IT353 : Deep Learning

Lab Assignment 1 - (PCA - SVD)

Kesanam Ashinee
211AI023

Dataset

A Wheat variety classification Dataset picked from kaggle is used

Link: [Wheat Classification Dataset]

Dataset Description:

The dataset comprised wheat kernels belonging to three different species of wheat:

- Kama
- Rosa
- Canadian

70 elements each.All of these parameters were real-valued continuous

Attribute Information:

To construct the data, seven geometric parameters of wheat kernel were measured:

- area A
- perimeter P
- compactness C = 4piA/P^2
- length of kernel
- width of kernel
- asymmetry coefficient
- length of kernel groove.

Example

	area	perimeter	compactness	length	width	asymmetry coefficient	groove length	category
0	15.26	14.84	0.8710	5.763	3.312	2.221	5.220	1.0
1	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1.0

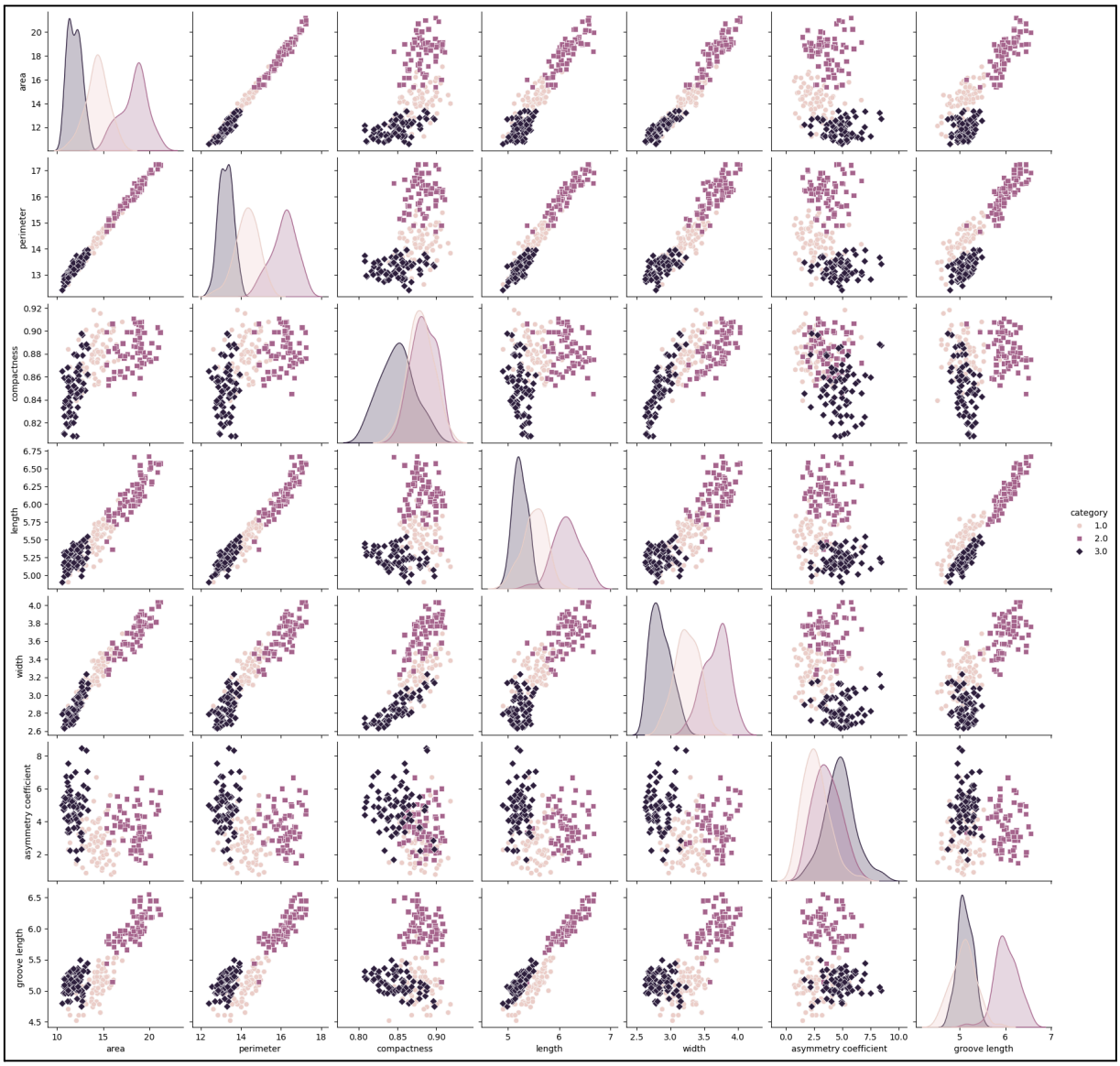
Google Colab Notebook

All code can be found here - https://colab.research.google.com/drive/14L4QJwnzvrQKLBveRiXJtsVyshYAM84A?usp=sharing

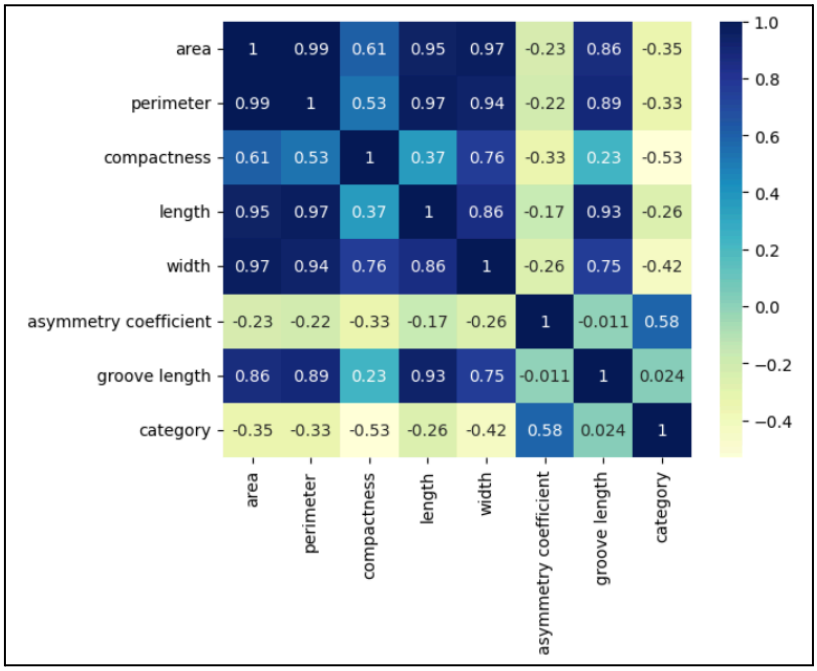
Tasks

A. Visualize the dataset using multiple dimensions and say why SVD and PCA should be used here.

T-SNE Visualization



Heatmap



In the above visualization, the scatter plots are plotted to identify the relationship between the independent columns of the datasets. Out of all the plots some graphs show collinearity with other graphs

Ex: area v/s groove length and perimeter v/s groove length

The correlation values of area v/s groove length and perimeter v/s groove length are 0.86 and 0.89 respectively

From the scatter plot visualization and the heatmap correlation values, both area v/s groove length and perimeter v/s groove length plots show collinearity

The collinearity should be mitigated in order to reduce the dimensions of the data. Hence **dimensionality reduction techniques such as SVD and PCA are useful in this case**

B. Implement SVD and PCA logic on your own and find the appropriate k-dimensions to represent this data.

PCA Approach:

- Mean Value Calculation:
  - Extracted values and calculatde the mean-centered data.
- Calculating Covariance Matrix:
  - Computed the covariance matrix of the mean-centered data.
- Computing Eigenvalues and Eigenvectors:
  - Used NumPy's linalg.eig to find eigenvalues and eigenvectors.
- Sorting and Normalizing Eigenvalues:
  - Sorted eigenvalues and corresponding eigenvectors in descending order.
  - Calculated explained variance and cumulative variance.
- Perform PCA:
  - Transformed the mean-centered data using the sorted eigenvectors.
- K- dimension selection:
  - Based on the variance values, chose 0.99 as the threshold and selected 2 components based on the threshold

Cumulative Variance:

```
cumulative variance:  [0.79559849 0.96589739 0.99493124 0.9987523  0.99973616 1.00049943
 1.00033163 1.        ]
```

PCA Components:

	PC_1	PC_2
0	0.764822	1.638737
1	0.426670	2.857632
2	-0.554333	1.421327
3	-0.944343	1.898546
4	1.721966	2.331480
...	...	...
205	-3.120477	0.241247
206	-4.186589	-0.274431
207	-2.610256	-4.300336
208	-3.409253	0.298024
209	-3.198376	-1.671991

210 rows × 2 columns

SVD Approach

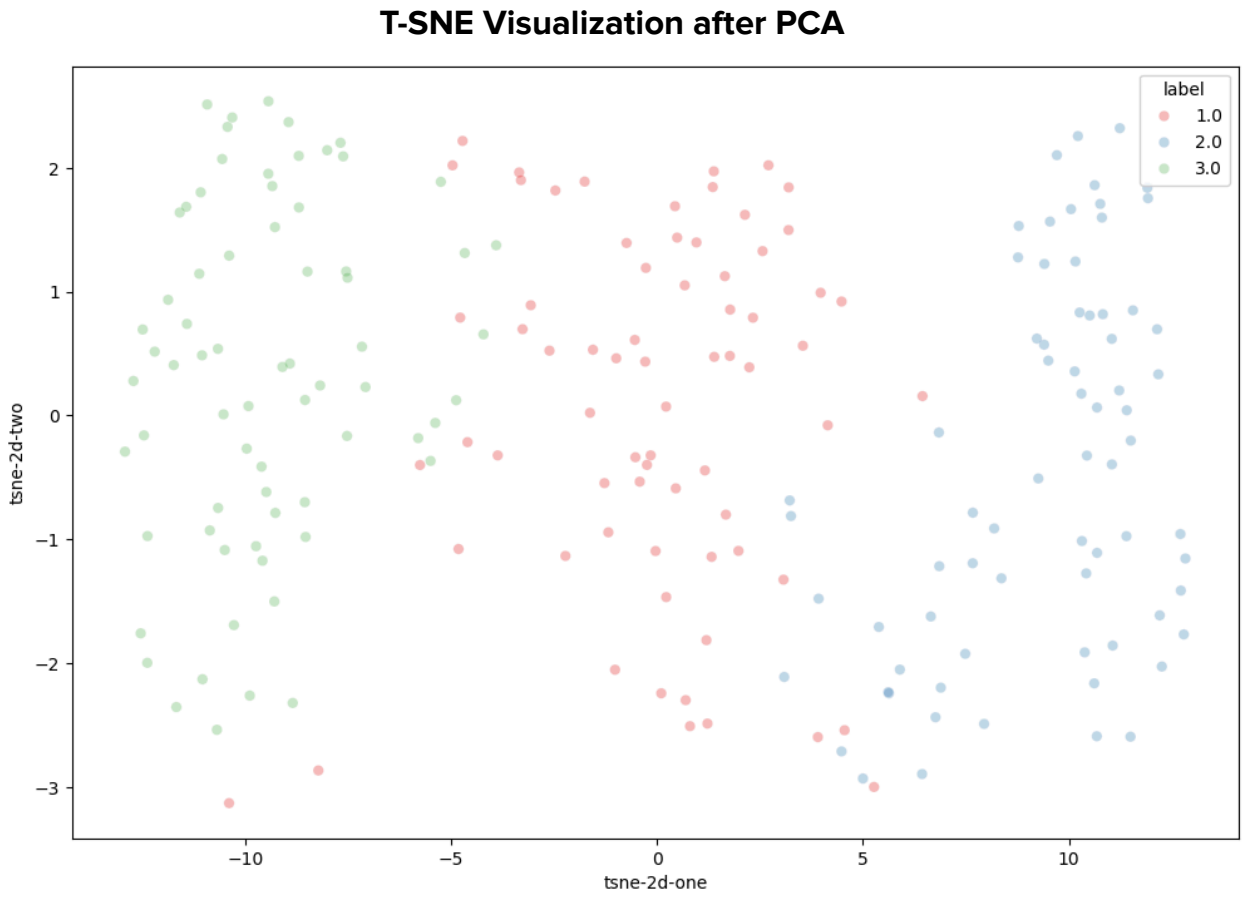
- Calculating Covariance Matrix:
  - Computed the covariance matrix from the features.
- Computing SVD:
  - Used NumPy's linalg.eig to calculate eigenvalues and eigenvectors.
  - Sorted eigenvalues and corresponding eigenvectors in descending order.
  - Computed singular values (σ), left singular vectors (U), and right singular vectors (VT).
- Number of Components:
  - Based on the cumulative variance values 2 components are chosen
- Reducing Dimensionality:
  - Reduced the dimensionality of the features based on the specified or calculated number of components.

SVD Components

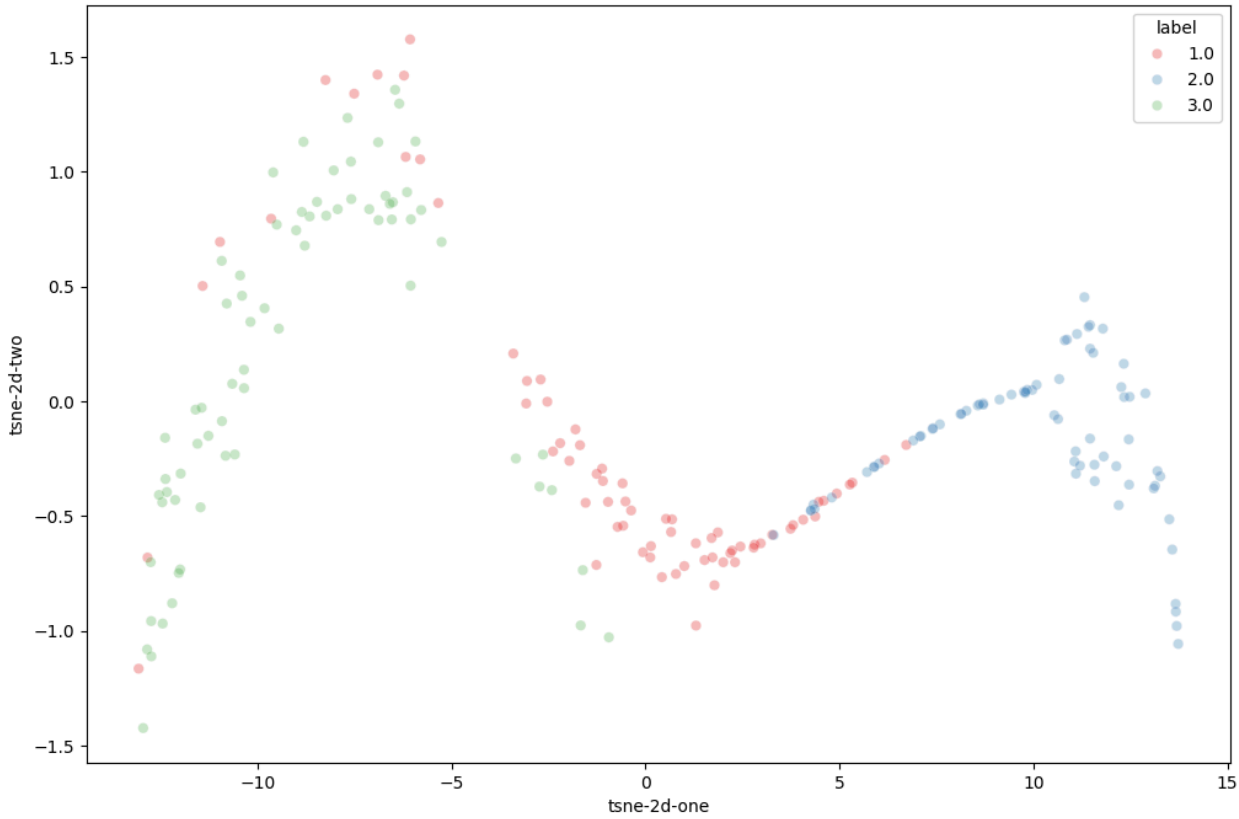
	Component_1	Component_2
0	7645.227662	33.573400
1	7404.775031	58.998092
2	7231.568106	16.909588
3	7081.157089	22.166579
4	7823.937085	63.590518
...	...	...
205	6606.723825	-25.272277
206	6367.139820	-51.588447
207	7207.007253	-122.655061
208	6540.088429	-29.410992
209	6789.662412	-71.126079

210 rows × 3 columns

C. Visualize the data (t-sne plot) after applying SVD and PCA



T-SNE Visualization after SVD



D. State your conclusions as to how SVD and PCA have helped here.

- In this use case, the dimensions of the dataset is reduced to
    - 2 in case of PCA technique from 8 columns
    - 2 in case of SVD technique from8 columns
- The collinearity issue which was prevalent in the dataset was mitigated using both the techniques and the T-SNE visualization shows the data is not collinear
- The T-SNE graph shows how the data is distributed after applying the dimensionality reduction technique and the color shows the data points associated with the wheat category label