CS- 7641 HW3: Unsupervised Learning and Dimensionality Reduction

Mtabesh6@gatech.edu

1. Introduction:

Clustering methods are unsupervised machine learning algorithms that group similar objects based on their characteristics or features. These methods aim to identify similarities within a dataset and separate dissimilar objects. Clustering is commonly used in a variety of fields, including data analysis, image processing, and market segmentation.

Dimension reduction methods are techniques used to reduce the number of variables or features in a dataset while retaining as much of the important information as possible. These methods are often used to simplify complex datasets, improve computational efficiency, and visualize high-dimensional data in a lower-dimensional space.

In this homework, we will examine and compare two different clustering methods including K-means and Expectation Maximization, and 4 different dimension reduction methods including PCA, ICA, Random Projection, and isomap.

K-means is a popular clustering algorithm that partitions a dataset into k clusters. The algorithm works by iteratively assigning data points to the nearest cluster center and updating the cluster centers based on the mean of the data points in the cluster. K-means is widely used for its simplicity, efficiency, and ability to handle large datasets.

Expectation Maximization (EM) is a popular algorithm used to find the maximum likelihood estimates of parameters in mixture models with hidden variables. EM works by iteratively assigning probabilities of cluster membership to each data point and then re-estimating the model parameters based on these probabilities. EM can be used for a variety of clustering applications, including Gaussian mixture models and hidden Markov models.

Principal Component Analysis (PCA) is a popular dimension reduction technique that transforms high-dimensional data into a lower-dimensional space. PCA works by identifying the principal components of the data, which are the directions in the data that capture the most variation. These principal components can then be used to represent the data in a lower-dimensional space while minimizing information loss.

Independent Component Analysis (ICA) is a computational technique used to separate a multivariate signal into independent, non-Gaussian components. ICA assumes that the observed signal is a linear combination of independent sources and aims to recover these sources from the observed signal. ICA has a wide range of applications, including signal processing, image processing, and data compression.

Random Projection is a simple yet effective dimension reduction technique that projects high-dimensional data onto a lower-dimensional space using a random matrix. The random matrix is typically generated by sampling from a random distribution, such as the Gaussian distribution. Random Projection is computationally efficient and can preserve the pairwise distances between data points, making it suitable for applications such as nearest-neighbor search and clustering.

Isomap (Isometric Feature Mapping) is a nonlinear dimension reduction technique that preserves the geodesic distances between data points in a lower-dimensional space. Isomap works by constructing a graph of nearest neighbors and then computing the shortest path distances between all pairs of points on the graph using graph theory algorithms. The resulting lower-dimensional representation of the data can capture the underlying nonlinear structure of the data, making it useful for applications such as image processing and natural language processing.

2. Datasets:

Adult Income Dataset: In this dataset, information from almost 50000 US residents in 14 features is gathered. The ground truth is whether the income of the person is greater than $50k or less than that. Data was initially extracted from the US Census Bureau [1]. The data sample were edited and uploaded to the UCIU machine learning repo [2] for educational purposes. The features are the combination of categorical and continuous variables. Feature includes basic information including demographic, education occupation marital status, etc.

In this step, the data were prepared for model training.

- Some of the categories with high category values have been regrouped. For example, education has been relabeled for are school grades as some school education. For a country of birth, the data have been regrouped to US and non-US citizens. The marital status was regrouped to married and not married.
- Converting all categorical variables to numerical values using one-hot encoding. **The total features are 68.**
- All the numerical values were normalized and standardized.
- The target variables were relabeled from { '>50K', '<=50K'} to {1,0}.
- Since there are many individuals with no capital gain or loss, I need to recreate a column for the individuals that have any capital gain or not as 0 and 1.
- Hours of working are continuous. I make them into three categories <40, 40, and >40 hours of work.
- Due to the low computational resource, **1000 random sample points** were used for analysis. 800 for training and 200 for a test.
- To deal with data imbalance, SMOTE method was used. In SMOTE, the minority class samples are used to create synthetic examples by taking the difference between a randomly chosen sample and its k nearest neighbors in the feature space. This difference is then multiplied by a random number between 0 and 1 and added to the original sample, producing a new synthetic sample. The process is repeated to generate more synthetic samples. The goal is to increase the number of samples in the minority class so that the classifier has more information to learn from and can make more informed predictions.
- Only training data was balanced:
    - the ratio of >50 in imbalance data: 0.2225, number of sample points: 800
    - the ratio of >50 in SMOTE balanced data: 0.5, number of sample points: 1244

Letter Recognition Dataset: The dataset includes features extracted from handwritten letters. The goal of this task is to classify a large number of rectangular pixel displays into one of the 26 capital letters in the English alphabet. The dataset consists of 20,000 unique stimuli, each based on one of 20 different fonts and randomly distorted. To represent each stimulus, 16 numerical attributes were extracted, including statistical moments and edge counts, and scaled to fit into a range of integer values between 0 and 15. The target classes are almost equally distributed.

3. Evaluation Metrics:

Inertia is a metric used to evaluate the quality of clustering in k-means models. It measures the sum of squared distances between each data point and its assigned cluster center. A lower inertia value indicates that the data points are more tightly clustered around their respective cluster centers, which is generally considered a better clustering performance.

Euclidean distance is a metric for measuring the similarity between data points and cluster centers. It is calculated as the straight-line distance between two points in a multi-dimensional space. In k-means, Euclidean distance is used to assign each data point to its nearest cluster center and update the center location in each iteration of the algorithm.

The silhouette score, or **si-score**, is a metric used to evaluate the quality of clustering in k-means models. It measures how well each data point fits into its assigned cluster compared to other clusters. The silhouette score ranges from -1 to 1, with higher values indicating better clustering performance, and a score of 0 indicating that a data point may belong to more than one cluster.

The F1 score is a metric commonly used to evaluate the performance of clustering algorithms such as k-means. It is calculated based on the precision and recall of the clustering results. Precision measures the

percentage of true positives among all the samples assigned to a cluster, while recall measures the percentage of true positives among all the samples that belong to a certain category. The F1 score combines these two metrics to provide a single score that represents the overall performance of the clustering algorithm.

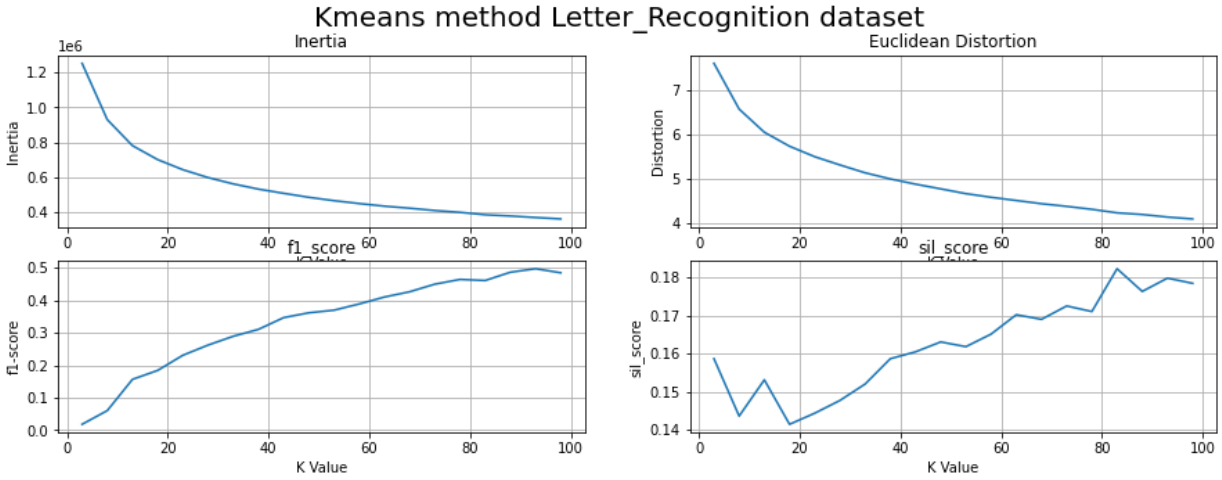
Akaike Information Criterion (AIC) is a statistical measure used in the Expectation Maximization (EM) algorithm to select the optimal number of clusters in a model. It is based on the trade-off between the goodness of fit of the model and its complexity, where a good model should fit the data well but not be too complex. AIC considers both the likelihood of the data and the number of parameters in the model and provides a score that can be used to compare models with different numbers of clusters. The model with the lowest AIC value is generally considered the best fit for the data.

Bayesian Information Criterion (BIC) is a statistical measure used in the Expectation Maximization (EM) algorithm to select the optimal number of clusters in a model. Like the Akaike Information Criterion (AIC), BIC is based on the trade-off between model fit and complexity, but it places a greater penalty on models with more parameters. BIC takes into account both the likelihood of the data and the number of parameters in the model and provides a score that can be used to compare models with different numbers of clusters. The model with the lowest BIC value is generally considered the best fit for the data.

4. Clustering Methods:

Kmeans Clustering- Letter Recognition dataset:

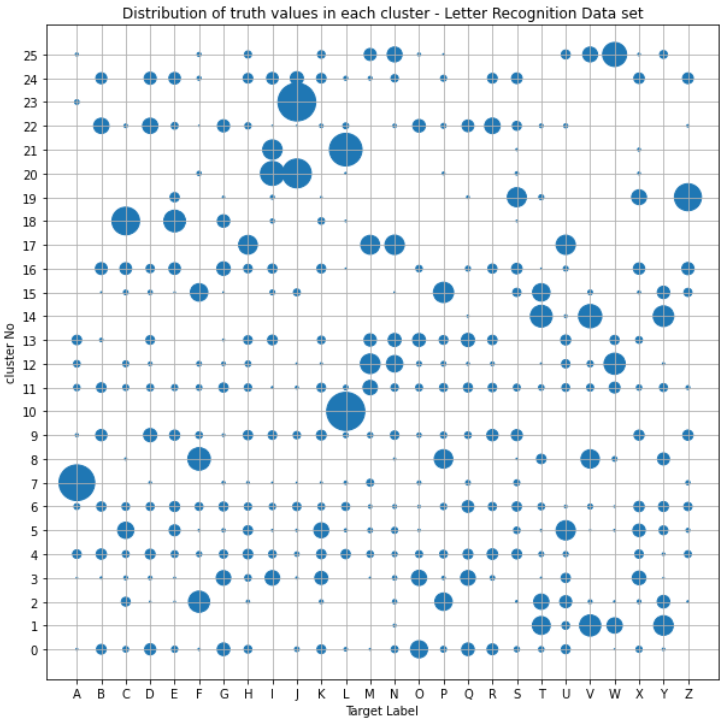
Kmeans clustering was applied to the dataset. The k value ranged from 1 to 100 with 5 intervals. The elbow method is a heuristic used to determine the optimal number of clusters in a k-means clustering model. The method involves plotting the within-cluster information as a function of the number of clusters and identifying the "elbow" or "kink" point in the plot where the rate of decrease in information slows down. The number of clusters corresponding to this elbow point is often chosen as the optimal number of clusters for the model. For this dataset 4 different metrics were analyzed and shown as a function of k. As shown below by increasing the number of clusters the inertia and Euclidian distance will decrease. The elbow point can be considered as **80** for this dataset. The maximum F1-Score is about % 50 which is a good number for this dataset.



However, since we have 26 classes including the letters, I tried to use 26 clusters in k-means as well. An innovative plot was generated for this dataset. Ground truth classes were shown on the x-axis and K-means cluster labels were shown on the y-axis. The size of the circles indicates the normalized number of instances in each set of ground truth and k-means cluster labels. It gives a lot of useful information. For example, Letter A was mainly labeled as cluster number 7 in k-means, meaning the k-means model can

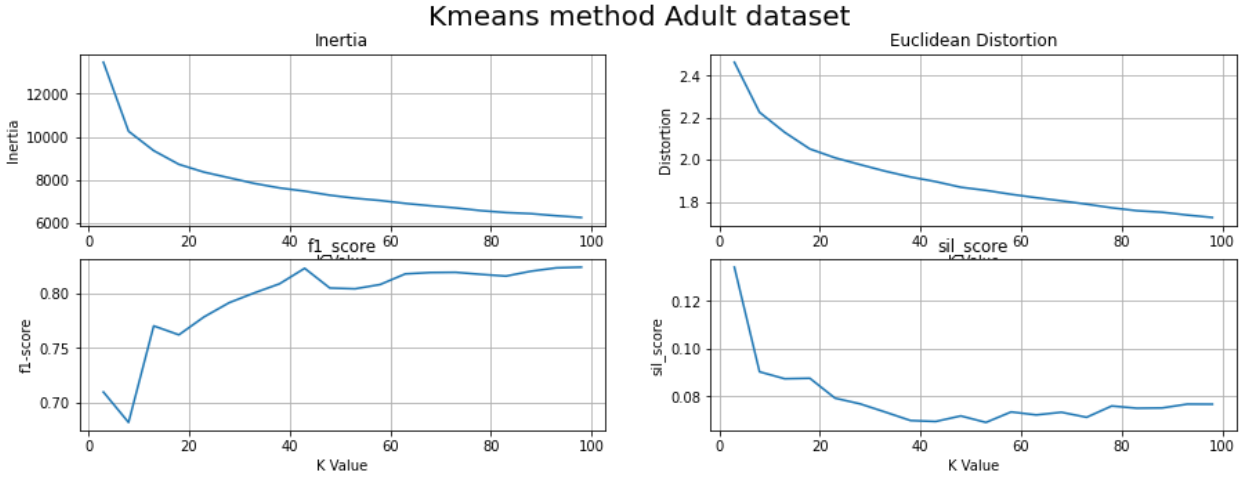
cluster many “A” instances in one cluster. However, the letter “B” were scattered between the clusters, meaning that k-means is not efficient in clustering features associated with the letter B.

Another interesting piece of information is that k-means cluster number 20 is mostly associated with the letter “I” and “J” which is due to the similarity of these two letters in writing.

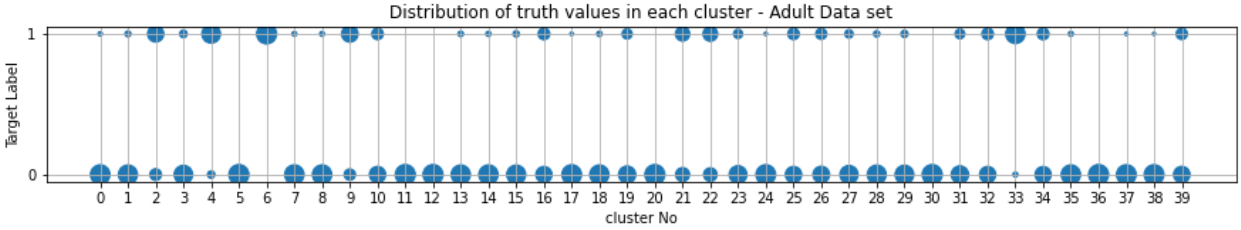


Kmeans Clustering- Adult Income dataset:

Kmeans clustering was applied to the dataset. The k value ranged from 1 to 100 with 5 intervals. The elbow method is a heuristic used to determine the optimal number of clusters in a k-means clustering model. For this dataset, 4 different metrics were analyzed and shown as a function of k. As shown below by increasing the number of clusters the inertia and Euclidian distance will decrease. The elbow point can be considered as 40 for this dataset. The maximum F1-Score is about % 80 which is a good number for this dataset.



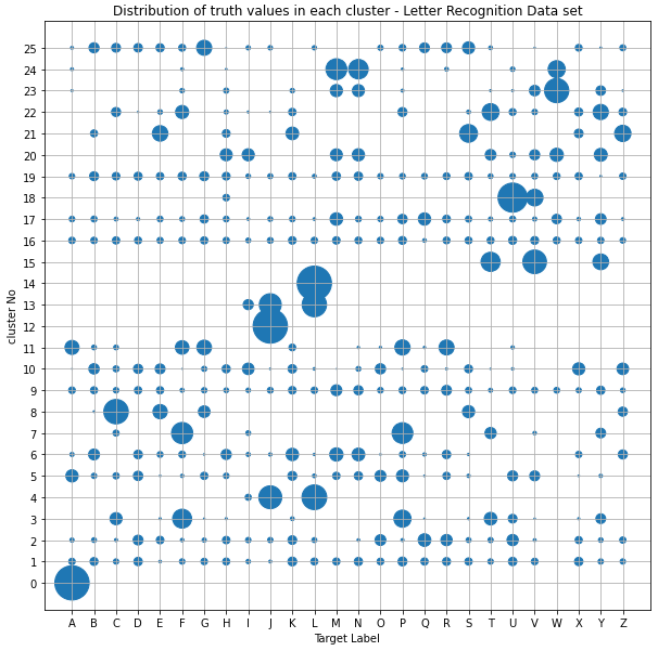
Based on the elbow method k= 40 clusters were chosen as several clusters. The plot below shows the distribution of ground truth classes in each cluster number. The f-1 score for this model in %83. This a very good number meaning the k-means work perfectly for this dataset. The data were balanced and as shown below, cluster numbers 4, 6, and 33 are mainly associated with class 1 or income over %50. The rest of the clusters are associated with class 0 or below 50k income.



Expectation Maximization Letter Recognition dataset:

Expectation maximizations were applied to the dataset. The number of clusters ranged from 1 to 100 with 5 intervals. The elbow method is a heuristic used to determine the optimal number of clusters in a k-means clustering model. For this dataset 4 different metrics were analyzed and shown as a function of k. As shown below by increasing the number of clusters the inertia and Euclidian distance will decrease. The elbow point can be considered as **60** for this dataset. The maximum F1-Score is about % 50 which is a good number for this dataset.

However, since we have 26 classes including the letters, I tried to use 26 clusters in k-means as well. Ground truth classes were shown on the x-axis and EM cluster labels were shown on the y-axis. Same as k-means, Letter A, was mainly labeled as cluster number 1 in EM, meaning the EM model can cluster many of “A” instances in one cluster. However, the letter “B” were scattered between the clusters, meaning that EM such as k-means is not efficient in clustering features associated with the letter B.

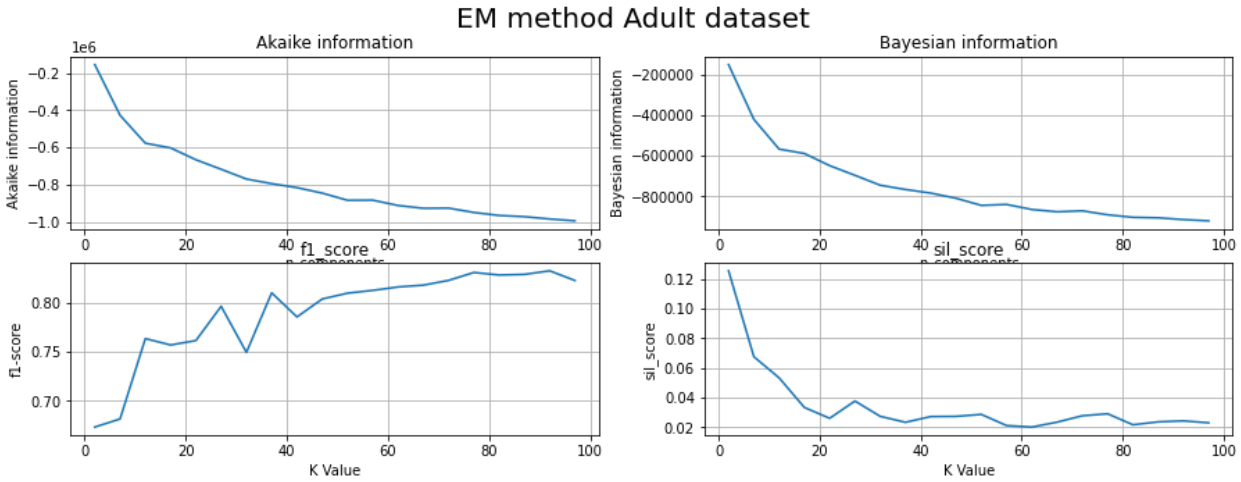


In opposition to k-means, the letter I and J are not labeled as the same in EM clustering. However, the letter J and L were labeled in the same cluster.

The EM method perfectly clusters letters A, L, and K, however, k-means perfectly cluster letters A, L, and J.

Expectation Maximization - Adult Income dataset:

Expectation maximizations were applied to the dataset. The number of clusters ranged from 1 to 100 with 5 intervals. The elbow method is a heuristic used to determine the optimal number of clusters in a k-means clustering model. For this dataset, 4 different metrics were analyzed and shown as a function of k. As shown below by increasing the number of clusters the inertia and Euclidian distance will decrease. The elbow point can be considered as **20** for this dataset. The maximum F1-Score is about % 90 which is a very high number for this dataset.



Based on the elbow method k= 40 clusters were chosen as several clusters. The plot below shows the distribution of ground truth classes in each cluster number. The f-1 score for this model in %80. A very good number means the EM works perfectly for this dataset. The data were balanced and as shown below, cluster numbers 6, 9,10 are mainly associated with class 1 or income over %50. The rest of the clusters are associated with class 0 or below 50k income.

Conclusion:

K-means and Expectation Maximization (EM) are two popular clustering algorithms that have some similarities and differences.

Both algorithms are iterative and involve assigning data points to clusters and updating cluster centers in each iteration. Both algorithms aim to minimize the sum of squared distances between data points and their assigned cluster centers. Both algorithms are sensitive to the initial conditions and can produce different results based on the initial random assignment of cluster centers.

K-means assumes that each data point belongs to one and only one cluster, while EM allows for the soft assignment of data points to multiple clusters based on the probabilities of belonging to each cluster. K-means uses Euclidean distance to measure the similarity between data points and cluster centers, while EM can use different distance metrics and probability distributions depending on the type of data being clustered. K-means converge to a local optimum, while EM tries to find the global optimum by maximizing the likelihood of the data given the model parameters.

Overall, k-means is faster and simpler to implement than EM, but EM is more flexible and can handle more complex data distributions. The choice of algorithm depends on the specific characteristics of the data and the goals of the analysis.

5. Dimension reduction algorithms:

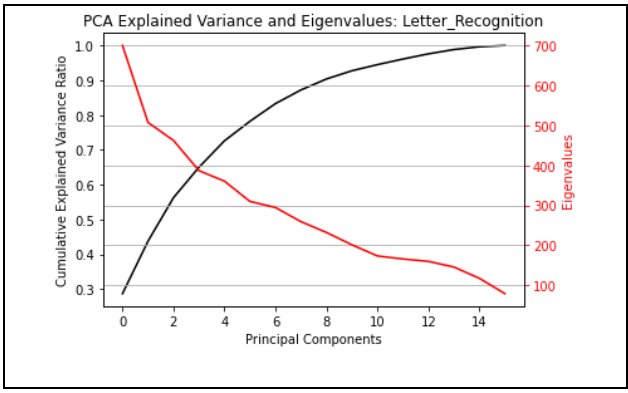
PCA:

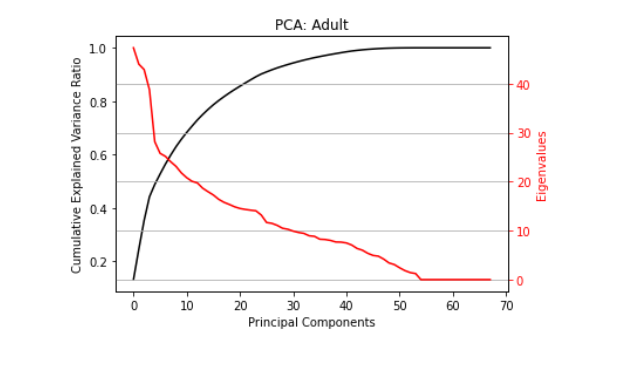
Cumulative explained variance in Principal Component Analysis (PCA) is a metric used to evaluate the amount of variance explained by each principal component in a dataset. It measures the total amount of variation captured by the first n principal components in descending order. The cumulative explained variance is useful for determining the minimum number of components required to capture a desired level of information and for visualizing the relative importance of each component.

Eigenvalue is a scalar that represents the amount of variance explained by a given principal component. It is obtained by solving the eigenvector equation for the covariance matrix of the data. The eigenvalues of the principal components are sorted in descending order so that the first principal component has the largest eigenvalue and explains the most variance in the data.

PCA method was applied to the dataset, as shown below, for the letter recognition dataset. The cumulative variance was shown in black and the eigenvalue was shown in red. As shown in the figure, almost 8 PC will explain %90 of the variance in the letter recognition dataset. Thus if we need to reduce the feature space, we can choose % a 90 variance or 8 PCs of the PCA method.

For the adult income dataset, 20 PCs will explain %90 variances in the dataset, That is a huge reduction in the number of features in the adult income data set by keeping the most important information. PCA will be more effective in the adult income dataset compared to the letter recognition dataset.

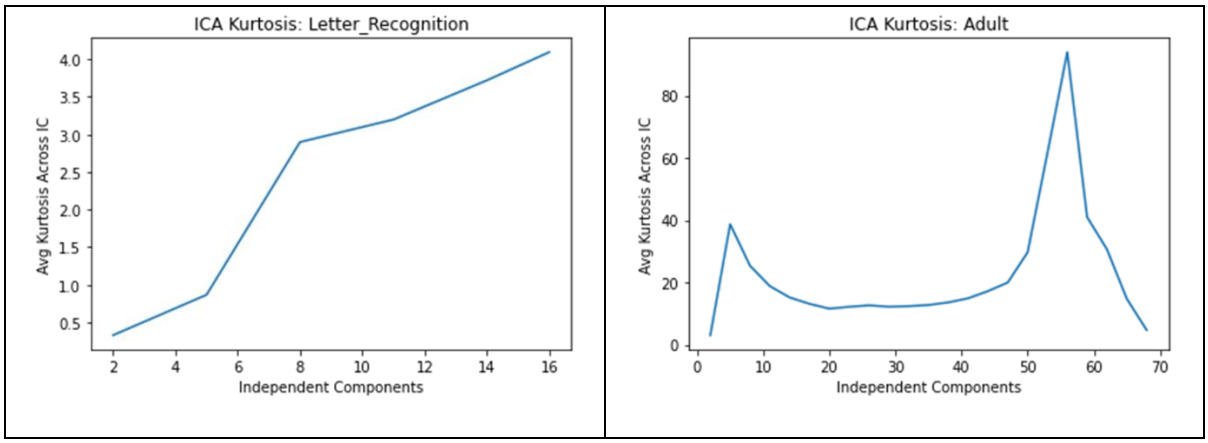




ICA:

Kurtosis is a statistical measure that describes the "peakedness" or "flatness" of a distribution relative to the normal distribution. In Independent Component Analysis (ICA), kurtosis is used as a measure of non-Gaussianity to identify independent components in a mixture of signals. The idea behind using kurtosis is that independent components tend to have non-Gaussian distributions, while the sum of dependent components approaches a Gaussian distribution. By maximizing the kurtosis of each component, ICA separates the mixture of signals into statistically independent sources.

The ICA method was applied to the dataset, as shown below, for the letter recognition dataset. As shown below Kurtosis will increase the number of independent components for letter recognition. However, for the adult income dataset, there are two peaks. The highest value is in 55 components. Compare to the PCA method, ICA needs more components, and dimensionality reduction using ICA does not work well for these datasets.



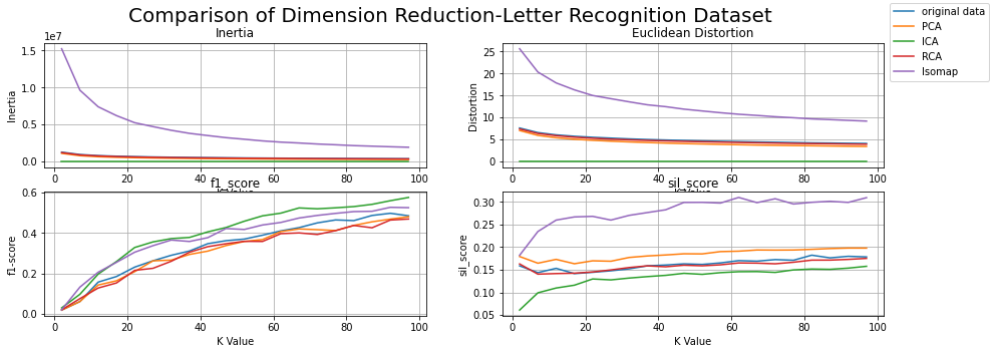
RCA and Isomap:

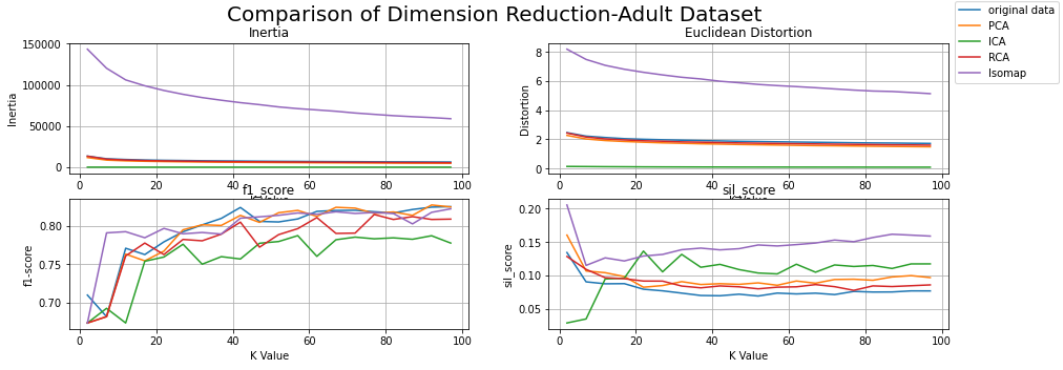
The same number of components as ICA was used for applying the RCA and Isomap, Since the figures were redundant the plots were not placed here.

6. Clustering dimension reduced data.

k-means

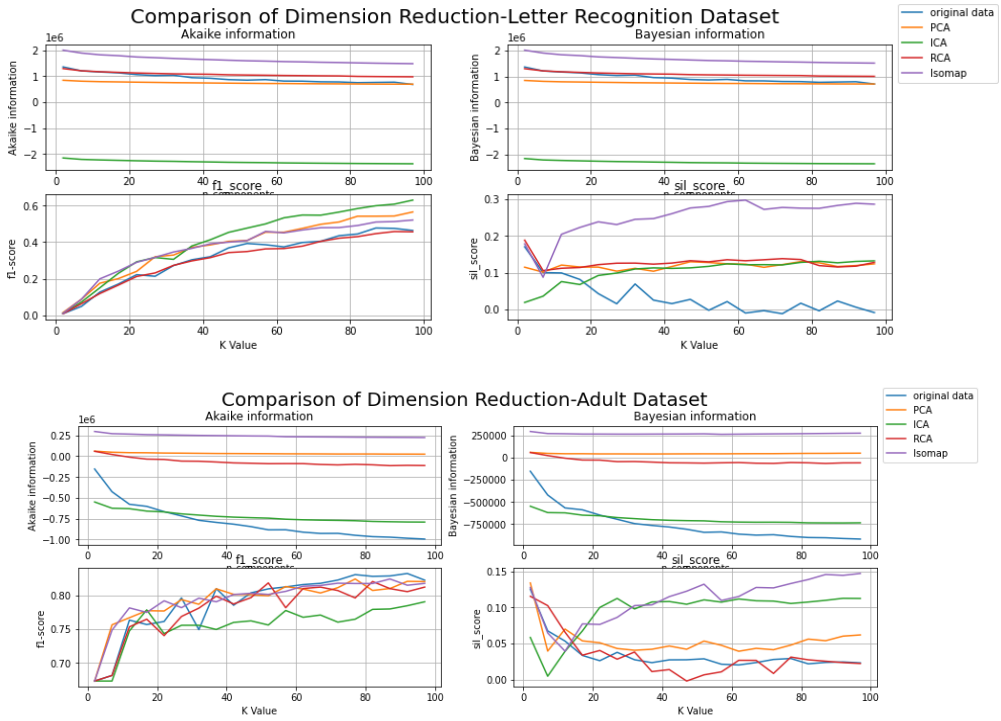
Four different dimension reduction methods were applied to the datasets. %90 variances in PCA were used for both datasets. The 8 and 55 components were used for ICA, RCA, and isomap for letter recognition and adult income dataset. Iso map does not work well with both datasets. PCA and ICA work well with letter recognition dataset and leads to very low inertia, however, PCA will lead to a higher F-1 Score.





Expectation Maximization:

Four different dimension reduction methods were applied to the datasets. %90 variances in PCA were used for both datasets. The 8 and 55 components were used for ICA, RCA, and isomap for letter recognition and adult income dataset. Iso map does not work well with both datasets. ICA works well with EM clustering. It yields the highest F-1 score for letter recognition. It increases the f1-score by %20 compared to using the original dataset. Again, Isomap does not work well here. For the Adult income dataset, the PCA method works better than the ICA method.



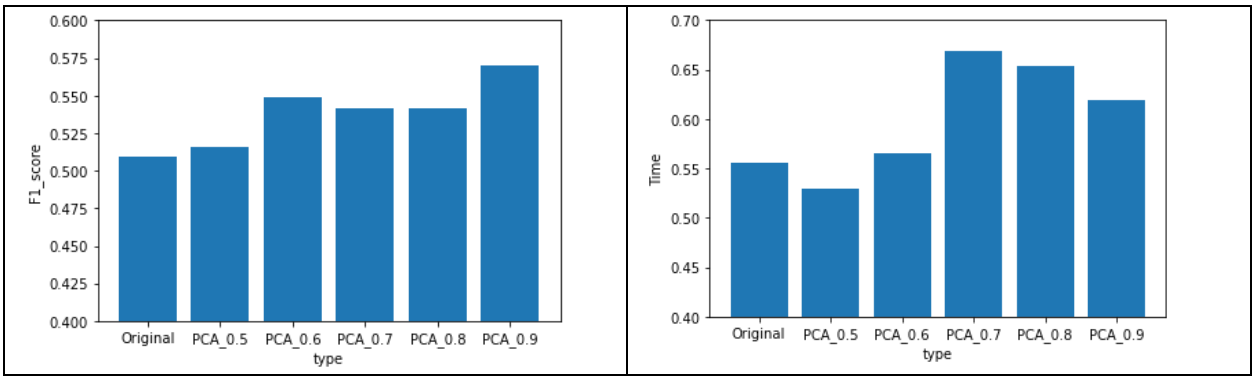
Conclusion:

Overall, the choice of dimensionality reduction method depends on the specific characteristics of the data and the goals of the analysis. Linear methods like PCA are useful for capturing the overall structure of the data, while non-linear methods like ICA and Isomap are useful for uncovering hidden structures or relationships. Random projection is useful for the fast processing of high-dimensional data with minimal assumptions.

7. ANN Model with Dimension Reduction:

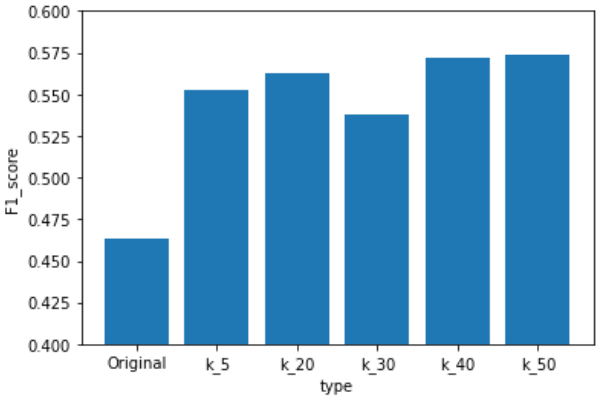
In the current example, the PCA method was applied to the adult income dataset and the f-I score on the test set was recorded for performance comparison. Model parameters include { hidden_layer_sizes=(10,2,), solver='adam', activation='logistic', learning_rate_init=0.1, max_iter=500}.

Five different levels of information variance (0.5,0.6,0.7,0.8,0.9) were used. First, the training time was reduced a little bit, as shown in the figure by decreasing the number of features(lower variance) the training time will decrease. However, the F1-score will increase by using the PCA dataset. A dataset with 0.9 variance components has the highest accuracy. This is because we have the most valuable information in fewer dimensions.



8. ANN Model with K-means cluster as a feature

The k = [5,20,30,40,50] clusters were used for k-means. The cluster labels were added as a new feature and as shown below the F-1 Score increased by almost %20. This is because the k-means add extra information to the dataset. The computation is not high but it will increase the model's accuracy considerably.



Conclusion:

Dimensionality reduction techniques can have several effects on datasets when used with Artificial Neural Network (ANN) models. Dimensionality reduction can reduce the number of features in a dataset, which can lead to faster training and testing of ANN models. This is particularly useful when dealing with high-dimensional datasets, where training a model on the full feature set can be computationally expensive. Reducing the dimensionality of a dataset can help to reduce overfitting and improve the generalization performance of ANN models. This is because high-dimensional datasets can be noisy, and reducing the number of features can help to remove some of the noise and irrelevant information.

One potential drawback of dimensionality reduction is that it can lead to a loss of information. This is because some of the features that are removed may contain important information for the model. Therefore, it is important to carefully choose the dimensionality reduction technique and the number of features to retain. There is often a trade-off between the dimensionality reduction and the performance of the ANN model. Depending on the dataset and the problem, reducing the number of features may improve or degrade the performance of the model. Therefore, it is important to carefully evaluate the performance of the model before and after applying dimensionality reduction techniques.