Exploring Clustering and Dim Redn Algorithms

# Datasets

For the project two datasets were chosen *Dry Beans* and *Wine Quality* sourced from UCI ML data repo. The datasets have all numeric features to circumvent requirement of any special treatment for categorical features. The datasets also have a target variable defined to help verify the unsupervised clustering against ground truth.

In pair plots of *Dry Beans* exhibit clear clustering and strongly correlated features. There are no outliers. *Wine Quality* data set however does not demonstrate clear clustering across features, no strong correlation and has outliers. This should provide for a good contrast on the algos performance.

# Clustering

We chose KNN and Gaussan – KNN sticks with circular blobs so has limitations unless there is clear groups.

Gaussian should be more adaptable to the

Clustering raw data – show the elbow + sil score, and the BIC 3 plots

Data with clustering2 plots

## Dry Beans

### KMeans

A graph with a line

Description automatically generatedA graph with a line

Description automatically generated A chart of different types of beans

Description automatically generated

### Gaussian Mixture

A graph with blue lines

Description automatically generated A diagram of a line graph

Description automatically generated with medium confidence

## Wine Quality

### KMeans

A graph of a line

Description automatically generatedA graph with a line

Description automatically generatedA diagram of a wine quality

Description automatically generated

### Gaussian Mixture

A graph with blue lines

Description automatically generatedA diagram of different colored dots

Description automatically generated

# Dimension Reduction

StandardScaler was applied to all datasets before all algorithms of clustering /dimension reduction.

***- WHY? Does this help with dealing with outliers?***

*Dim Redn – how many components were required –*

*Give the skree plots 4 plots*

*Give the distrb plot after dim redn 4 plots*

*Explain if clusters formed – if not why might they not have formed*

## Dry Beans

A diagram of different colored dots

Description automatically generatedA diagram of a number of different colored dots

Description automatically generated with medium confidenceA diagram of different colored dots

Description automatically generatedA map of different colored dots

Description automatically generated

A graph with a line

Description automatically generatedA graph with a line

Description automatically generated A graph with a line graph

Description automatically generated A graph of error

Description automatically generated with medium confidence

A graph with a line

Description automatically generatedA graph of different colored lines

Description automatically generated

## Wine Quality

A diagram of a variety of wine quality

Description automatically generated with medium confidenceA diagram of a number of purple dots

Description automatically generatedA diagram of a variety of wine quality

Description automatically generated with medium confidence A diagram of a wine quality scatterplot

Description automatically generated

A graph with a line

Description automatically generatedA graph with a line

Description automatically generated A graph with a line

Description automatically generated A bar graph with different colored bars

Description automatically generated

A graph of a line

Description automatically generatedA graph of different colored lines

Description automatically generated

# 

# Clustering with Dim reduction

*KNN Clustering with dim reduction –*

*Based on dim reduction analysis – chose the 2 primary factors.*

*KNN – what is the best cluster – we don’t see an elbow*

*Gaussian Mix – BIC is all over the place*

*High dimensionality leads to issue in clustering. No adv*

*Compare - how did you choose the number of components to use? – did you see improved clustering/grouping in plot?*

## Dry Beans

### Original vs KMeans vs GMM

A diagram of different colored dots

Description automatically generatedA chart of different colored dots

Description automatically generatedA chart with different colored dots

Description automatically generatedA diagram of a variety of colored dots

Description automatically generated with medium confidence

A diagram of different colored circles

Description automatically generatedA diagram of different colored circles

Description automatically generatedA diagram of a cluster of dots

Description automatically generated with medium confidenceA diagram of a cluster of colored dots

Description automatically generated

A diagram of different colored circles

Description automatically generatedA diagram of different colored dots

Description automatically generatedA diagram of a clustering of dots

Description automatically generated with medium confidenceA diagram of different colored dots

Description automatically generated

KMeans underperforms as ther eis no clear boundary for clusers ven after dim red. GMM excels here as thee is a clear clustering of the points an

TSNE suggested 12 groups

## Wine Quality

Dim reduction required almost all the features (8) to keep the reconstr err low.

Reconstruction error also does not decrease monotonically

Does PCA work well with outliers?

### Original vs KMeans vs GMM

A diagram of a diagram with colored dots

Description automatically generated with medium confidenceA diagram of a number of colored dots

Description automatically generatedA chart with a diagram of different colored dots

Description automatically generated with medium confidenceA diagram of a wine quality

Description automatically generated

A diagram of different colored dots

Description automatically generatedA chart with many colored dots

Description automatically generatedA chart of different colored dots

Description automatically generatedA diagram of different colored dots

Description automatically generated

A diagram of a cluster of dots

Description automatically generated with medium confidenceA diagram of a cluster of colored dots

Description automatically generatedA diagram of a cluster of dots

Description automatically generatedA diagram of a cluster of wine quality

Description automatically generated

In kmeans ther eis no clear elbow and silhouette suggested 4 groups. Score was also below 0.3

# References

<https://towardsdatascience.com/kmeans-hyper-parameters-explained-with-examples-c93505820cd3>

<https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/>

<https://towardsdatascience.com/gaussian-mixture-model-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4>

<https://towardsdatascience.com/gaussian-mixture-model-clearly-explained-115010f7d4cf>

<https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html>

<https://scentellegher.github.io/machine-learning/2020/01/27/pca-loadings-sklearn.html>

<https://towardsdatascience.com/introduction-to-ica-independent-component-analysis-b2c3c4720cd9>

<https://medium.com/@violante.andre/an-introduction-to-t-sne-with-python-example-47e6ae7dc58f>

<https://distill.pub/2016/misread-tsne/>