CS 7641: Machine Learning
Assignment 3: Unsupervised Learning and Dimensionality Reduction

I have used German Credit data and Australian Credit Card Approval data that I had used in assignment 1 for this assignment too.

Description of Classification Problems

1. **German Credit Data :** The German credit data is used to classify people described by a set of attributes as good or bad credit risks. There are 1000 instances and 20 attributes. The attributes are both categorical and numerical. The target variable consists of only two classes, 1 and 0 with 1 being as bad risk and 0 being good. Although this is a classification problem, I am applying the clustering algorithms and dimensionality reduction techniques on the data excluding the given labels and then comparing the results produced by the clustering algorithms to the actual distribution of the two different classes present in the data.

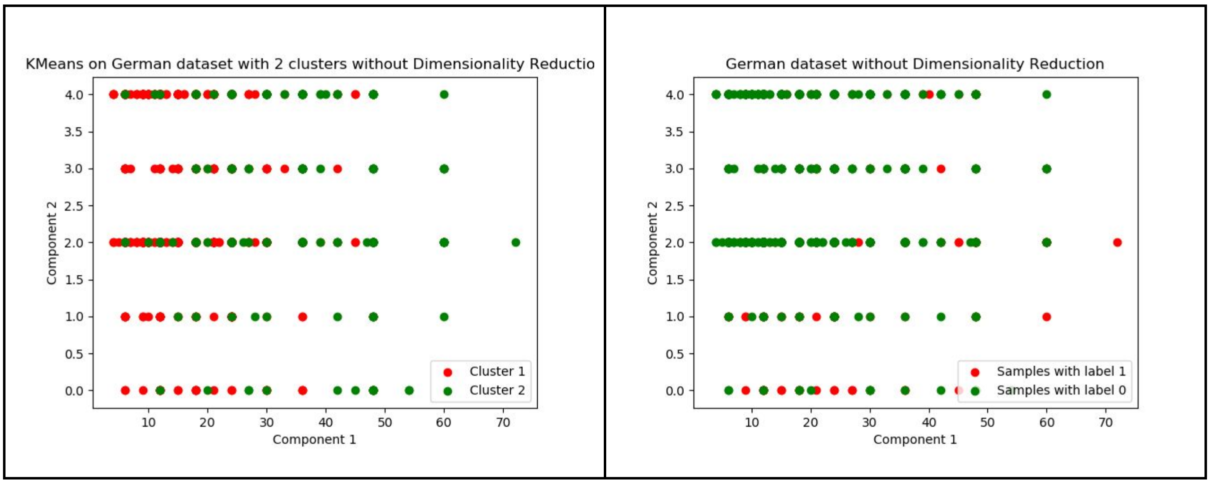
2. **Australian Credit Card Approval Data :** This dataset consists of credit card applications and their corresponding labels wherein they have either been approved or denied. There are 690 instances and 14 attributes. The attributes are both categorical and numerical. This dataset does not have data imbalance problem, the instances with label 1 (people whose credit card application got denied) and the instances with label 0 (people whose credit card application got approved) are almost equal in number, that is, 383 instances are with label 0 and 307 instances are with label 1. The evaluation metric that I chose for comparing the different classification algorithms is accuracy for this dataset. Since, the two classes of the target variable are almost equally present in the dataset, accuracy can give a good idea of the performance of each algorithm. This is also a classification problem and even though labels for each sample are available, I am applying the clustering algorithms and dimensionality reduction algorithms on the data excluding the already given labels and comparing and the results produced by them to the actual distribution of the given labels.

Explanation of methods for choosing ‘k’ in KMeans clustering, ‘n_components’ in Expectation Maximization and the number of features for dimensionality reduction algorithms.

For KMeans clustering and Expectation Maximization I used Silhouette score to choose the value of ‘k’ and ‘n_components’ respectively. I ran the clustering algorithms on the two datasets with different values of ‘k’ and ‘n_components’ and selected the values corresponding to the highest Silhouette score. I decided to use Silhouette score as a metric because I had the correct labels. In case of the dimensionality reduction algorithms namely Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections (RP) and the fourth one that I selected, Recursive Feature Elimination (RFE), I decided on the value of the number of components as 2 since 2 components can be used to visualize the results using a scatter plot. I tried using more components and then plotted the various combinations of those components but I always got the best clustering visualization corresponding to the first two components as those were the best components given by the above mentioned algorithms.

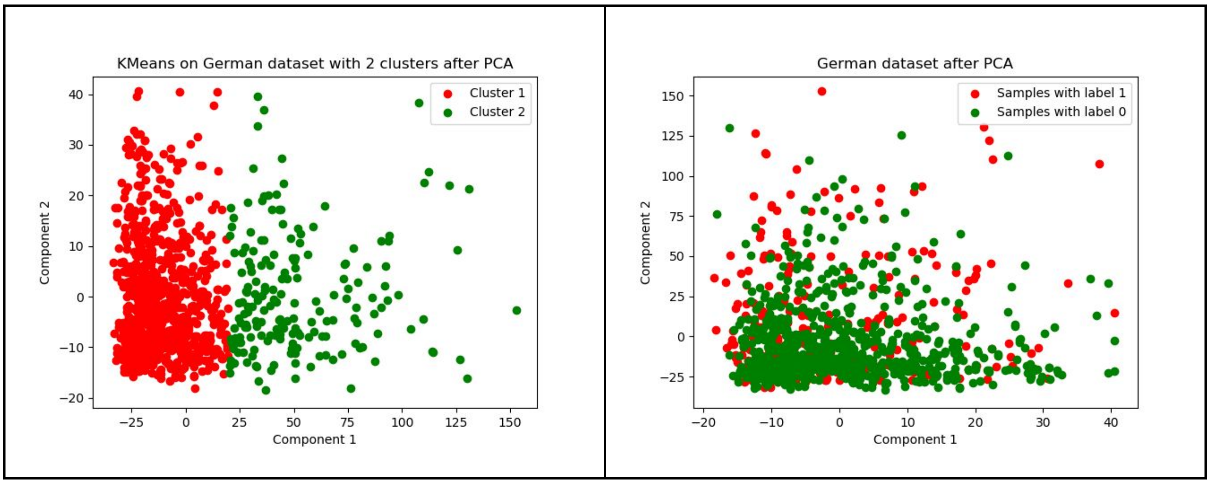
Description of the Clusters

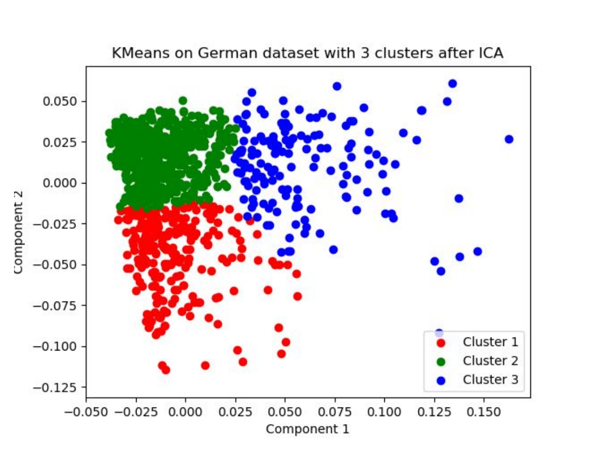
German Dataset



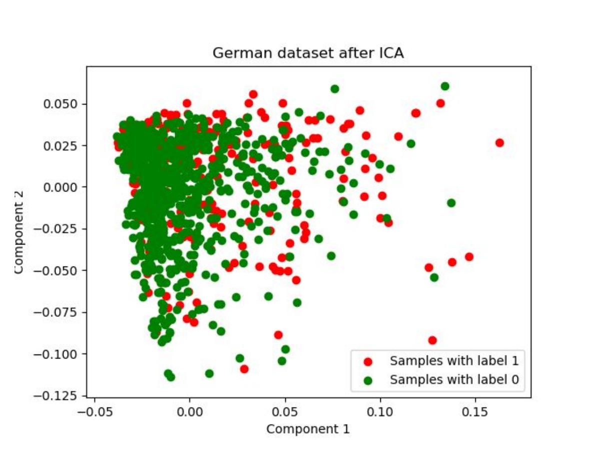
The image on the left shows the result of KMeans clustering on German dataset without performing Dimensionality Reduction and the image on the right is the actual distribution of the two classes in German dataset without any dimensionality reduction.

Note: For the graphs that follow similar structure will be followed. Image on the left will be the result of Clustering algorithm and the one on the right will be the actual distribution of the data.

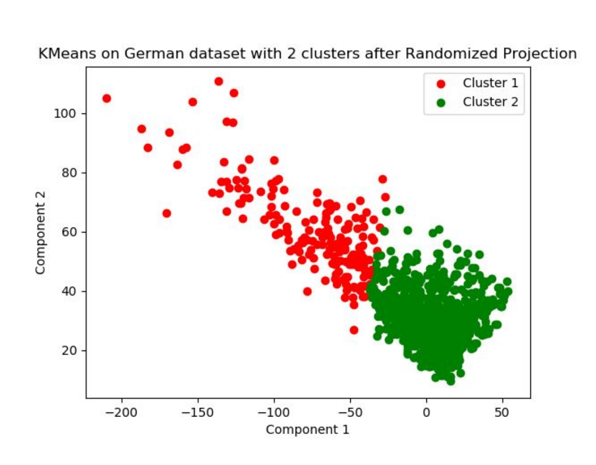




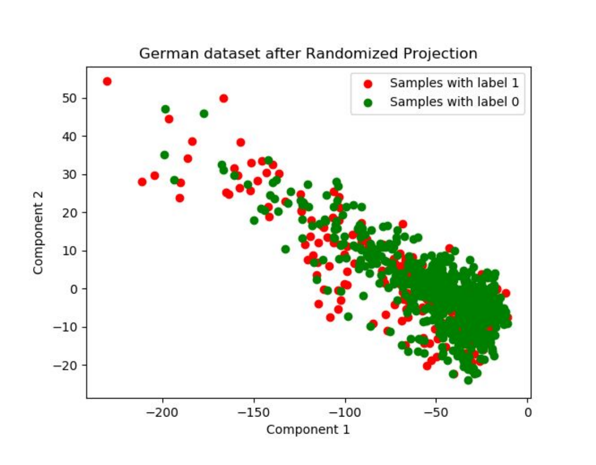
KMeans on German dataset with 3 clusters after ICA



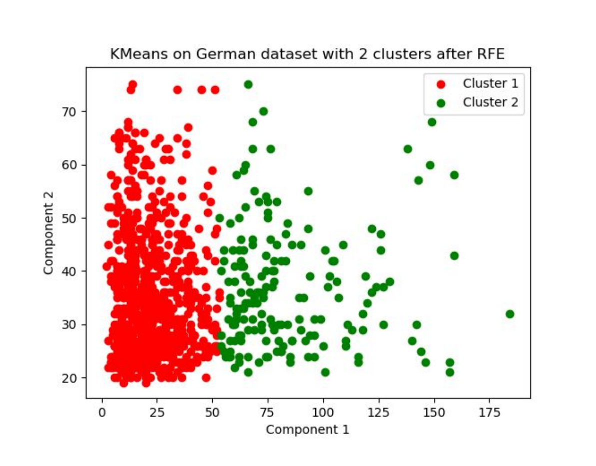
German dataset after ICA



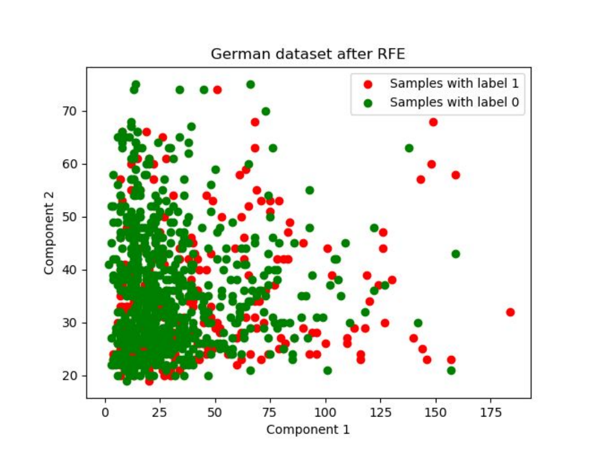
KMeans on German dataset with 2 clusters after Randomized Projection



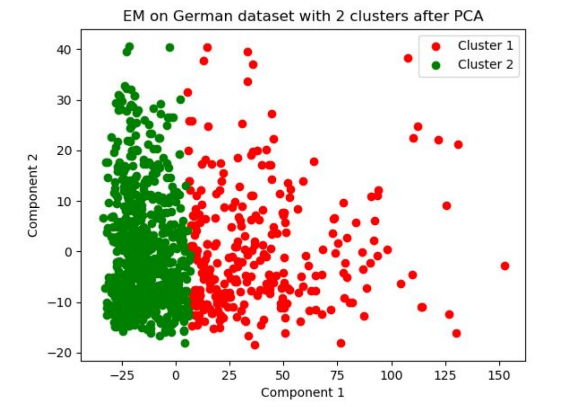
German dataset after Randomized Projection

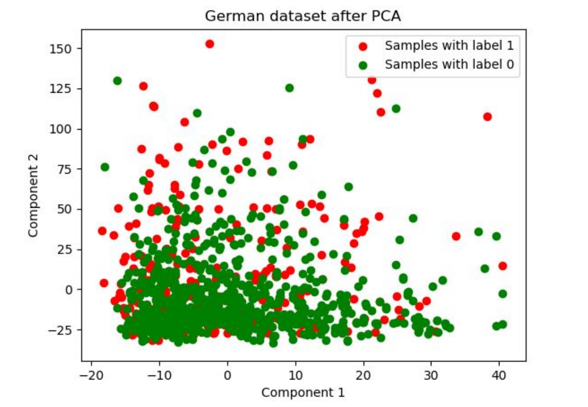


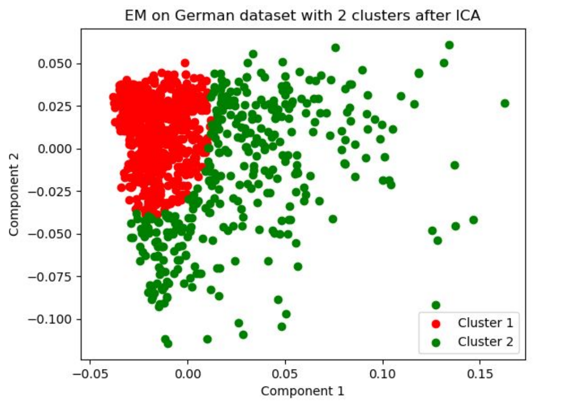
KMeans on German dataset with 2 clusters after RFE

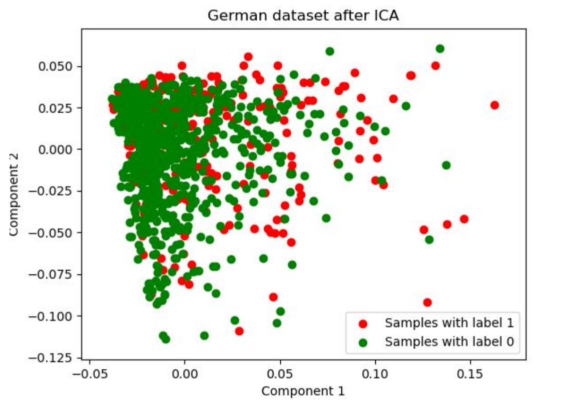


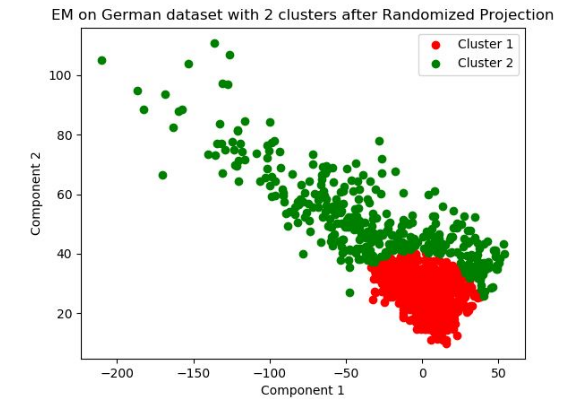
German dataset after RFE

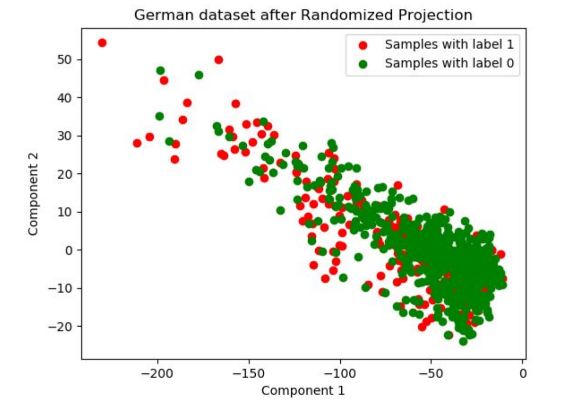

EM on German dataset with 2 clusters after PCA
Cluster 1
Cluster 2
Component 1
Component 2
−25
0
25
50
75
100
125
150
−20
−10
0
10
20
30
40

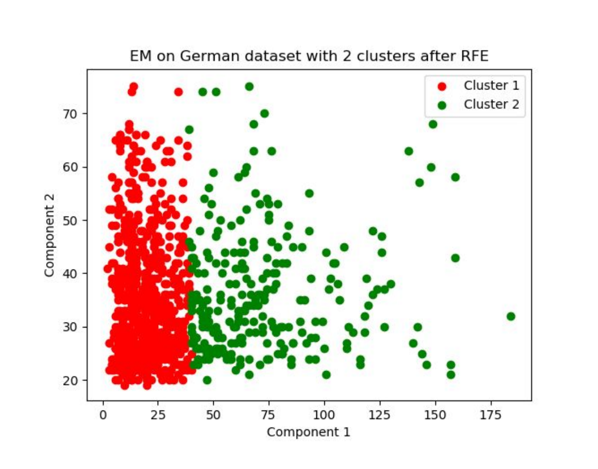

German dataset after PCA
Samples with label 1
Samples with label 0
Component 1
Component 2
−20
−10
0
10
20
30
40
−25
0
25
50
75
100
125
150

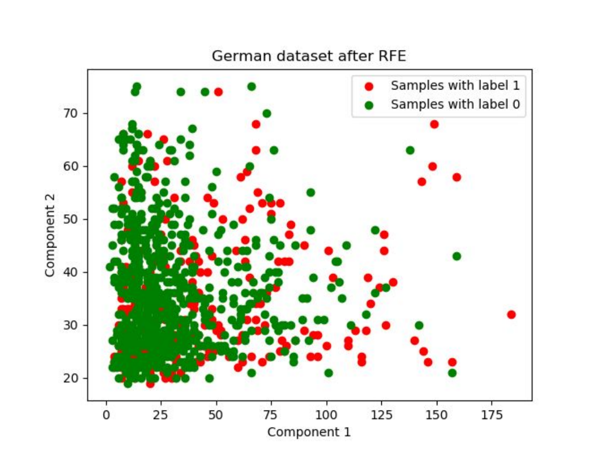

EM on German dataset with 2 clusters after ICA
Cluster 1
Cluster 2
Component 1
Component 2
−0.05
0.00
0.05
0.10
0.15
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100


German dataset after ICA
Samples with label 1
Samples with label 0
Component 1
Component 2
−0.05
0.00
0.05
0.10
0.15
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
−0.125

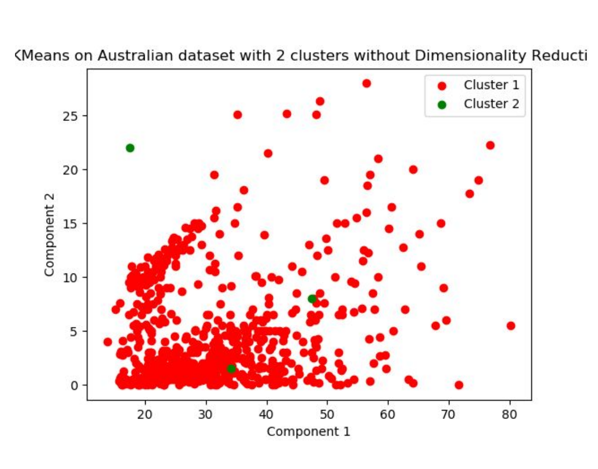

EM on German dataset with 2 clusters after Randomized Projection
Cluster 1
Cluster 2
Component 1
Component 2
−200
−150
−100
−50
0
50
20
40
60
80
100

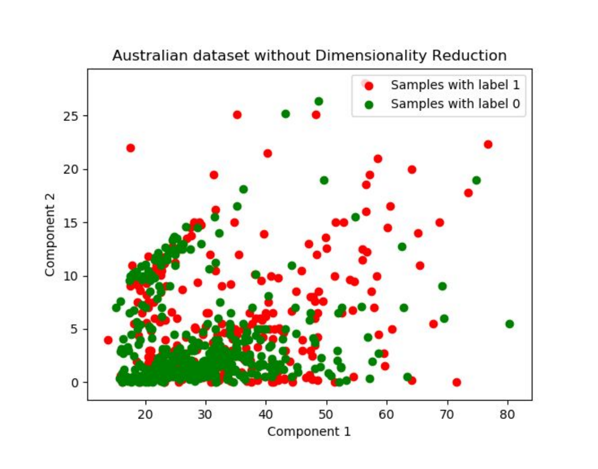

German dataset after Randomized Projection
Samples with label 1
Samples with label 0
Component 1
Component 2
−200
−150
−100
−50
0
−20
−10
0
10
20
30
40
50

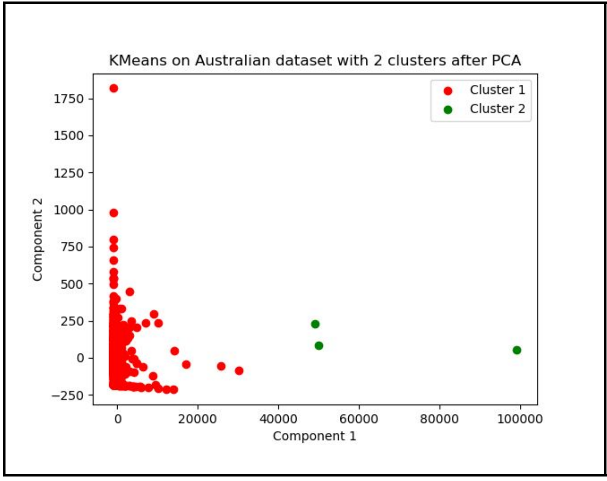


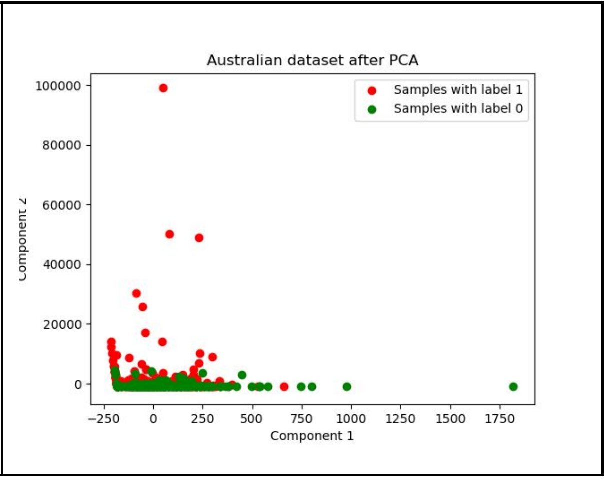


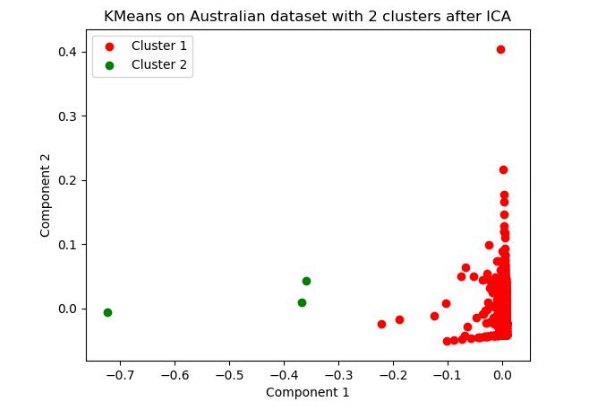
Australian Dataset

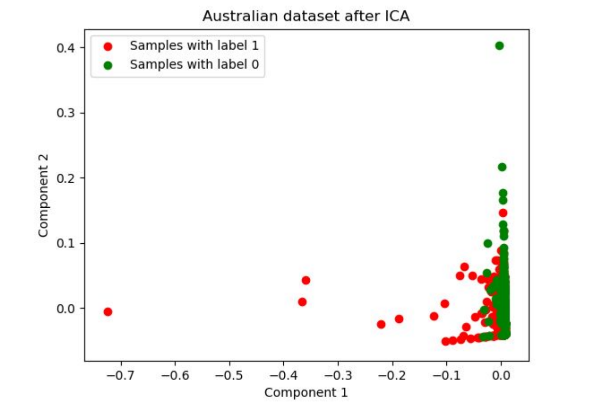


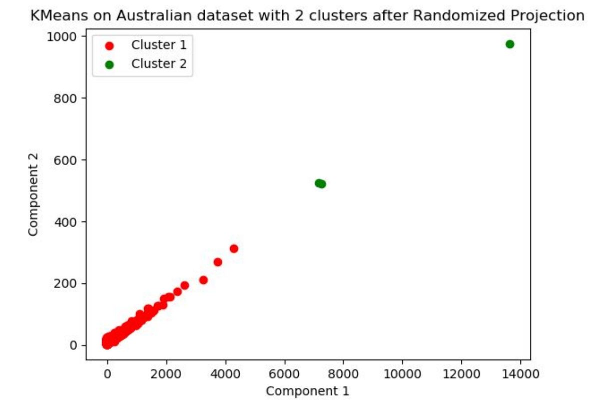


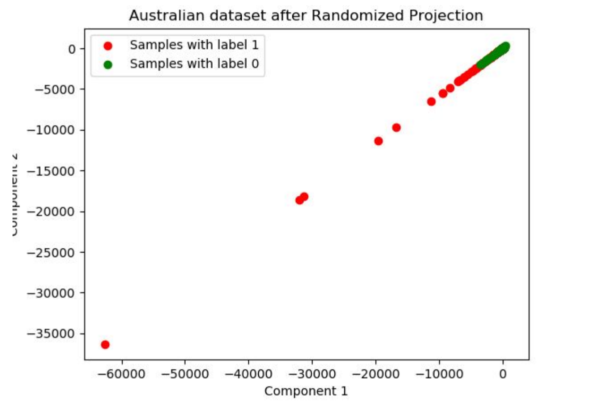


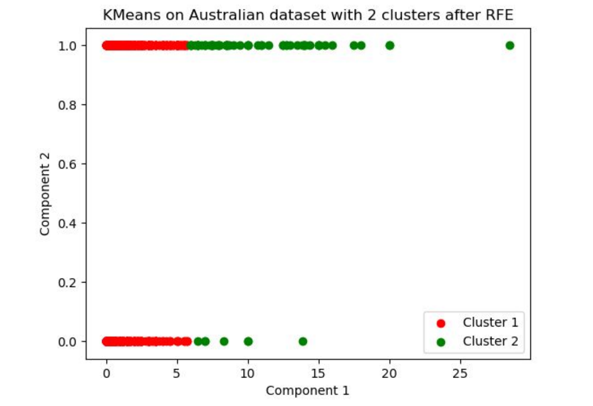


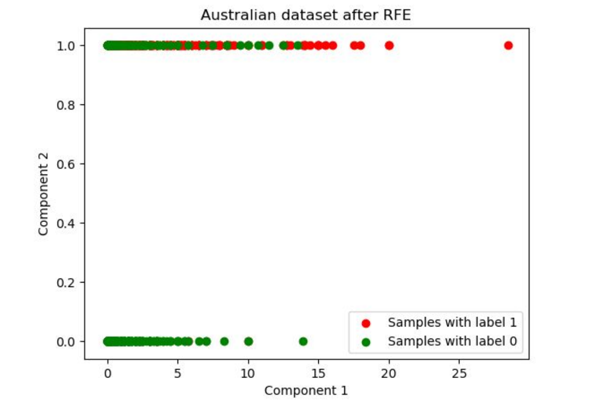

KMeans on Australian dataset with 2 clusters after ICA

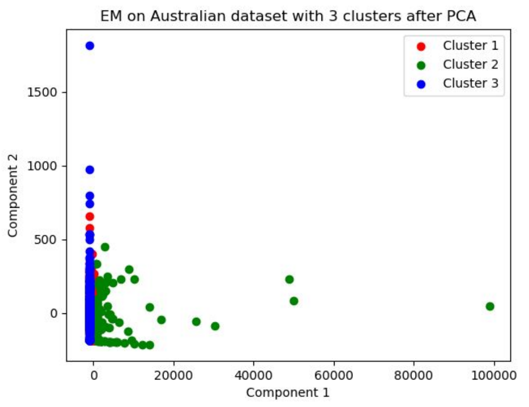

Australian dataset after ICA

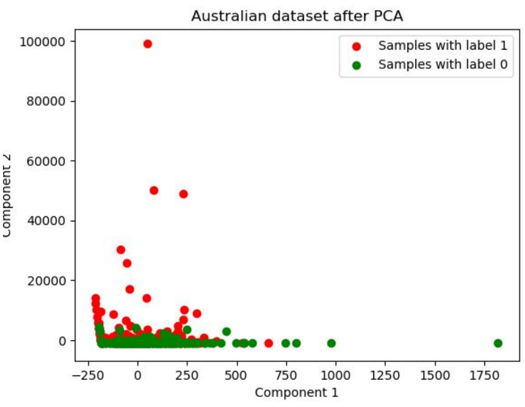

KMeans on Australian dataset with 2 clusters after Randomized Projection

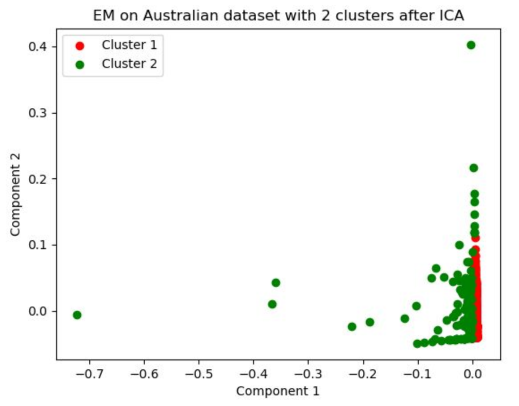

Australian dataset after Randomized Projection

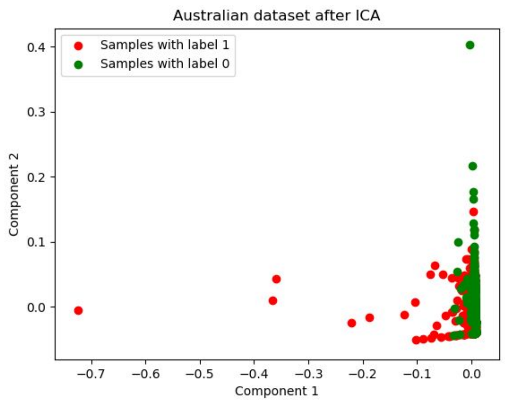

KMeans on Australian dataset with 2 clusters after RFE

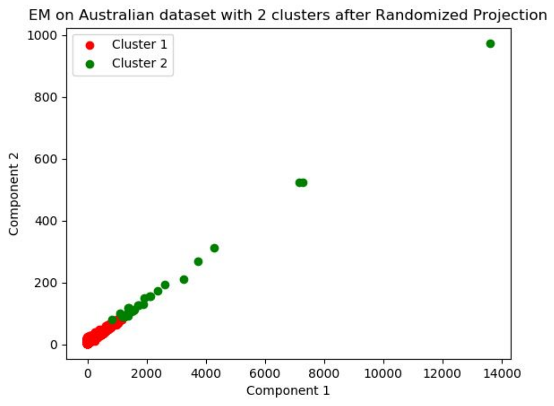

Australian dataset after RFE

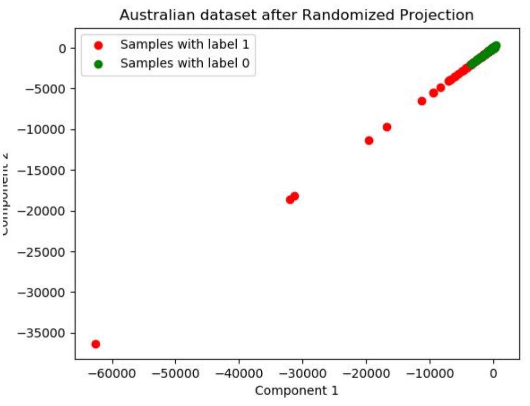

EM on Australian dataset with 3 clusters after PCA

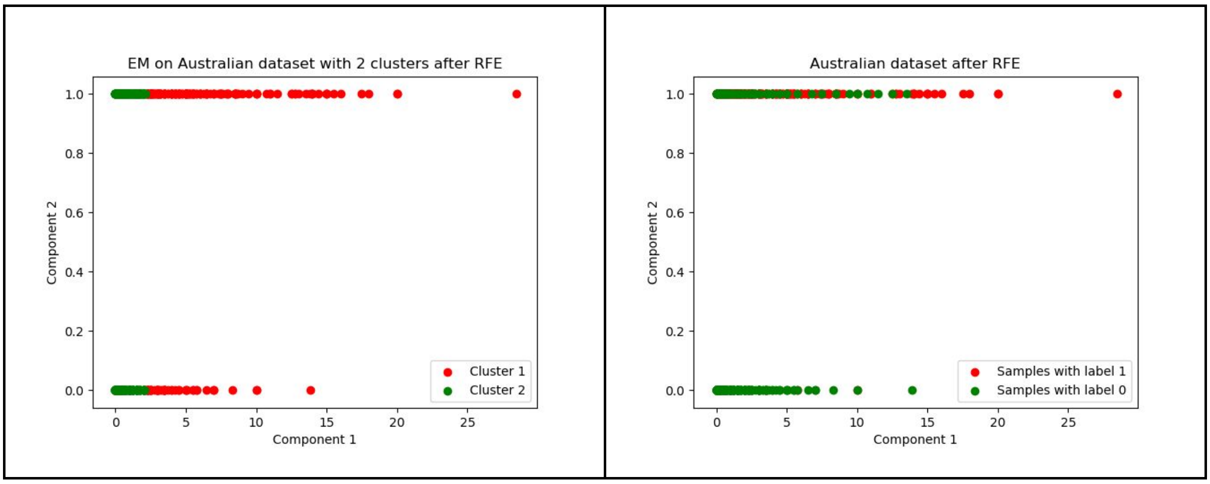

Australian dataset after PCA


EM on Australian dataset with 2 clusters after ICA


Australian dataset after ICA


EM on Australian dataset with 2 clusters after Randomized Projection


Australian dataset after Randomized Projection

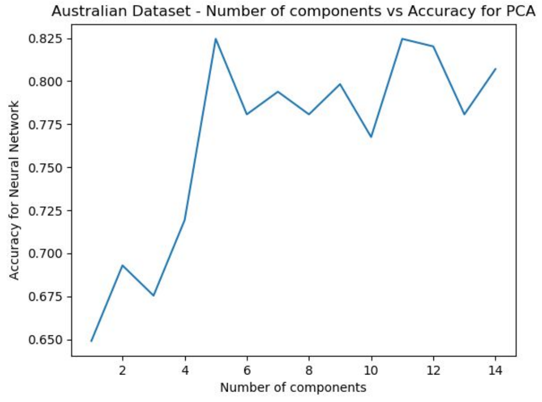


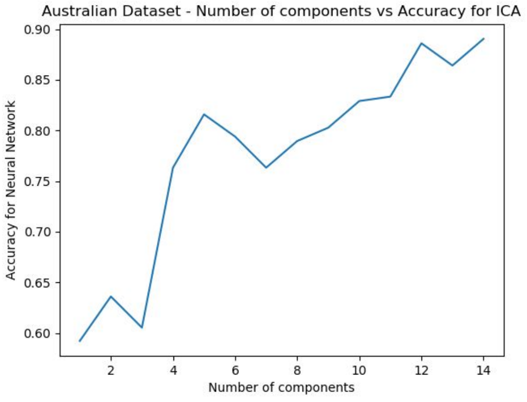
Analysis of the results of Clustering algorithms and Dimensionality Reduction algorithms

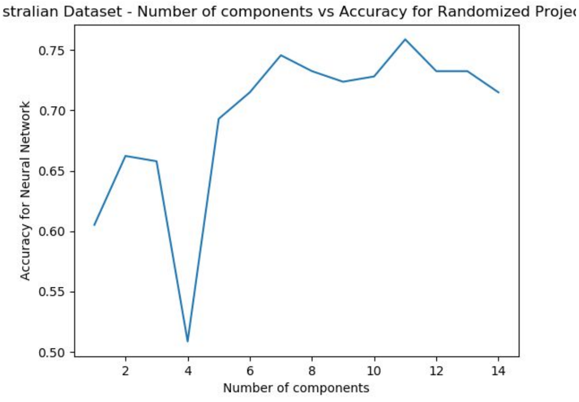
Australian and German Dataset : Without dimensionality reduction algorithms both KMeans and EM performed poorly as we can clearly see that the identified clusters are far from accurate when compared to the actual distribution of data. This might be happening because the clustering algorithms take into account features that contribute less significantly towards classification but are given equal weightage by the clustering algorithms. After applying dimensionality reduction algorithms the quality of the clusters identified by both KMeans and EM improves significantly especially after using ICA and Randomized Projections. In terms of the quality of clusters EM has identified clusters better than KMeans. This could be due to the fact that the datasets are highly mixed up and assigning samples to clusters only on the basis of euclidean distance is not sufficient but rather a soft clustering algorithm like EM is better able to approximate the probabilities of samples belonging to different clusters. For KMeans algorithm changing the similarity comparing function might improve performance.

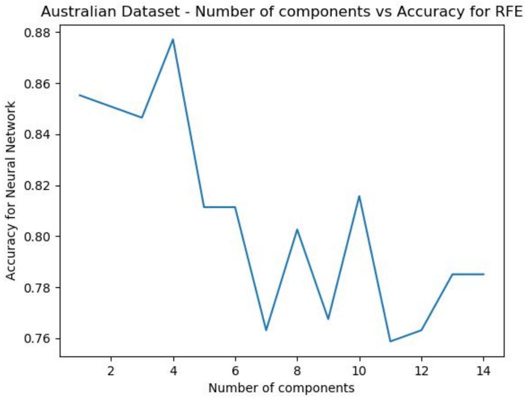
German data after dimensionality reduction looks somewhat similar to a normal distribution with some skewness for PCA, ICA and RFE. Australian data after dimensionality reduction is highly skewed and does not resemble normal distribution. Running clustering algorithms on data after applying PCA, ICA and RP give different clusters. This could be due to the fact that these different dimensionality reduction algorithms use different techniques to find the required number of components. Running RP multiple times produces different clustering results. Running RP and EM multiple times and then combining the results might reduce the problem of instability in terms of clustering.

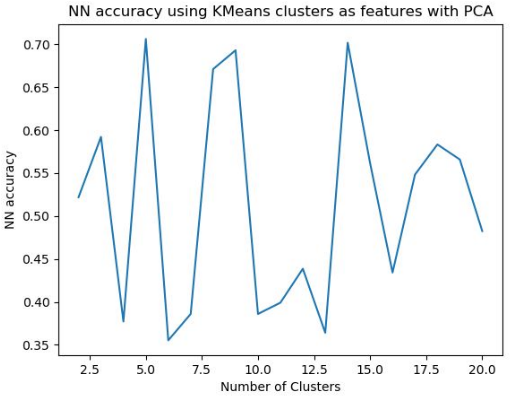
Applying dimensionality reduction algorithms and running Neural Network on the reduced data and then applying clustering algorithms on the reduced data and running Neural Network on the output of the clustering algorithms.

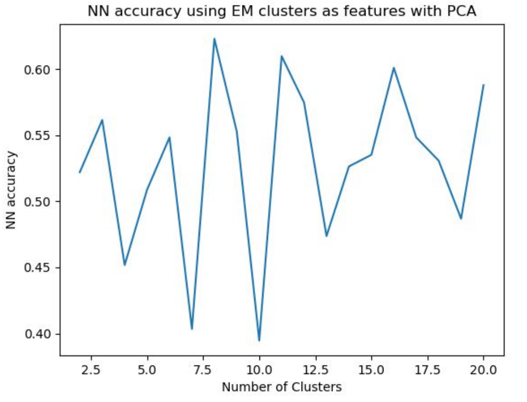

Australian Dataset - Number of components vs Accuracy for PCA
Accuracy for Neural Network
0.825
0.800
0.775
0.750
0.725
0.700
0.675
0.650
2
4
6
8
10
12
14
Number of components

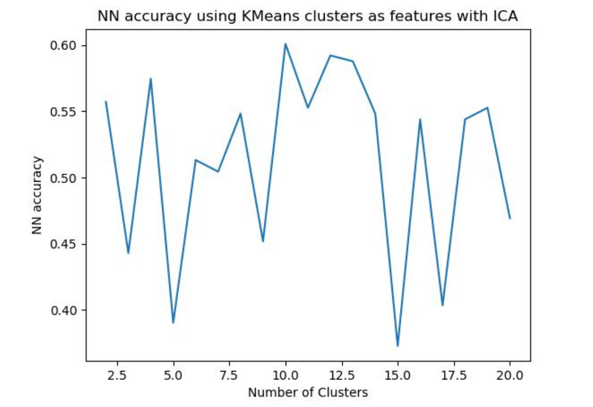

Australian Dataset - Number of components vs Accuracy for ICA
Accuracy for Neural Network
0.90
0.85
0.80
0.75
0.70
0.65
0.60
2
4
6
8
10
12
14
Number of components

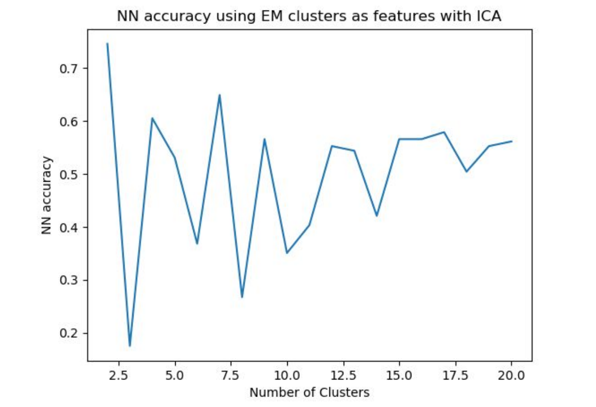

stralian Dataset - Number of components vs Accuracy for Randomized Projec
Accuracy for Neural Network
0.75
0.70
0.65
0.60
0.55
0.50
2
4
6
8
10
12
14
Number of components

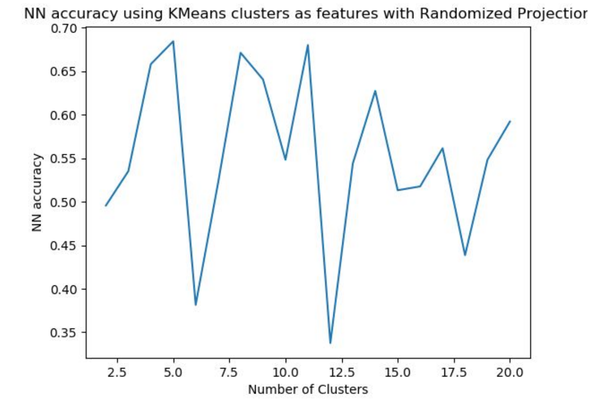

Australian Dataset - Number of components vs Accuracy for RFE
Accuracy for Neural Network
0.88
0.86
0.84
0.82
0.80
0.78
0.76
2
4
6
8
10
12
14
Number of components

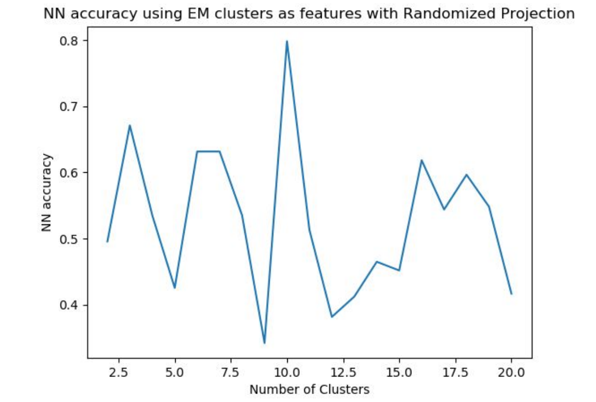

NN accuracy using KMeans clusters as features with PCA
NN accuracy
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Number of Clusters

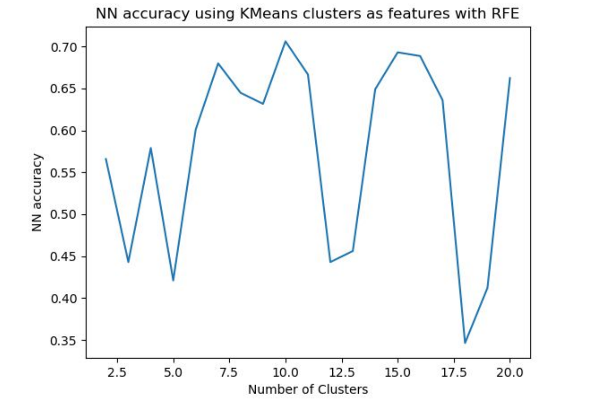

NN accuracy using EM clusters as features with PCA
NN accuracy
0.60
0.55
0.50
0.45
0.40
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Number of Clusters

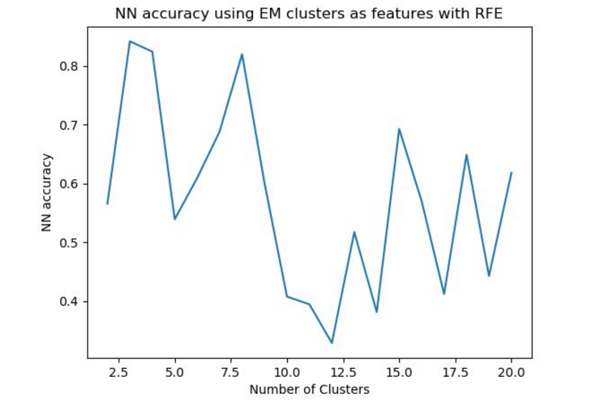

NN accuracy using KMeans clusters as features with ICA


NN accuracy using EM clusters as features with ICA


NN accuracy using KMeans clusters as features with Randomized Projectior


NN accuracy using EM clusters as features with Randomized Projection


NN accuracy using KMeans clusters as features with RFE


NN accuracy using EM clusters as features with RFE

First, I performed dimensionality reduction on Australian dataset while varying the number of features and plotted the accuracy of the neural network for the different values of the features for all the four dimensionality reduction algorithms. Then, I applied all four dimensionality reduction algorithms on Australian dataset with the number of features corresponding to the highest accuracy from the last task and then applied KMeans and EM on that reduced data. Then, the result of the KMeans and EM was used as input for the Neural Network and the number of clusters was varied and the accuracy of Neural Network was plotted against the number of clusters. Neural Network completed the training faster after dimensionality reduction. For ICA NN achieved approximately 90% accuracy