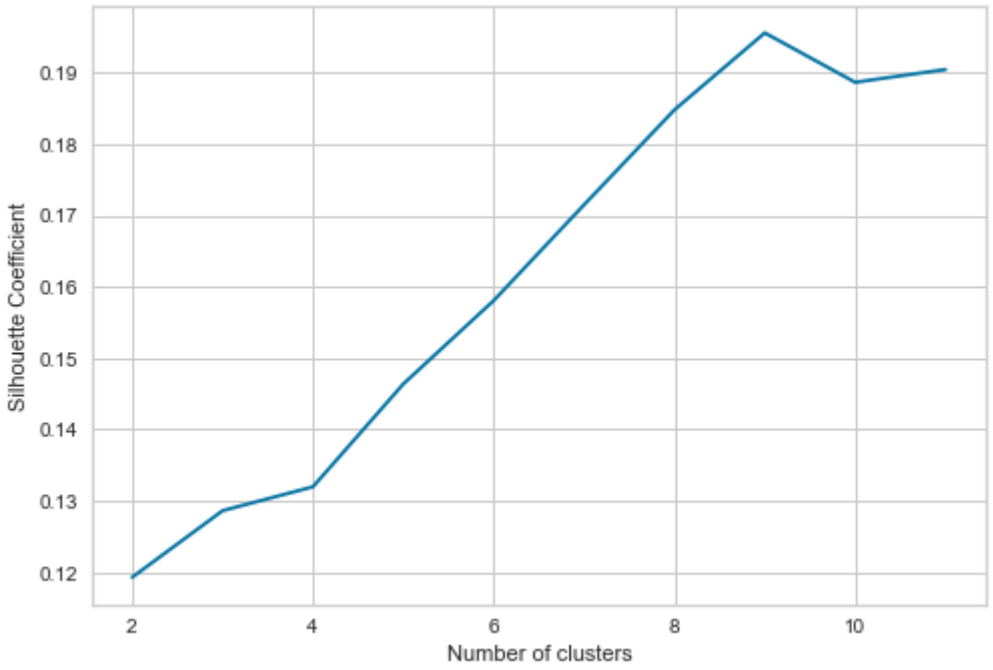
Target Grade : Pass

Task Details –

Part - I

Load data from digitData2.csv file.
The last column of each file presents the label and the rest of the columns are features.

1. Selecting the optimum k value using Silhouette Coefficient and plot the optimum k values.



Based on the average silhouette_score is : 0.1977 the optimum K value is chosen to be 9. (K = 9)

2. Create clusters using Kmeans and Kmeans++ algorithms with optimal k value found in the previous problem. Report performances using appropriate evaluation metrics. Compare the results.

Optimum K value found was 9.

Purity score for Kmeans algorithm with n_clusters = 9 is 0.74

Purity score for Kmeans++ algorithm with n_clusters = 9 is 0.737

3. Now repeat clustering using Kmeans for 50 times and report the average performance. Again, compare the results that you have obtained in Q3 using Kmeans++ and explain the difference (if any).

I haven't attempted this section. If I am given a chance for resubmission I will work it out and submit it.

4. Apply DBSCAN on this dataset (digitData2.csv) and find the optimum "eps" and "min_samples" value. Is the number of clusters the same as the cluster found in Q2? Explain the similarity or differences that you have found between two solutions.

I haven't attempted this section. If I am given a chance for resubmission I will work it out and submit it.

Part – II

Load "diabetes" datasets from SKlearn and do the PCA for n=3, and plot the data.

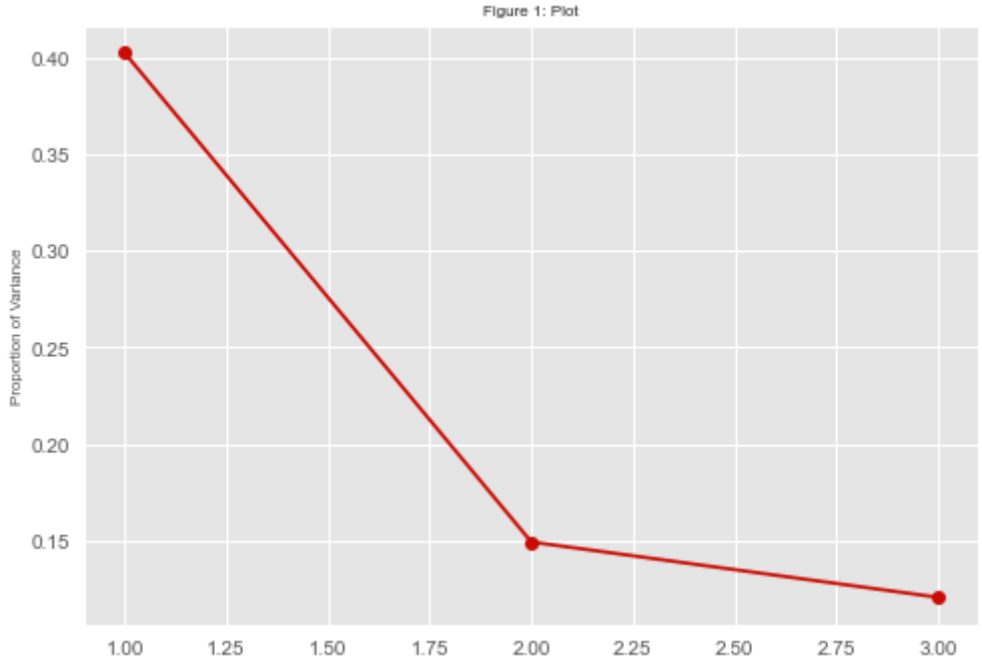
1. What is the variance (%) covered by the first 3 components? How is this percentage of variance computed?

First 3 components variance (%)

0.40 percentage of variance for n=1

0.15 percentage of variance for n=2

0.12 percentage of variance for n=3



2. Is there any correlation between these three components?

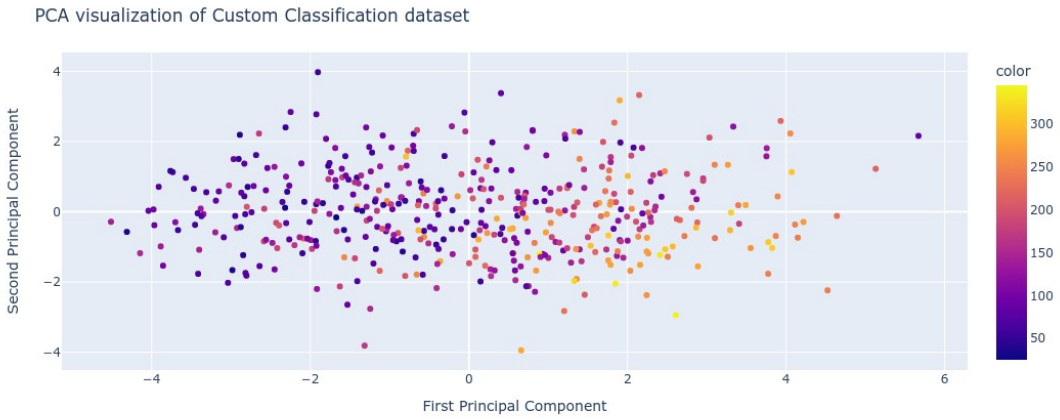
I haven't understand what I have to do for this. So I haven't attempted.

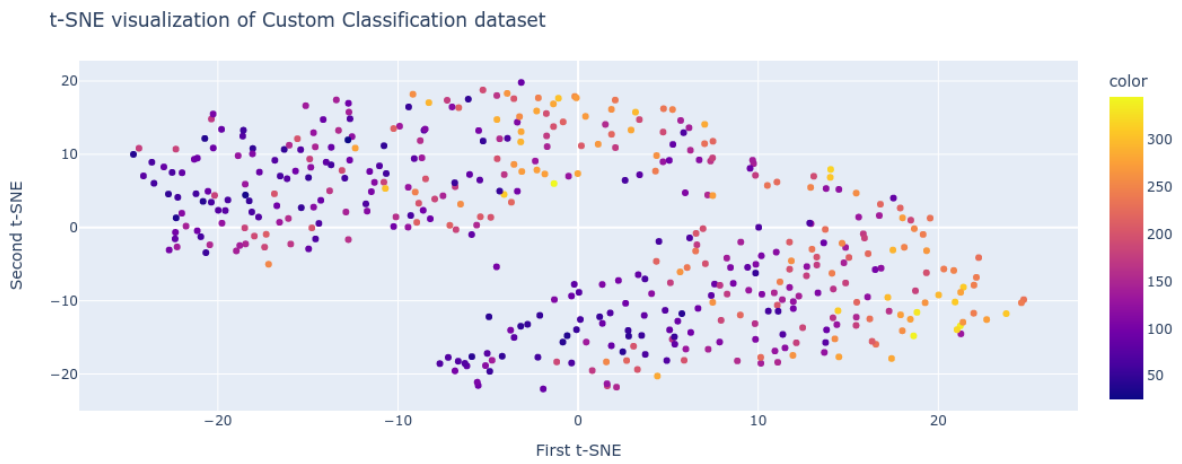
3. Apply t-SNE on the same dataset, select the first 3 components and plot them.

Yes. It’s done.

4. Report the comparison between the results obtained using t-SNE and PCA

For t-SNE we can see it as two cluster data whereas for PCA we cannot see any cluster difference.




t-SNE visualization of Custom Classification dataset
Second t-SNE
First t-SNE
color
20
10
0
−10
−20
−20
−10
0
10
20
300
250
200
150
100
50