Assignment 3: Unsupervised Learning and Dimensional Reduction

Shailesh Tappe: stappe3@gatech.edu

Objective: The objective of the assignment is to analyze and explore different unsupervised learning algorithm with two different datasets. An analysis is performed on each learning algorithm i.e. k-means clustering and Expectation Maximization and reduce dimensionality by applying feature reduction algorithms principal component analysis (PCA), independent component analysis (ICA), randomized projections (RP) and also experimented with Random Forest Classifier. Another objective of experiment is to apply these feature reduction algorithms to neural network algorithm and observe its performance in terms of cross validation accuracy and prediction time.

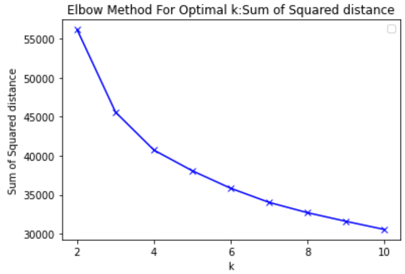
Datasets: Same as in assignment 1, the experiment uses two datasets for the analysis i.e. Wine Quality dataset and Wisconsin Breast Cancer dataset from UCI library and openml.org dataset repository.

1. **Wine Quality dataset:** The wine quality datasets are variant of Portuguese 'Vinho Verde' wine, sample taken from a study performed by University of Minho in Portugal. Data contains 6497 instances and 12 attributes with classification attribute "*quality*". In preprocessing quality rating of wine is set to low quality (i.e. 0 for rating less than 6) and high quality (i.e. 1 for rating 6 or above).

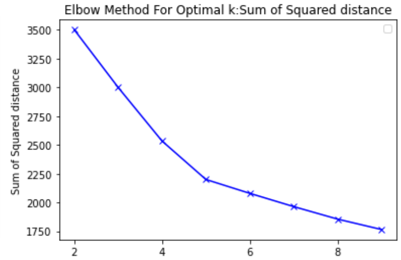
2. **Wisconsin Breast Cancer dataset:** The data was originally created by Dr. William Wolberg from University of Wisconsin Hospital in Madison WI. Data features are from digitalize image of FNA (fine needle aspirant), and represent feature characteristic of cell nuclei presented in the image. Data contains 699 instances with classification attribute "*Class*". Some of the data values are missing from datasets (i.e. marked as '?' instead of numeric value for attribute). In preprocessing all missing values are replaces with propagated fill forward values using python panda library.

Experiments: In unsupervised learning clustering is to analyze data instances and derive it with similar groups i.e. measure objects and see closeness with defined distance metrics. The experiments were performed to analyze performance of different unsupervised clustering algorithms. Experiment was developed in python and scikit-learn library, pandas and numpy library to analyze clustering algorithms k-mean clustering and expectation maximization (EM).

1. **K-Means Clustering:** In K-means clustering with given set of data, K data points randomly gets selected and each K computes closest neighborhood point with distance metrics and group them together. K-Means algorithm repeats the process until it converges and to get optimal K numbers of clusters for a data set. That is, data points distance to its centroid is minimal. Similar to optimization algorithm hill climbing, K-means uses closeness to neighborhood point to recompute centroid and forms cluster groups.
   Analysis for K-mean was performed with different metrices to find optimal cluster value for both datasets.

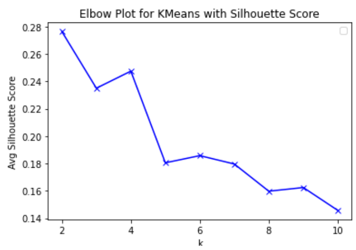


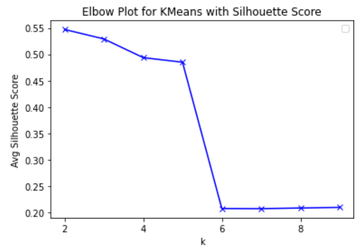
Wine Quality Dataset



Breast Cancer Dataset

Elbow methodology was used to analyze sum of squared error (SSE) for analyzing optimal K cluster value from the range of clusters in both datasets. That it, if straigth line drwan is between first and last cluster is draw, then longest distance that line is considered elbow point. This elbow point is generally consider as optimal number of cluster for dataset. Looking at graph SSE is descending and elbowed out near 4 and 5 for wine quality dataset and 5 for breast cancer dataset for opitmal k value of cluster.

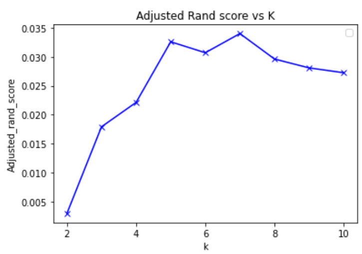


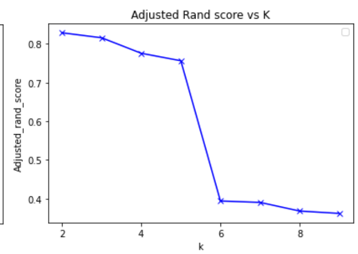


Wine Quality Dataset

Breast Cancer Dataset

Silhouette score within a cluster calculates mean of similarities between data points in the same cluster to the data points in the next closest cluster. Score values of sinhouette ranges from -1 to 1 and higher the value is better. Observing sinhouette score for wine quality dataset, 4 is optimal value for K, eventhough higher score is for K=2. But score is descending before it goes back up again at K=4. Similar observations can be made for breast cancer dataset and K is optimal close to 5 using sinhouette score. Ideally highest peak of score is indicator is optimal K value, but datasets need more instances to generalize the observation.





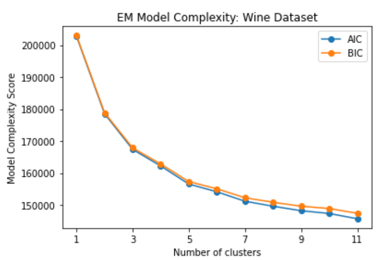
Wine Quality Dataset

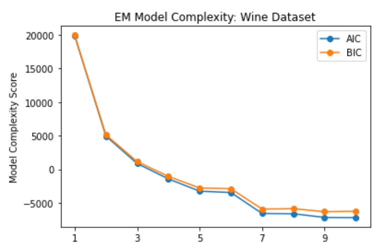
Breast Cancer Dataset

Observing adjusted rand index, which computes similarities in 2 clustering assignment, it is found that K values are peaking at 4 and 8 for wine quality dataset and for breast cancer dataset it stabilizes around K value of 4 and 5. Experiment needs more data instances to generalize ARI index.

Although more data instances and features could be helpful in generalizing K-mean clustering for both datasets, observing them with above explained metrices, K is close to 4 for wine quality dataset and 5 for breast cancer dataset.

2. **Expectation Maximization (EM):** Unlike K-mean clustering, expectation maximization models with gaussian probability distribution. Algorithm first evaluates data points probabilities from a gaussian cluster, then maximize cluster centroids using gaussian probability of another cluster. Analysis for EM was performed with different metrices to find optimal cluster value for both datasets.

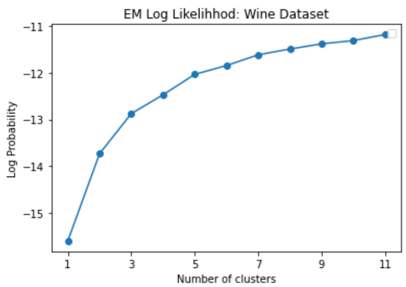


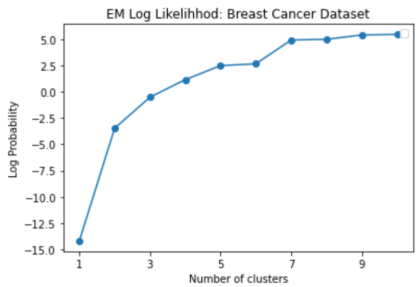


Wine Quality Dataset

Breast Cancer Dataset

To analyze model complexity score for clusters AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are used to gauge model likelihood. Elbow method is used to find optimal value of K for both AIC and BIC. Observing model complexity for both datasets and measured using AIC and BIC, it is found that for both dataset optimal value for is near to 5 and 7.

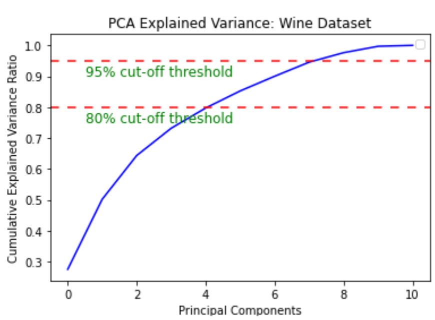




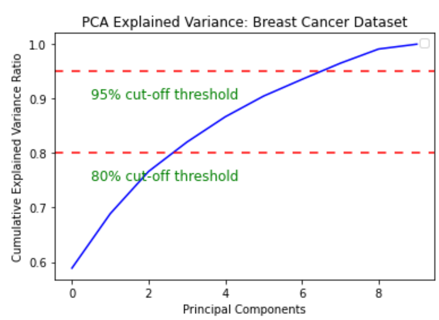
Observing log likelihood (log probability) i.e. probability of conformed data in the model, it is observed that maximum log likelehood is at around 7 for both datasets.
So observing both metrices it is found that for EM optimal value for cluster is near 6 and 7 for both datasets.

Dimensionality Reduction and Clustering: Dimensionality reduction is useful in reducing or restructuring features from dataset. This reduce dimensional data is then used to train any learning algorithm like discussed in supervised learning such as decision tree, neural network, boosting, SVD, KNN. One of the main benefits is to reduce complexity in modelling algorithm generally geared to curse of dimensionality i.e. with increase dimension search space expands exponentially. The experiment looks for some of the dimensionality reduction (feature reduction) techniques such as principal component analysis (PCA), independent component analysis (ICA), randomized projection (RP) and random forest classification to reduce number of features from classification.

1. **Principal Component Analysis (PCA):** PCA technique reduces dimensionality of data into independent data components in dataset while keeping most variances of datasets.

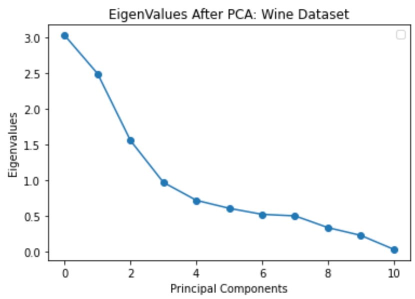


Wine Quality Dataset

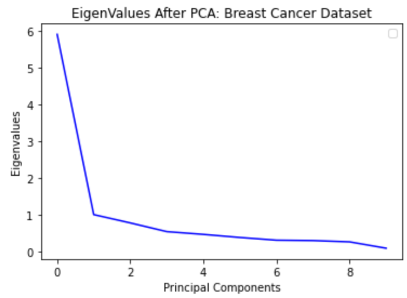


Breast Cancer Dataset

Observing variance threshold of 80% to 95% with number of components, in wine quality dataset PCA feature reduction constitutes between 4 and 7 out of total 10 features, but closer to 6 while maintaining 90% of variance. Similarly, for breast cancer dataset features can reduce till 4 from total 10 features, while maintain 90% of variance.



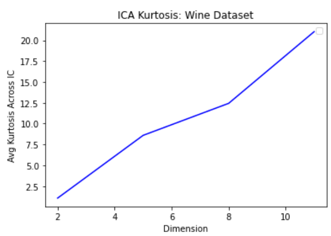
Wine Quality Dataset



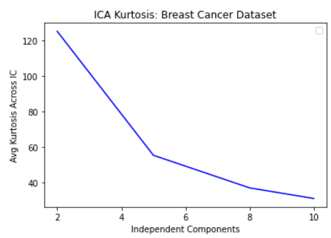
Breast Cancer Dataset

Using elbow method for eigenvalue with components for PCA, it is observed that for wine quality dataset feature reduction levels off near 4 and 5, while for breast cancer dataset close to 4. So observing PCA optimal number of feature reduction is consistant with both mertices.

2. **Independent Component Analysis (ICA):** Contrary to PCA, independent component analysis (ICA) disassociates maximum variant sets of components into individual component.



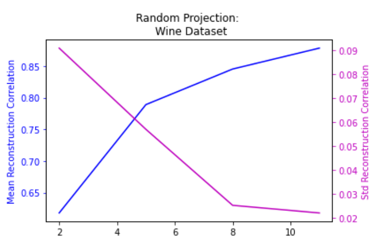
Wine Quality Dataset



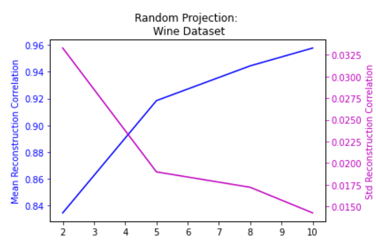
Breast Cancer Dataset

Observing kurtosis for both datasets, wine quality have high kurtosis for max dimension of 10, but for breast cancer dataset kurtosis is high at low dimension of 1. That indicates, components wine quality datasets are independent of each other and can help cluster data, but for breast cancer dataset, with low dimension kurtosis indicates, data can be clusstered with combination of features.

3. **Randomized Projection:** Randomized projection technnique reduces features by projecting data in euclidean space on matrix projectec ramdomly in gaussian distribution.



Wine Quality Dataset

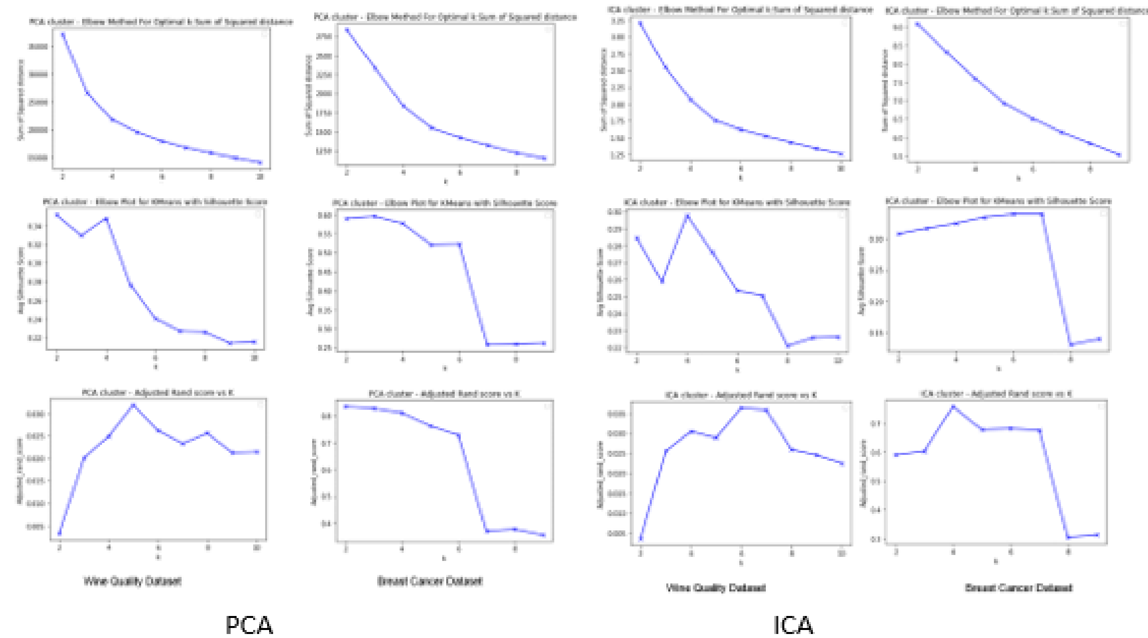


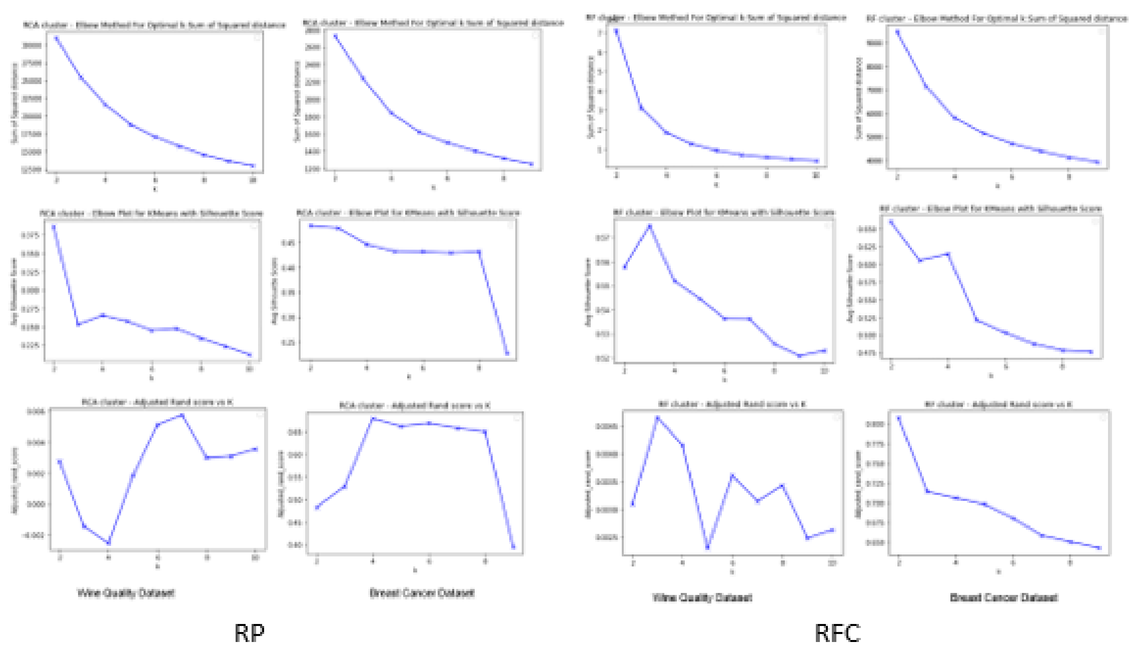
Breast Cancer Dataset

Observing reconstruction error for both datasets, mean of reconstruction error decreases with increased dimension and level off close to component 8 for wine quality dataset and 6 to breast cancer dataset.

4. **Random Forest Classification:** With random forest classification fetures were selected by building decision trees and ensamble them for classification. Feautre importance was calculated using cumulative sum of important features and thresholded to 87%.

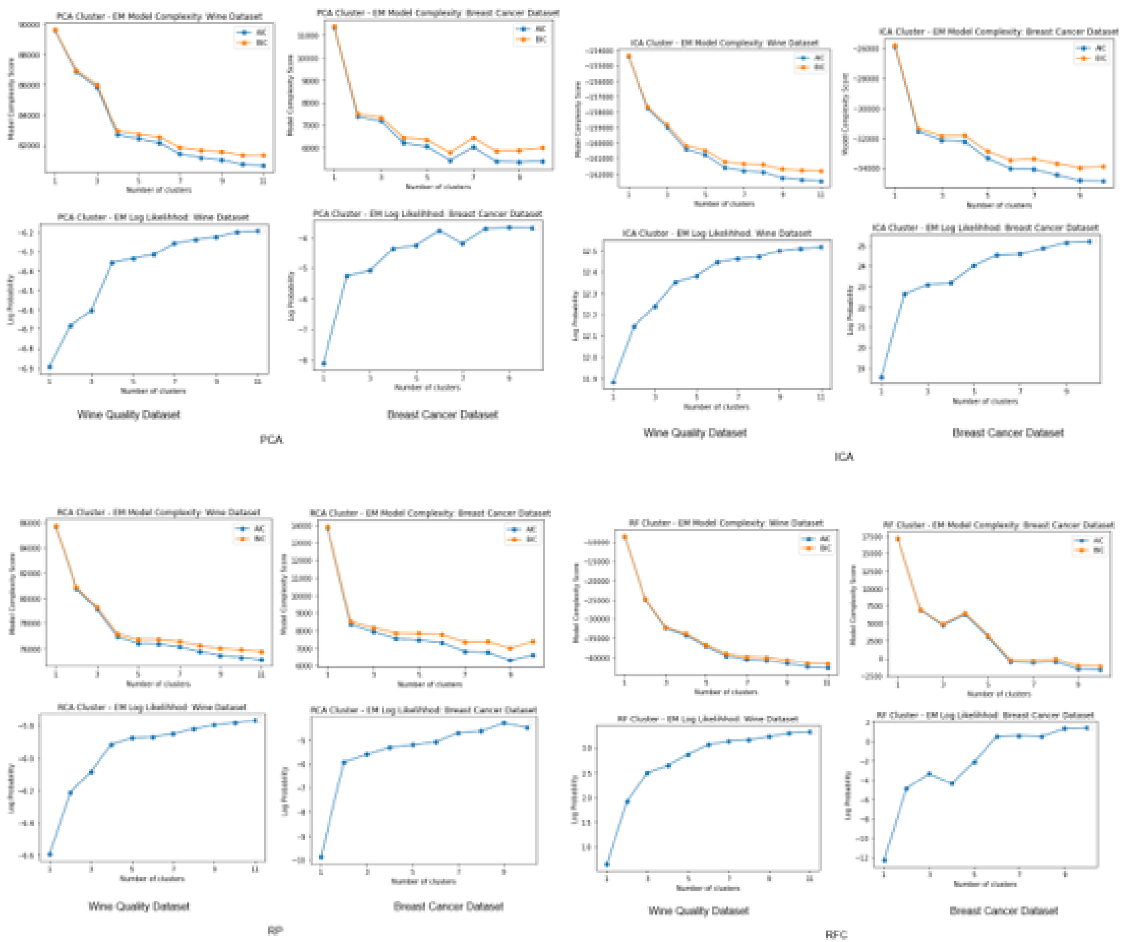
Clustering Analysis for PCA using K-Means Clustering:





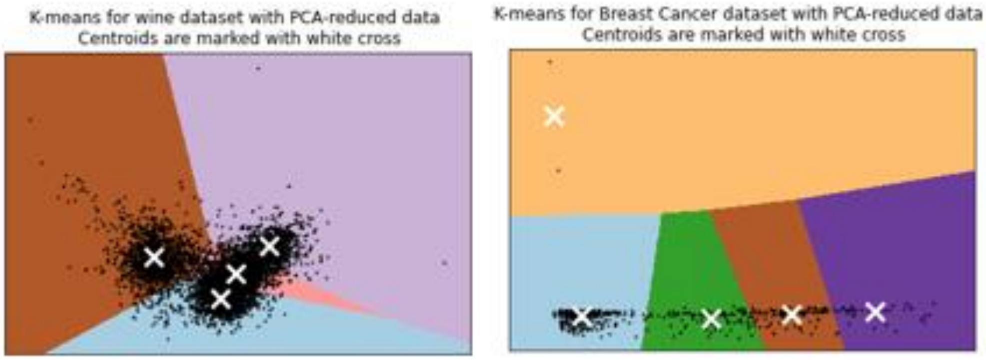
K-means clustering technnique were applied to dimensional reducede data and compared with full dataset. Observing and comparing it with full dataset it is found that in general optimal values for clusters are very much same in both the cases. But one in perticular i.e. silhouette score has improved with reduced dimension and getting higher value close to optimal value of K in both dataset. Similar observation is for ARI index. These observations hypothesies that even with reduced dimension is perfoming well in identifying optimal K and little better in observing silhouette acore and ARI index.

Clustering Analysis for PCA using Expectation Maximization:

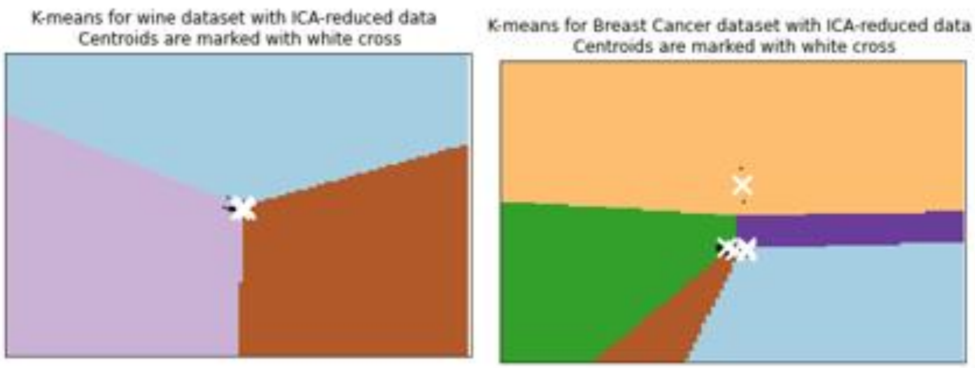


Expectation Maximization clustering technnique were applied to dimensional reducede data and compared with full dataset. Model complexity in reference to AIC and BIC with number of clusters is similar to full dataset, same is log likeliness is similar, but improved with PCA and RFC for wine quality dataset.

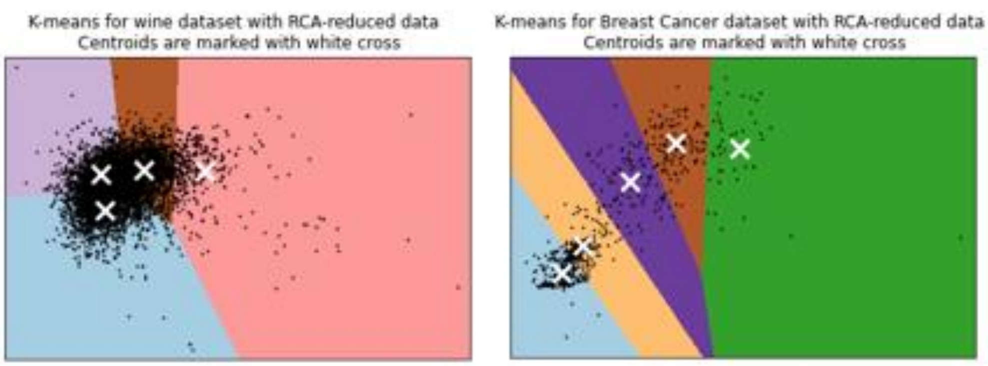
Observing Centroids for Dimensionally Reduced technique: After optimal cluster was indetified cluster were formed with its centroid. Looking at clusters and centroids PCA and RP converging cluster and most of data points are close to centroids, but for ICA clusters are close together and data points can belong with any other cluster and may not be optimal in classifying data. Howevewr for RFC for wine quality datasets data is more descritized and not providing optimal K-means for number of clusters. More features and data sets can help generalize pattern in clustering og data. Over all PCA and RP are close to convergance.



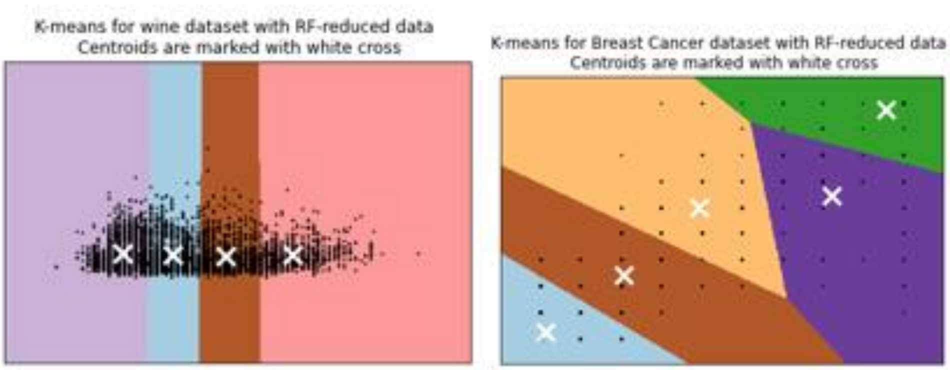
PCA



ICA



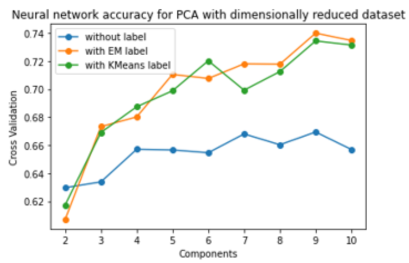
RP



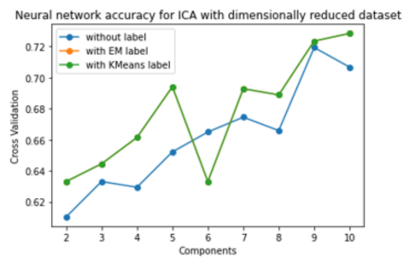
RFC

Wine Quality Dataset **Breast Cancer Dataset**

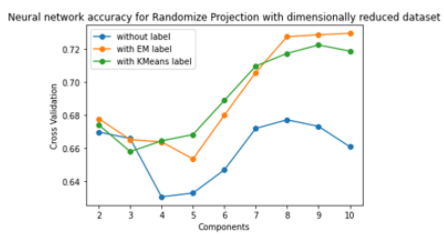
Neural Network Analysis for dimensionality reduced data and clustered data for Wine Quality dataset: With this experiment, objective is to apply dimensionally reduce technique to datasets used in experiment 1 (wine dataset was used for the experiment) and learn neural network algorithm with dimensionally reduced dataset withour cluster label and dimesionally reduced dataset eith clustered label.



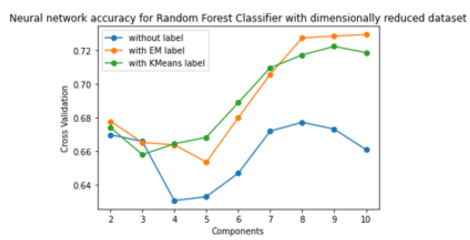
PCA



ICA



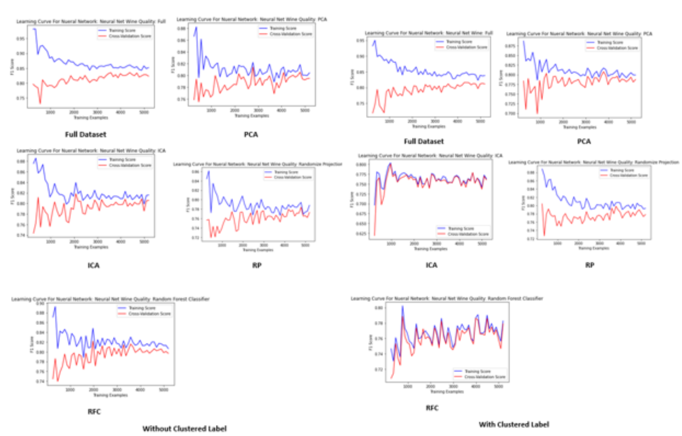
RP



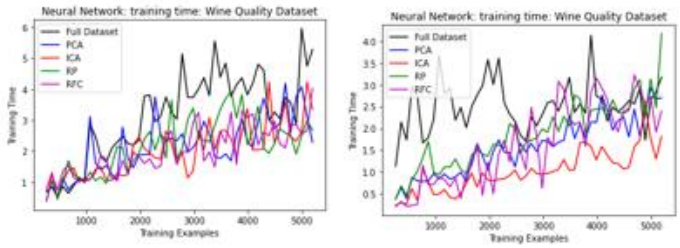
RFC

In the experiment all 4 dimensionality reduced algorithm and with clustering label, it is observed that cluster accuracy is higher to accuracy without cluster, indicates better accuracy of NN for labeled dataset.

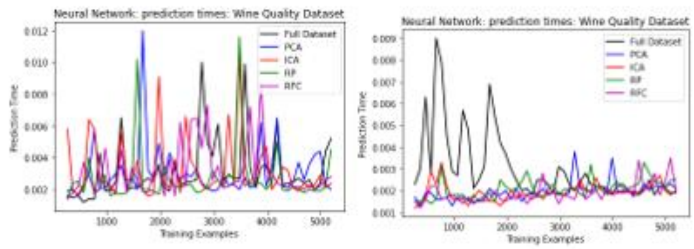
Other part of experiment was to run NN for full dataset and dimeisionally reduced dataset with all 4 techniques and observe cross validation score i.e. learning curve, cross validation curve, prediction time and fit time.



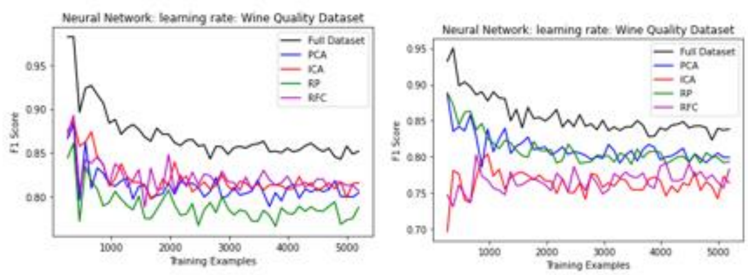
Observing learning and cross validation curve for all datasets (full and reduced) with and without cluster label, NN is converging well with high F1 score, except for ICA anf RFC reduction with clustered data where NN is underfitting. This means for these technique K needs to be optimized and needs more dimensions and hyperparameter tunning.



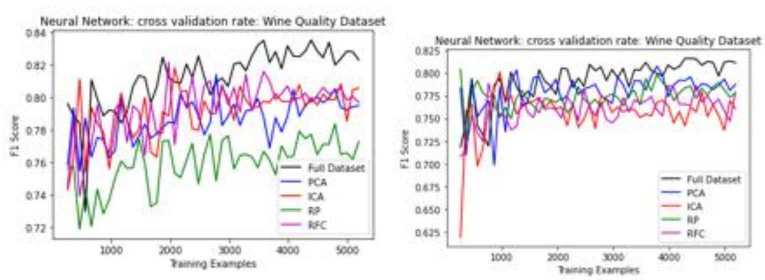
Training Time



Prediction Time



Learning Curve



Validation Curve

Without cluster lavel With cluster label

Observing with and without cluster labeled reduced dimension dataset, it is found that for wine quality datasets with reduced dimension there is no significant reduction in training and prediction time for withour cluster dataset, but dataset instances are low and can lead to significant time reduction in more instances of data. However with clustered labeled there is significant improvement in time for reduced dataset as compare to full dataset. Also with reduced dimension of data there is no significant loss in learning and cross validation curve. It significes that even with reduced data neural network is perfoming well with all redution techniques. This indicates that applying reduced dimensinal datasets to different learning algorithm such as neural network, decision tree, SVD etc. is useful in generalization and also reduces complixity and time singificantly.

Conclusion: Overall project experience was unique and interesting. Although I got good hands on with this exercise, I'm still not done with learning more on clustering and feature reduction. Honestly these experiments are very good and helps in converging concept very well, but for me I still need more analysis to get converged well with the concept. I intent to study and explore more on different datasets, possibly multi classification datasets and observe patterns around it

.

References:

- Tim Mitchell: book – Machine Learning
- Dr. Charles Isabell and Dr. Michael Littman: video lectures supervised learning chapter SL1 – SL10, UL1 – UL4
- Dr. William Wolberg from University of Wisconsin Hospital
- Paulo Cortez, University of Minho, Guimarães, Portugal
- UCI Machine Learning Repository
- Openml.org
- scikit-learn library and documentation
- TA's during office hour and piazza forum