Dimensionality Reduction

Mid-term Project Evaluation

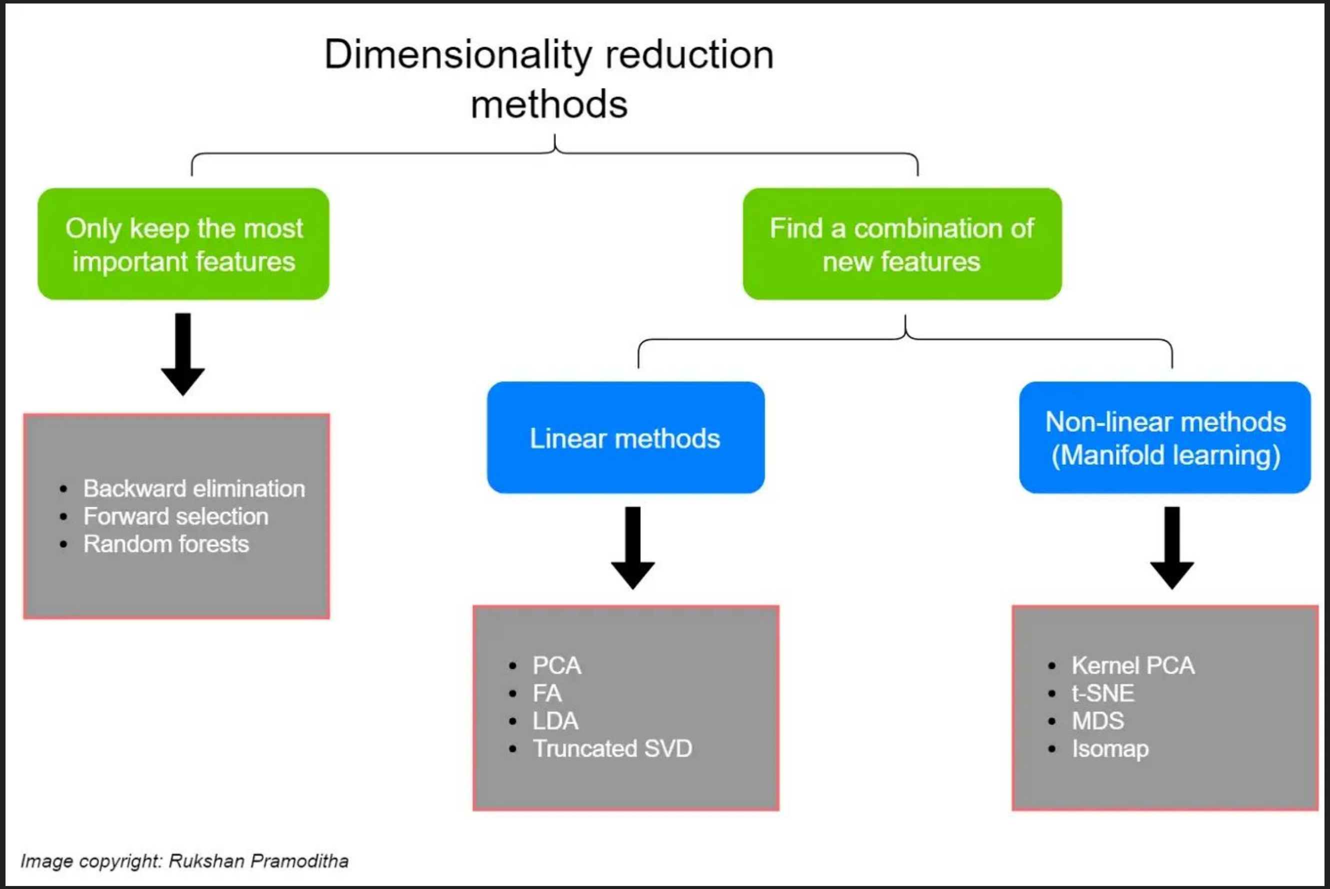
Pankajh Jhamtani
Palak Mishra
Kumar Kanishk Singh
Shlok Mishra

What is Dimensionality Reduction?

Dimensionality reduction refers to techniques that reduce the number of input variables in a dataset. More input features often make a predictive modelling task more challenging to model, more generally referred to as the curse of dimensionality.

High-dimensionality statistics and dimensionality reduction techniques are often used for data visualization. Nevertheless these techniques can be used in applied machine learning to simplify a classification or regression dataset in order to better fit a predictive model.

Dimensionality Reduction Techniques



Dimensionality reduction methods

Only keep the most important features

- Backward elimination
- Forward selection
- Random forests

Find a combination of new features

Linear methods

- PCA
- FA
- LDA
- Truncated SVD

Non-linear methods (Manifold learning)

- Kernel PCA
- t-SNE
- MDS
- Isomap

Image copyright: Rukshan Pramoditha

Principal Component Analysis (PCA)

Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.

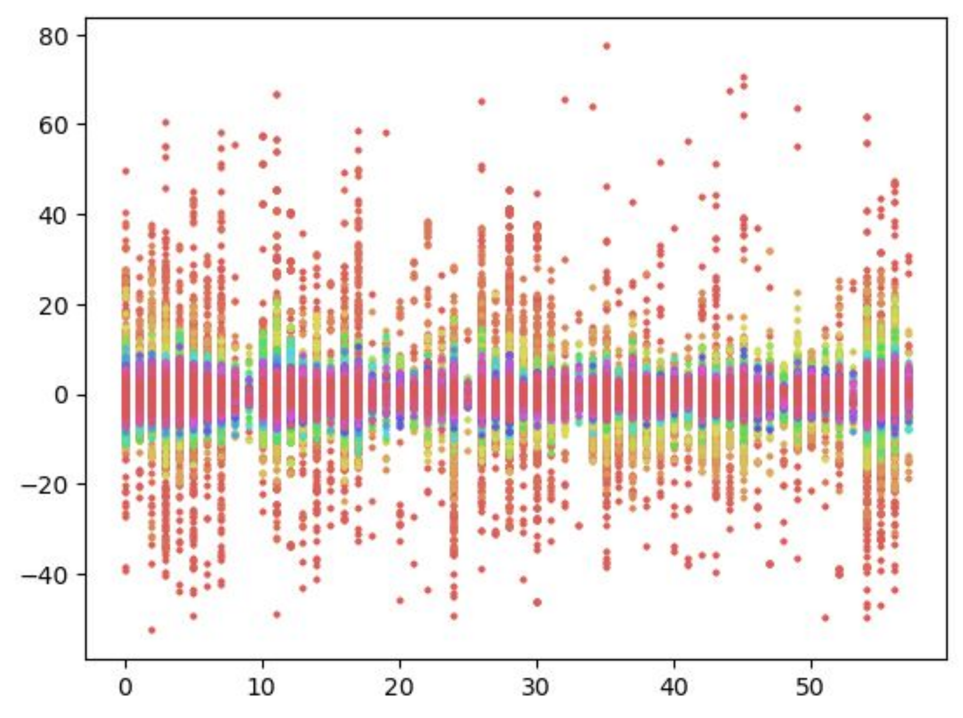
These new transformed features are called the **Principal Components**. It is one of the popular tools that is used for exploratory data analysis and predictive modelling. It is a technique to draw strong patterns from the given dataset by reducing the variances.

Implementing PCA

We have performed PCA on some datasets:

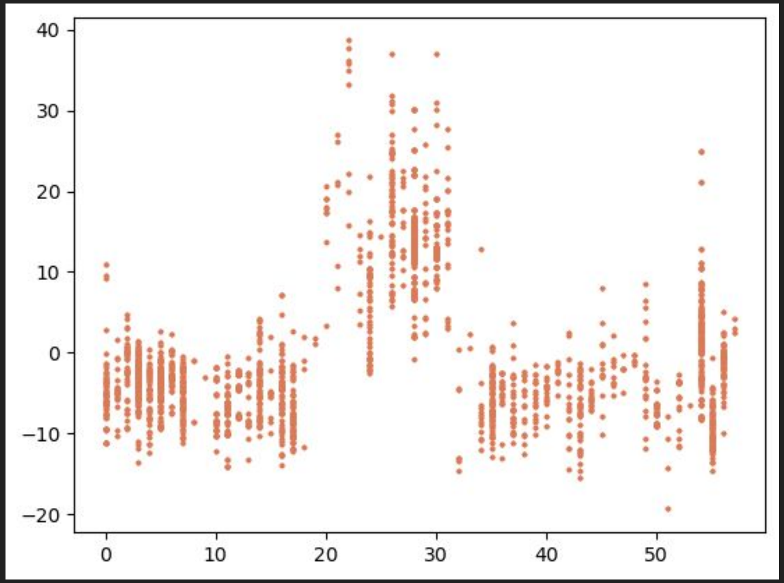
- Chess games Data set
- Traffic signs images
- Pizza Dataset
- Covid Dataset
- Flower Image Dataset

Traffic Signs Data Set



The data set had around 1390 images and 58 different types of Traffic signs. On performing PCA the dimensions were reduced to 58(from 1390*3). The image on the left has a plot, where each color represents one of the 58 components. The X-axis has 58 classes of data(Traffic Signs). Y-axis represents the values of the the components.

The image on the right shows that a particular component of a image can point that the image might belong to a certain group of Traffics Signs

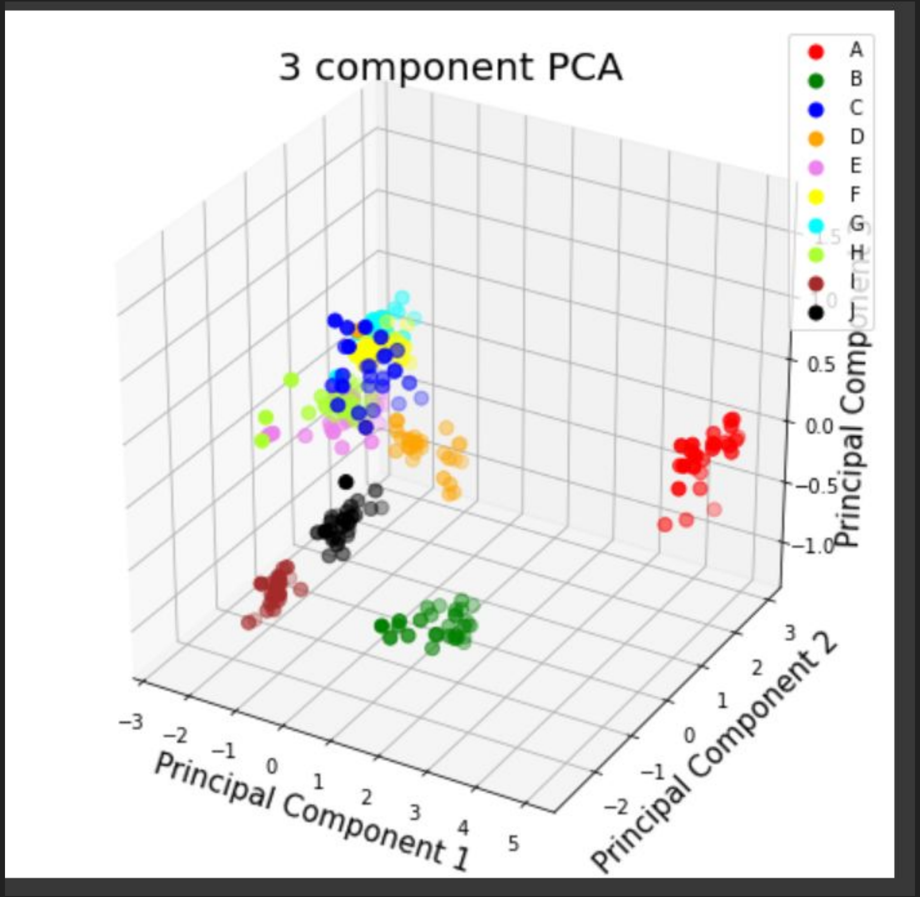


We can say that computation required for performing classification has been reduced a lot for the Traffic Signs Dataset.

Chess games data set

The dataset has record of around 20k games. On performing PCA we were able to reduce dimension of the data set to around 8. Orignal dataset had 16 columns of information for each game. The Information lost in the dimensionality reduction was around 11 percent.

Data Visualization of Pizza Dataset



Data visualization can give good intuition to how your data “looks” and what are the directions you’ll need to consider when building the right model.

In our dataset there are 300 rows and 9 columns of 10 different types of pizzas and 8features like moisture, fat, etc. therefore the data is 9 dimensional. We will now reduce the dimensions of our data to 3 using PCA, which is much easier to visualize.

We can see from the given plot that pizza of type A, B, I and J are separated from other types of pizzas which are clustered. So, we can say that pizzas of t type A, B, I and J are easily distinguishable from other types.

Covid Dataset

We worked on the analysis of the covid data too. This dataset is web scraped from Wordometer, using R language.

There were 5 numerically-relevant dimensions initially. Using PCA, the number of dimensions were reduced down to three principal components.

Following is the figure of the 3D visualisation of the modified new dataset, colour-coded based on the continent a country belongs to.

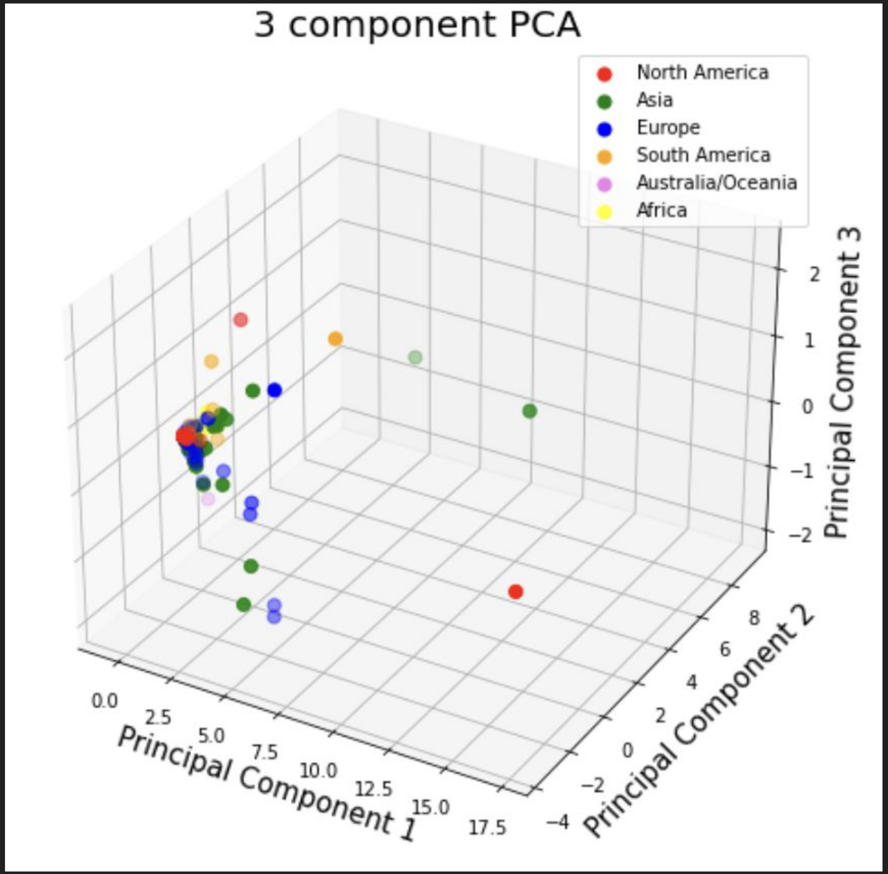
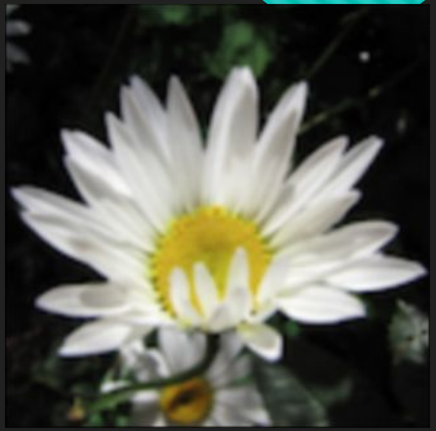
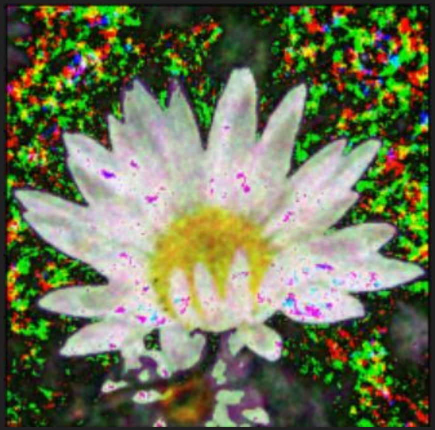


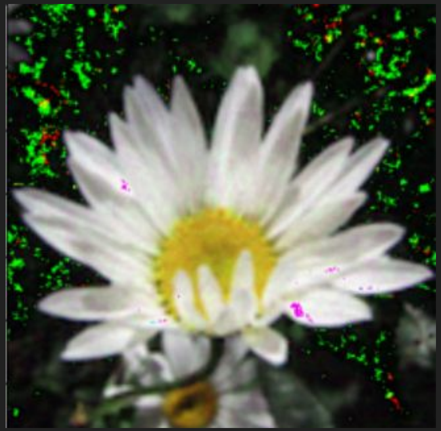
Image Compression on Flower Image Dataset



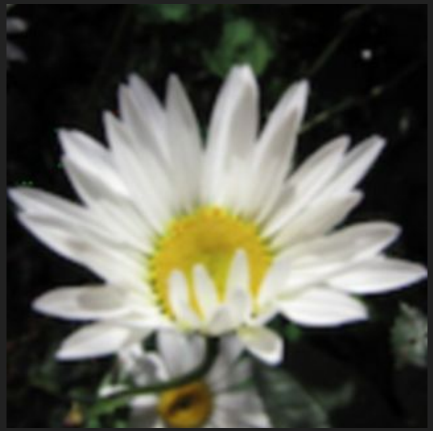
Original Image



Compressed Image with 400 components



Compressed Image with 550 components



Compressed Image with 600 components

Images have a huge number of features and take up a lot of disk space. A solution to this problem is compressing them. To do this, we reduce the dimensionality using PCA.

Our data set contains 813 images of various flowers in color. On applying PCA with a specified number of principals.

We can see from the above images that the quality improves with increase in the principal components.

Future Prospects

We are planning to perform dimensionality reduction using auto-encoders on the same data sets. We would then compare auto-encoding and PCA. We can explore more methods of Dimensionality reduction.