Assignment 4

A. General Guidelines about Models used in the Report:

For this report, neural network models are run on features created by 1) clustering algorithms, namely – k-Means Clustering and Expectation Maximization, 2) feature selection technique using Random Forest Feature Importance and feature transformation techniques such as Principal Component Analysis, Independent Component Analysis and Random Projections, and 3) k-Means Clustering features with features selected using Random Forest Feature Importance and Principal Components.

Grid Search is used to select the number of neurons required in the first hidden layer of the network which results in best 5-fold Cross Validation scores. Following are the key points to note before constructing the neural network:

- For learning algorithm (back propagation), Stochastic Gradient Descent is used in all experimental models.
- Sigmoid activation function is used as the output node.
- Binary Cross entropy is used as the loss function.

B. Bike Sharing Dataset:

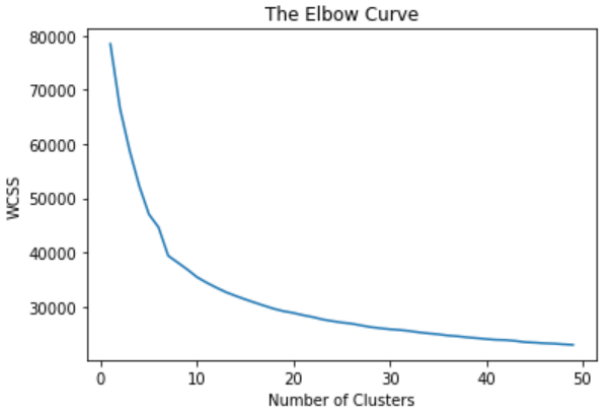
The Seoul Bike dataset contains information regarding number of bikes rented on an hourly basis for a year and includes prevailing weather conditions. The aim is to predict whether the bike rentals on a particular hour of a day are higher than the median value.

Following are the parameters of neural network learner for the Bike Sharing Dataset:

- **Weight Update:** batch_size = 10, epochs = 100
- **Hidden Layer 1:** activation = ReLU, dropout = 0.2, weight_constraint = maxnorm(3)
- **Output Layer:** neurons = 1, activation = sigmoid
- **SGD Optimizer:** learning_rate = 0.1, momentum = 0.1

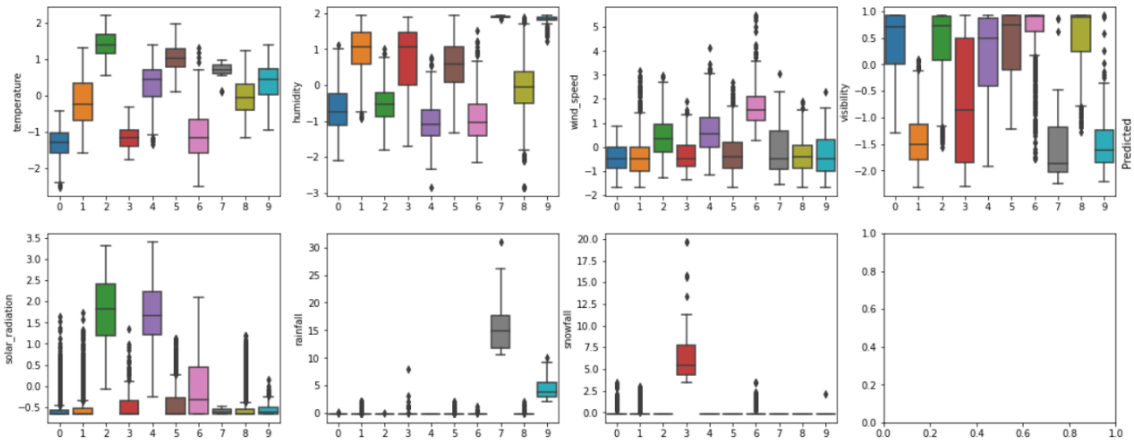
B1. Clustering Algorithms

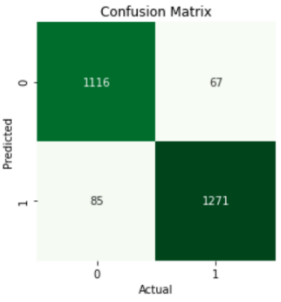
B1.1. k-Means Clustering



The elbow curve suggests that with added cluster, the within cluster sum of squares reduces. The drop in WCSS reduces gradually as the number of clusters increase. The optimum number of clusters chosen are 10. Below boxplots show the characteristic distribution of the data points with respect to the 10 classes. The boxplot distributions are mostly symmetrical across median values suggesting that the clusters are compact. Cluster no. 2 and 4 have higher temperatures, higher solar radiation, moderate wind speeds and lower humidity. These two clusters have higher number data points attributed to positive class which is corroborated by the table shown on the right. These clusters are added as dummy features (total of 49 features) that are fed to the neural network. Chosen parameters: # of Neurons = 25. The average CV accuracy is 0.91. The accuracy, sensitivity, and specificity of neural network learner on test set are very high.

class cluster	False	True
0	937	111
1	1203	561
2	8	757
3	163	3
4	79	785
5	411	1062
6	632	184
7	18	1
8	531	912
9	103	4



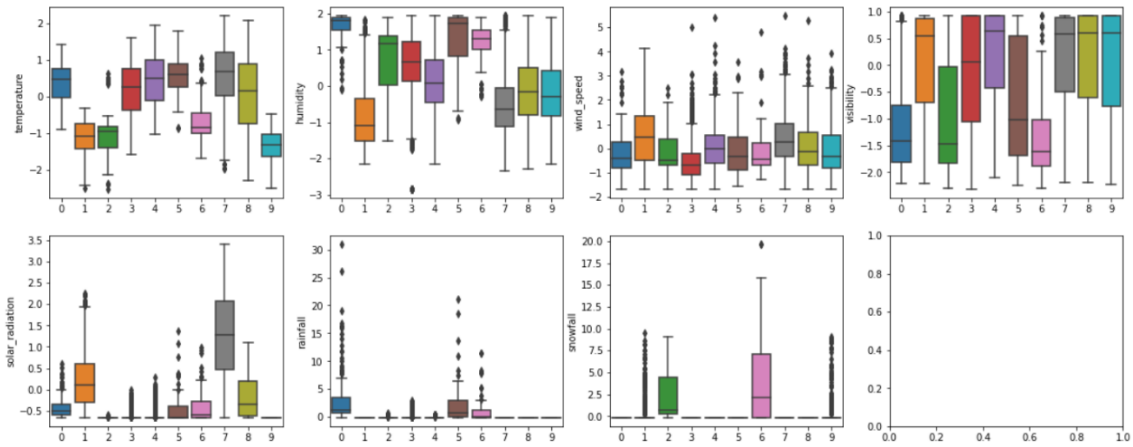


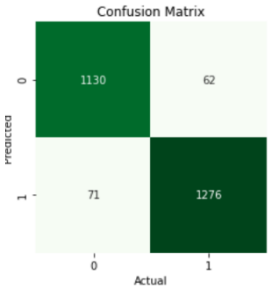
Accuracy: 0.94
Senstivity: 0.95
Specificity: 0.93

B2.2. Expectation Maximization

Maximum likelihood estimation is an approach to density estimation for a dataset by searching across probability distributions and their parameters. Maximum likelihood becomes intractable if there are variables that interact with those in the dataset but were hidden or not observed, so-called latent variables. The expectation-maximization algorithm is an approach for performing maximum likelihood estimation in the presence of latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. The optimum number of clusters chosen in the k-means clustering algorithm gives us a good idea of many clusters are needed in EM algorithm. Cluster no. 0, 1, 2, 5, 6 and 9 have significant distribution for rainfall and/or snowfall which validates unfavorable conditions for bike renting. Original features with dummy features (49) are fed to the neural network with same hyperparameters that were chosen for the k-means clustering. Chosen parameters: # of Neurons = 10. The average CV accuracy is 0.91. The performance on test set is similar to that of k-mean clustering.

class	False	True
cluster		
0	152	20
1	669	23
2	58	14
3	1407	649
4	112	1088
5	95	39
6	68	14
7	280	2012
8	113	513
9	1131	8





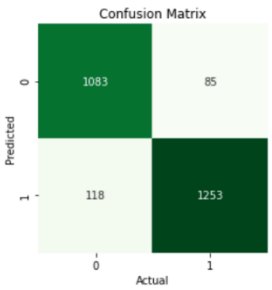
Accuracy: 0.95
Senstivity: 0.95
Specificity: 0.94

B3. Dimensionality Reduction Techniques

B3.1. Feature Importance using Random Forest

Feature selection using random forest is a filtering-based dimensionality reduction technique. The default method to compute variable importance is the mean decrease in impurity (or Gini importance) mechanism: At each split in each tree, the improvement in the split-criterion is the importance measure attributed to the splitting variable, and is accumulated over all

	importance		
feature			
temperature	0.238940	visibility	0.049075
hour	0.221591	wind_speed	0.045246
season	0.174046	day	0.042867
humidity	0.101825	rainfall	0.036526
solar_radiation	0.076247	snowfall	0.009651
		holiday	0.003986

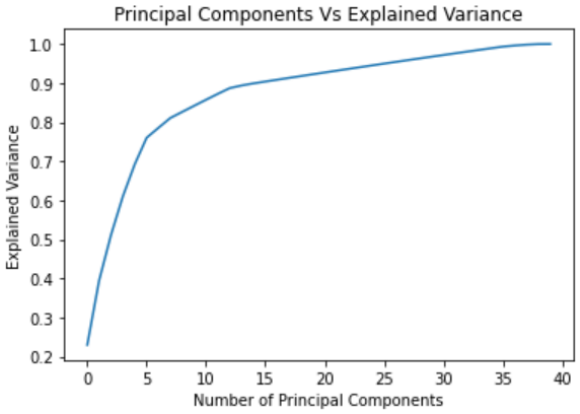


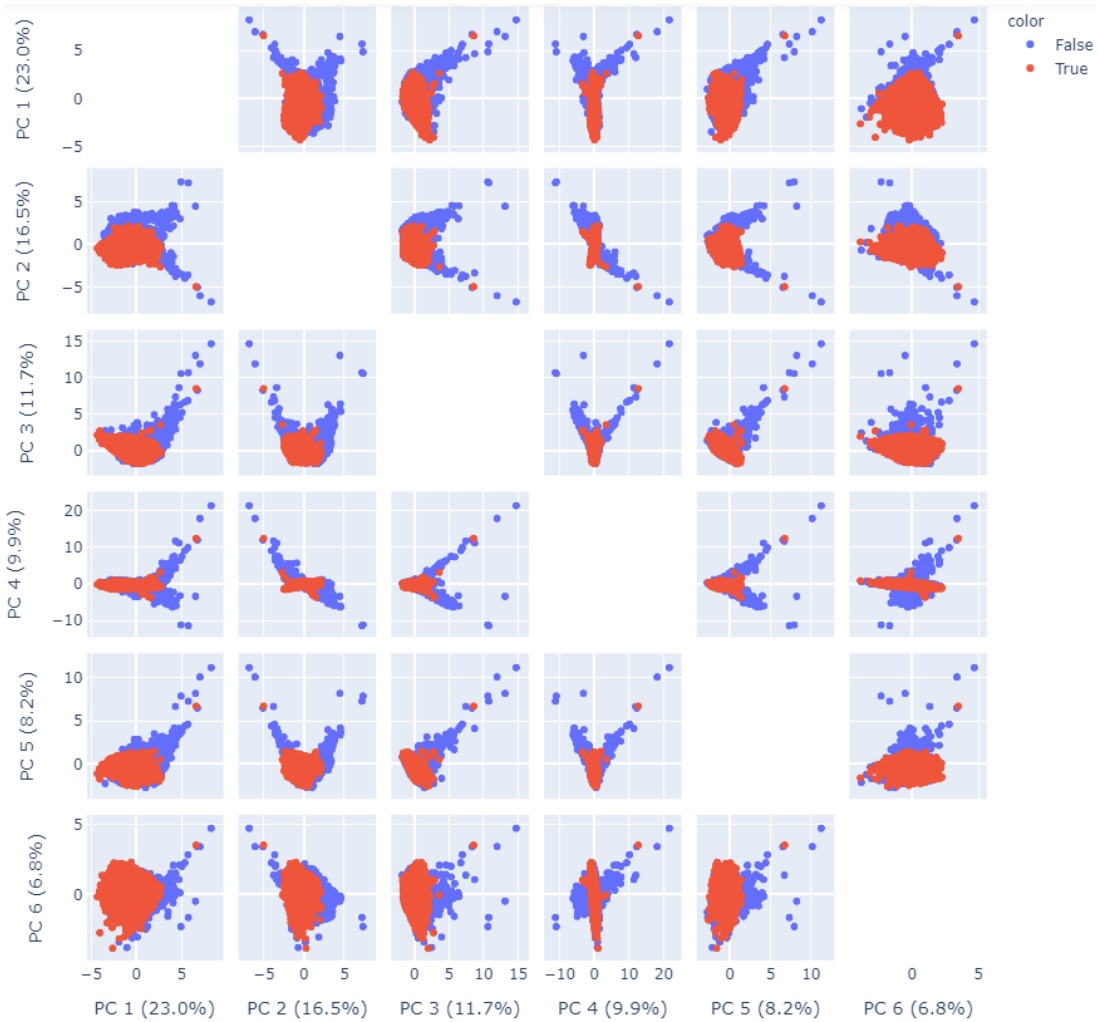
the trees in the forest separately for each variable. The table shows the decreasing order of feature importance. For the purpose of feature selection, the features that have importance over 5% are selected. These are – hour, temperature, season, humidity, solar radiation and visibility. There are a total of 29 features which are fed to neural network learner. The hyper-parameters are same as before. Chosen parameters: # of Neurons = 30. The average CV accuracy is 0.9. Since the neural network is tuned with only a subset of dataset the performance has significantly deteriorated. The specificity reduced by 4 points which dragged down the accuracy.

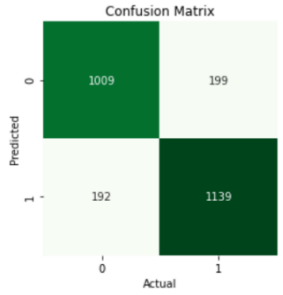
```
Accuracy: 0.92
Senstivity: 0.94
Specificity: 0.90
```

B3.2. Principal Component Analysis

Principal Component Analysis is a method to linearly transform the features from one space to another. PCA finds direction of maximum variance of the data. There are same number of principal components as the lower value among of columns (features) or rows (records). The components power of explaining variance reduces in a decreasing fashion. The graph to the right shows the cumulative sum of explained variance of the principal components. After 13 components, the explained variance graph becomes almost flat. Below grouped plots shows how the first 6 principal components separate the data points belonging to the binary classes.





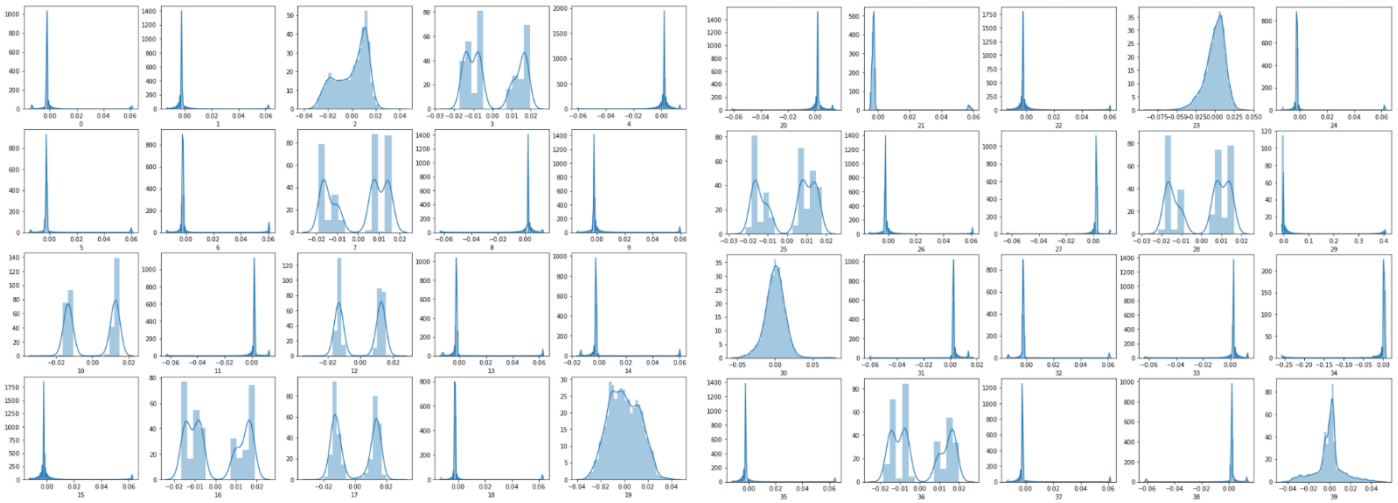



```
Accuracy: 0.85
Senstivity: 0.85
Specificity: 0.84
```

Chosen parameters: # of Neurons = 5. The average CV accuracy is 0.82. With the selected 13 features out of 40, the performance is not the best. However, by selecting all the features, accuracy, sensitivity, and specificity increases close to 95%.

B3.3. Independent Component Analysis

Unlike PCA, Independent Component Analysis tries to maximize independence. Linear transformation from our feature space to a new feature space such that each of the new features are mutually independent. In ICA, the mutual information between all the new features Y and the original features X is as high as possible. In ICA the features are not ordered (just statistically independent). Kurtosis is used to determine importance of features in ICA. When kurtosis is positive, the variable is said to be super-Gaussian or leptokurtic. Super-Gaussians are characterized by a spiky pdf with heavy tails, i.e., Laplace pdf. When kurtosis is negative, the variable is said to be sub-Gaussian or platykurtic. Sub-Gaussians are characterized by a rather flat pdf. Below charts show the distribution of all independent components.

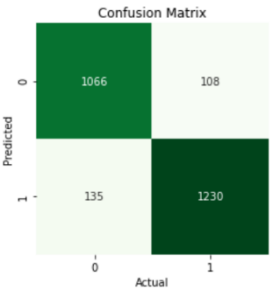


Only 27 out of 40 features fit the criterion having spiky pdfs. So, these features are used to find the local features. However, the dataset is linearly separable globally, the ICA does not perform well in classifying the labels.

B3.4 Random Projections

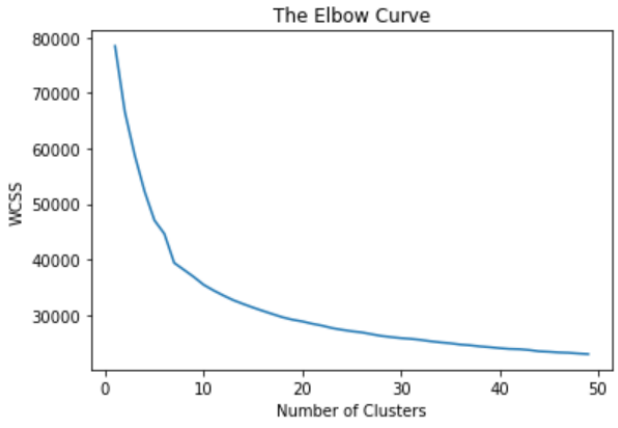
Random Component Analysis picks up random directions and projects the data on to these random directions. RCA runs faster than PCA and ICA. For this analysis, 13 random dimensions are chosen (same number as the principal components). Chosen parameters: # of Neurons = 30. The average CV accuracy is 0.85. The performance on test set is better than that of principal component analysis.

```
Accuracy: 0.90
Senstivity: 0.92
Specificity: 0.89
```

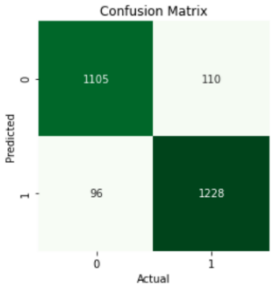


B.4. Cluster Analysis on Dimensionality Reduction Algorithms

B4.1. k-Means on Features Selected using Random Forest

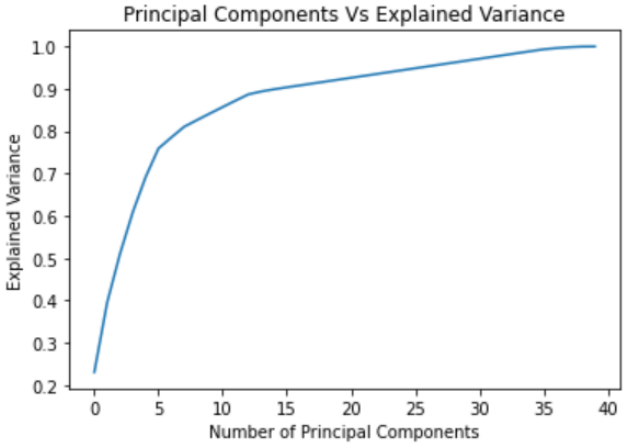


k-Means clustering is done on the dataset where the features are selected using feature importance. The suitable number of clusters are selected using the graphical approach of Elbow curve between number of clusters and WCSS. The number of clusters chosen is 10. The cluster membership feature is converted to dummy variables and added to the features selected using Random forest. The total number of features are 38. Chosen parameters: # of Neurons = 15. The average CV accuracy is 0.89. This model performed similar on the test set compared to the model with features selected using the random forest.

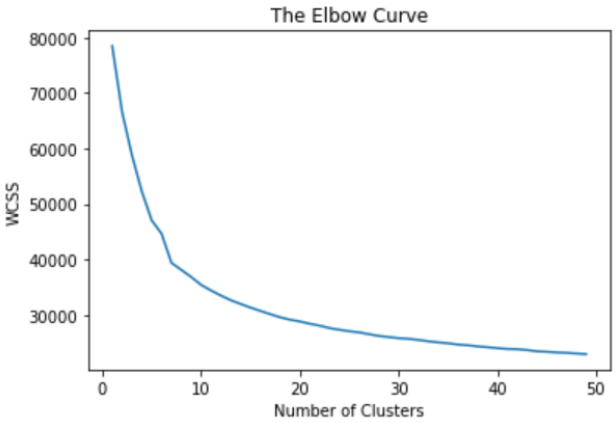



```
Accuracy: 0.92
Senstivity: 0.92
Specificity: 0.92
```

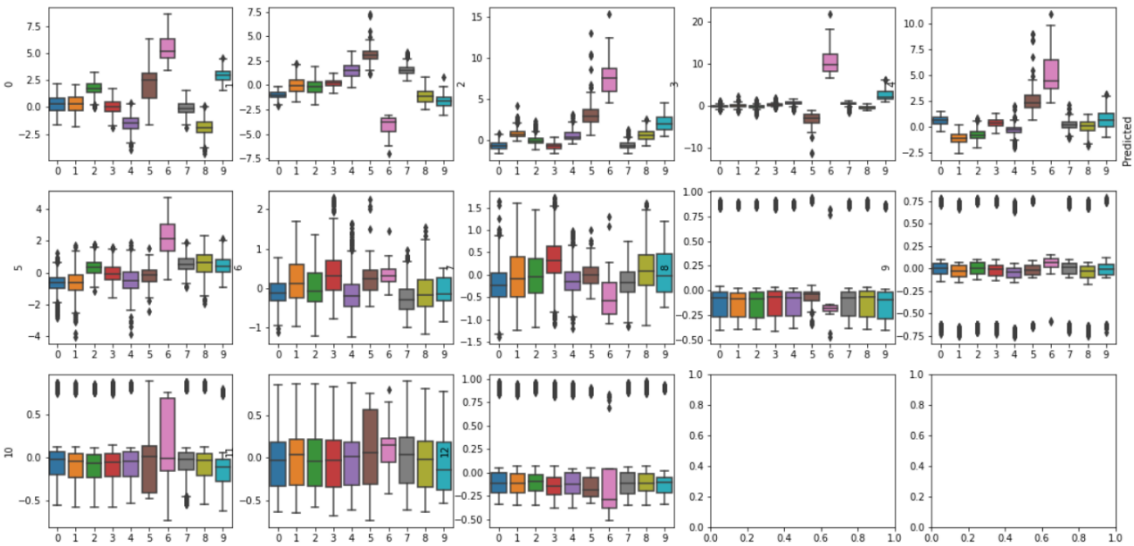
B4.2. k-Means on Features Transformed using PCA

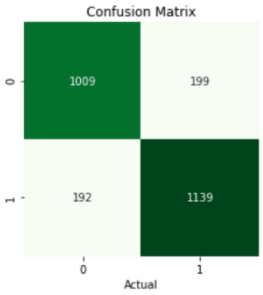


k-Means clustering is done on the dataset where the features are transformed using Principal Component Analysis. The first 13 PCA features are selected as the graph flattens after the optimal value. The suitable number of clusters are selected using the graphical approach of Elbow curve between number of clusters and within cluster sum of squares. The number of clusters chosen is 10.



The cluster membership feature is converted to dummy variables and added to the features transformed using PCA. Below boxplots suggest that the clusters are circular. The total number of features are 24. Chosen parameters: # of Neurons = 30. The average CV accuracy is 0.82. This model performed better on the test set compared to the model with features selected using the PCA feature transformation.





Accuracy: 0.87
Senstivity: 0.85
Specificity: 0.84

B5. Comparison of Models

Model	# of Neurons in 1st hidden layer	5-fold CV Accuracy	Test Accuracy	Test Sensitivity	Test Specificity
k-Means Clustering (10 clusters with 49 features)	25	0.91	0.94	0.95	0.93
Expectation Maximization (10 clusters with 49 features)	10	0.91	0.95	0.95	0.94
Feature Importance using Random Forest (29 features)	30	0.90	0.92	0.94	0.90
Principal Component Analysis (13 features)	5	0.82	0.85	0.85	0.84
Independent Component Analysis (27 features)	-	-	-	-	-
Random Projections (13 features)	30	0.85	0.90	0.92	0.89
k-Means on Features Selected using Random Forest (10 clusters with 38 features)	15	0.89	0.92	0.92	0.92
k-Means on Features Transformed using PCA (10 clusters with 24 features)	30	0.82	0.87	0.85	0.84

C. Chronic Heart Diseases Dataset:

C1. EDA and Data Pre-processing:

The chronic heart diseases dataset contains information regarding the chance of contracting a chronic heart disease after 10 years based on following parameters:

- Demographic: male, age, education
- Behavioral: current_smoker, cigs_per_day
- Medical (history): bp_meds, prevalent_stroke, prevalent_hyper, diabetes
- Medical (current): tot_chol, sys_bp, dia_bp, bmi, heart_rate, glucose

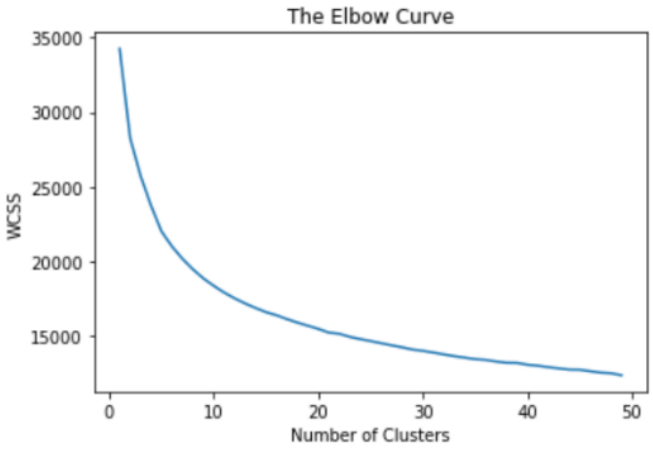
Following are the parameters of neural network learner for the Chronic Heart Diseases Dataset:

- **Weight Update:** batch_size = 20, epochs = 50
- **Hidden Layer 1:** activation = Linear, dropout = 0.1, weight_constraint = maxnorm(3)
- **Output Layer:** neurons = 1, activation = sigmoid
- **SGD Optimizer:** learning_rate = 0.001, momentum = 0.0

Since the dataset is not balanced, Mathew Correlation Coefficient (MCC) is used instead of Accuracy as a metric to choose the best neural network.

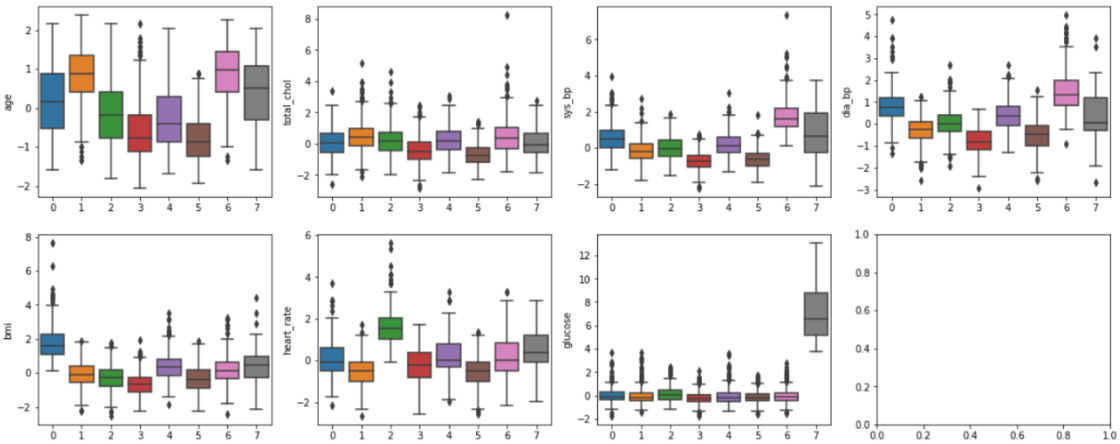
C2. Clustering Algorithms

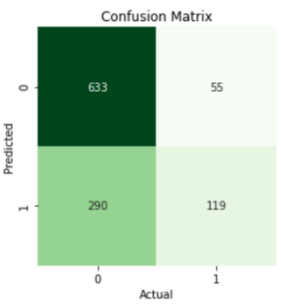
C2.1. k-Means Clustering



The elbow curve suggests that the optimum number of clusters is around 8. However, there is no clear breakpoint where the line becomes flat. This results in clusters which do not clearly attribute to individual classes as shown on the table to the right. The boxplots further suggest that the clusters are compact but do not clearly attribute to target classes. These clusters are added as dummy features (total of 24 features) that are fed to the neural network. Grid Search Cross Validation is used to choose the correct number of neurons in the 1st hidden layer. Chosen parameters: # of Neurons = 5. The average CV MCC is 0.25. The overall classification performance on the test set is decent.

ten_year_chd	0	1
cluster		
0	257	56
1	618	130
2	324	28
3	575	72
4	376	85
5	641	27
6	290	139
7	20	20



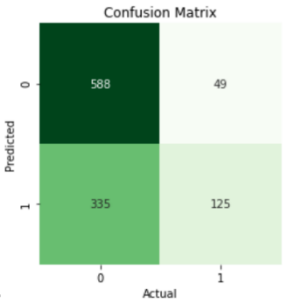


Accuracy: 0.69
Sensitivity: 0.68
Specificity: 0.69

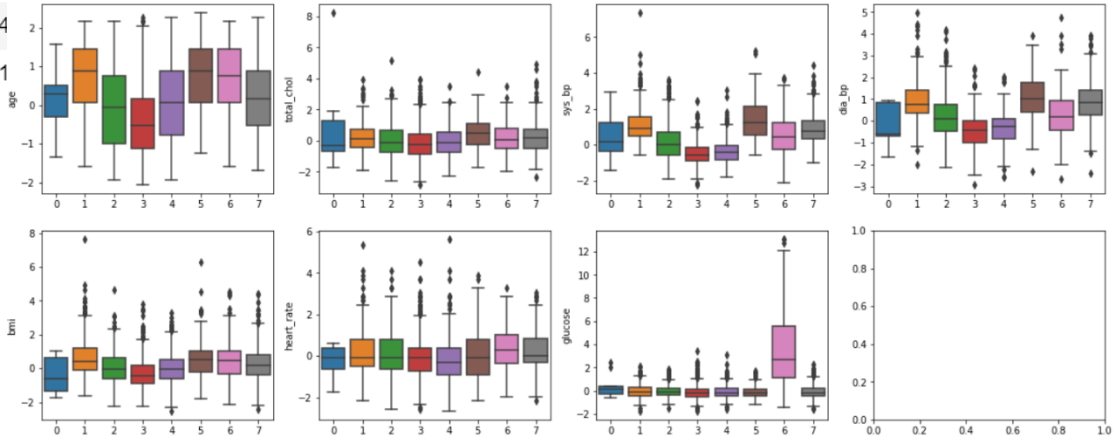
C2.2. Expectation Maximization

ten_year_chd	0	1
cluster		
0	5	4
1	242	83
2	503	63
3	1141	160
4	761	76
5	75	36
6	64	34
7	310	101

The number of clusters chosen by inspecting Elbow curve in k-Means Clustering gives a good starting point in selecting number of clusters for Expectation Maximization model. The cluster membership and shape of clusters are almost similar to the k-means clustering. The network hyper-parameters are same as that of k-Means Clustering. Chosen parameters: # of Neurons = 25. The average CV MCC is 0.26. The model performance is worse than the previous one.



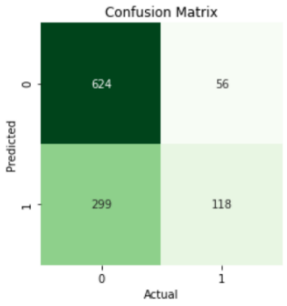

```
Accuracy: 0.65
Sensitivity: 0.72
Specificity: 0.64
```



C3. Dimensionality Reduction Techniques

C3.1. Feature Importance using Random Forest

The table on the right shows the decreasing order of feature importance. For the purpose of feature selection, the features that have importance over 5% are selected. These are – sys_bp, total_chol, bmi, age, dia_bp, glucose, hear t_rate, cigs_per_day. There are a total of 8 features which are fed to neural network learner. Chosen parameters: # of Neurons = 15. The average CV MCC is 0.23.

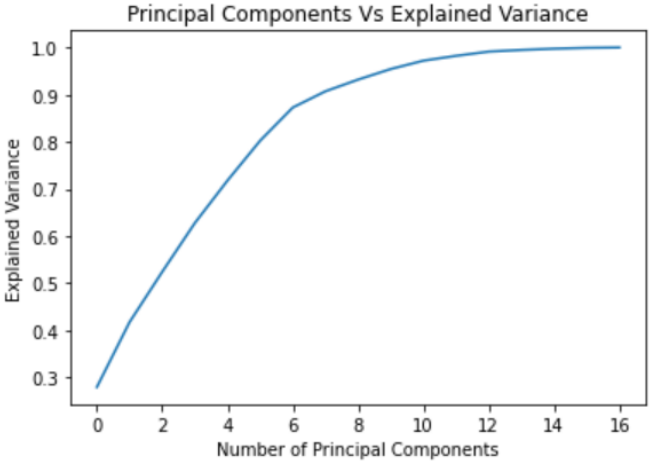


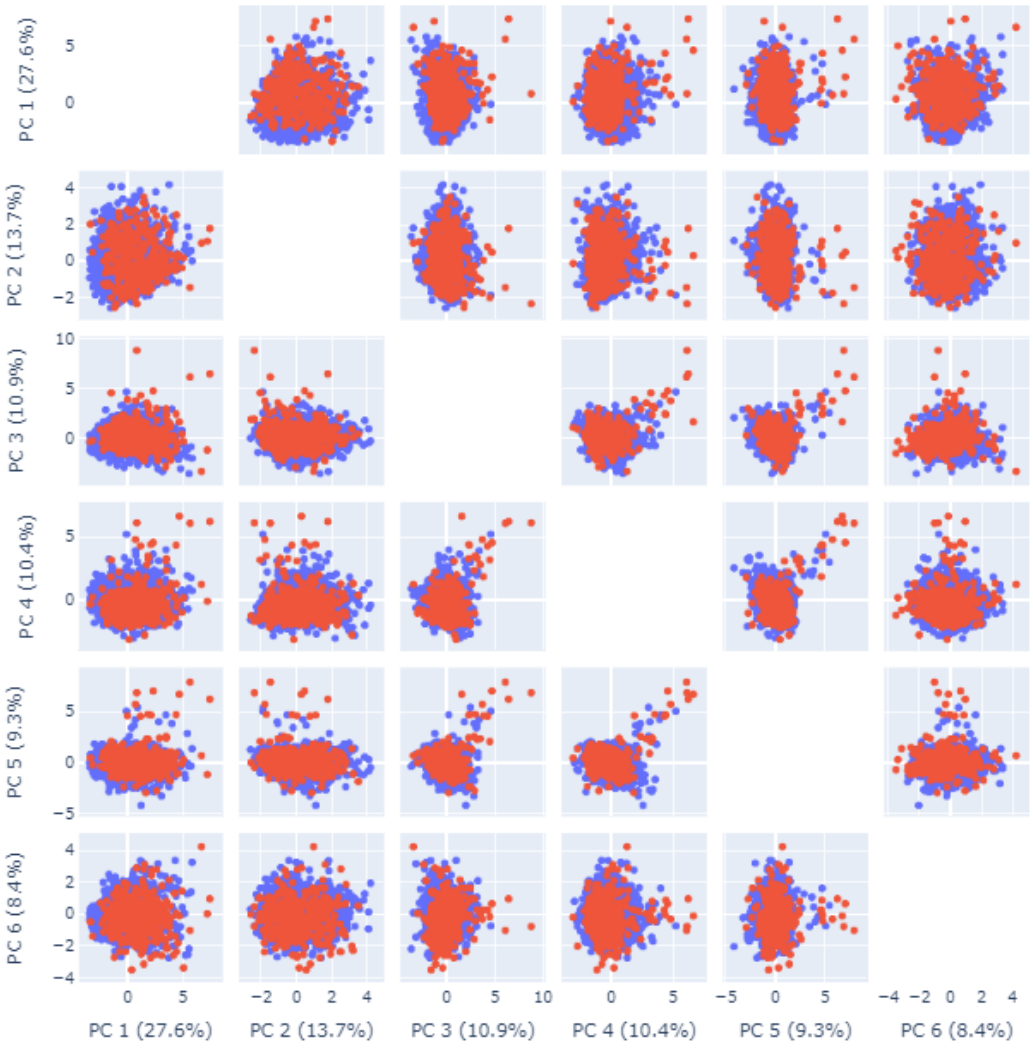

```
Accuracy: 0.68
Sensitivity: 0.68
Specificity: 0.68
```

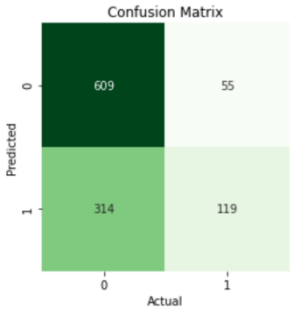
feature	importance
sys_bp	0.135932
age	0.124991
bmi	0.124494
total_chol	0.124226
dia_bp	0.117772
glucose	0.116643
heart_rate	0.093633
cigs_per_day	0.050690
edu	0.041783
male	0.021274
prevalent_hyper	0.016891
current_smoker	0.013145
bp_meds	0.007396
diabetes	0.007014
prevalent_stroke	0.004116

C3.2. Principal Component Analysis

The power of explaining variance for principal components reduce in a decreasing fashion. The plot on the right shows the increase in explained variance as number of principal components increase. The graph becomes almost flat after 10 principal components. These features are chosen as inputs to the learner. Chosen parameters: # of Neurons = 10. The average CV MCC is 0.24. The performance on test set did not deteriorate by dropping the last 7 principal components.



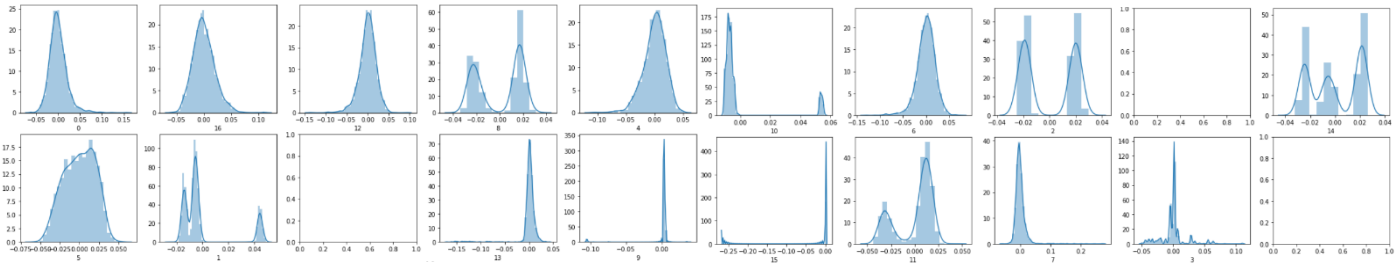




Accuracy: 0.66
Sensitivity: 0.68
Specificity: 0.66

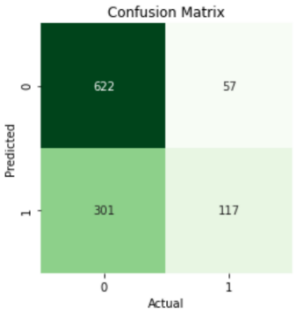
C3.3. Independent Component Analysis

6 out of 17 features fit the criterion having spiky pdfs and have positive kurtosis. So, these features are used to find the local features. However, the neural network learner is not trained on these components.



C3.4. Random Projections

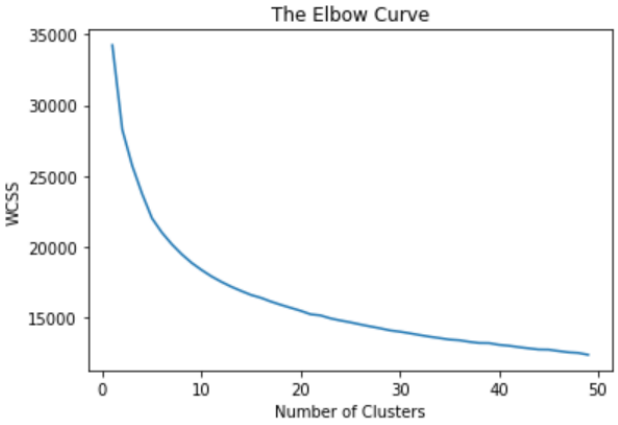
Random Component Analysis (RCA) picks up random directions and projects the data on to these random directions. For this analysis, 10 random dimensions are selected. Chosen parameters: # of Neurons = 10. The average CV MCC is 0.24. Sensitivity is the highest compared to all other models, but this has significantly affected the specificity. This is due to the projections in random directions could not model in a better way compared to others.



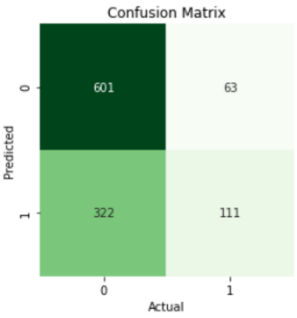
Accuracy: 0.67
Sensitivity: 0.67
Specificity: 0.67

C4. Cluster Analysis on Dimensionality Reduction Algorithms

C4.1. k-Means on Features Selected using Random Forest

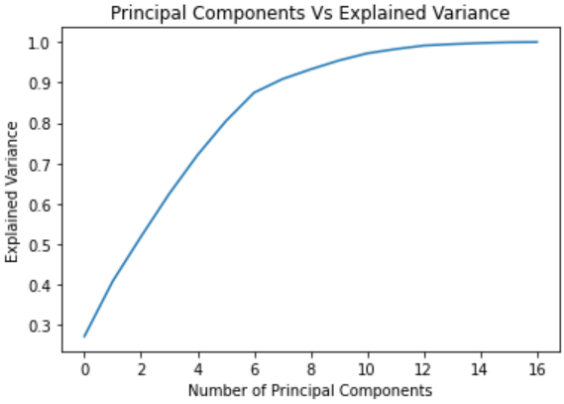


k-Means clustering is done on the dataset where the features are selected using feature importance attribute of random forest algorithm. The suitable number of clusters are selected using the graphical approach of Elbow curve between number of clusters and WCSS. The number of clusters chosen is 8. The cluster membership feature is converted to dummy variables and added to the features selected using Random forest. The cluster membership remains similar to the first k-mean model. The total number of features are 15. Chosen parameters: # of Neurons = 20. The average CV MCC is 0.24. This model performed worse on the test set compared to the model with features selected using the random forest.

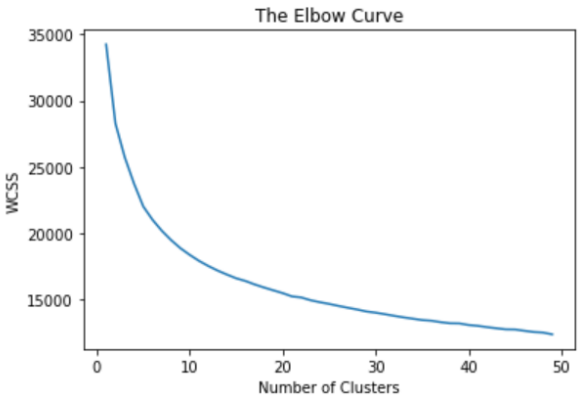


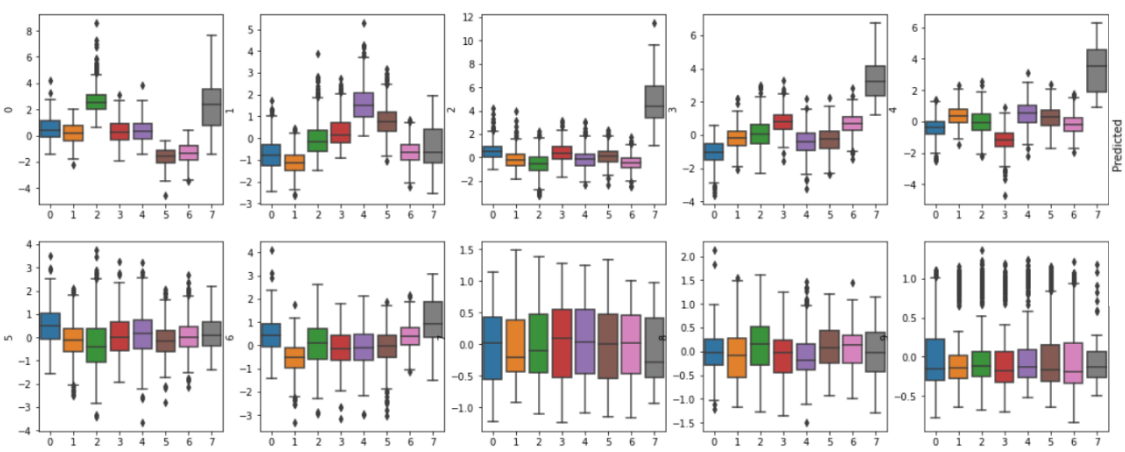

```
Accuracy: 0.65
Sensitivity: 0.64
Specificity: 0.65
```

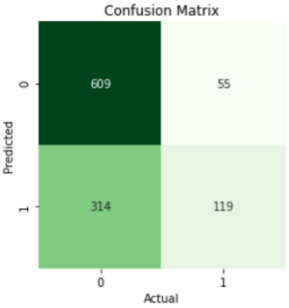
C4.2. k-Means on Features Transformed using PCA



k-Means clustering is done on the dataset where the features are transformed using Principal Component Analysis. The first 10 PCA features are selected as the graph flattens after the optimal value. The suitable number of clusters are selected using the graphical approach of Elbow curve between number of clusters and WCSS. The number of clusters chosen is 8. But none of these clusters do not clearly attribute to the class with lower occurrences. The cluster membership feature is converted to dummy variables and added to the features transformed using PCA. There are a total of 17 input variables to the neural network learner. Chosen parameters: # of Neurons = 1. The average CV MCC is 0.24. There is no improvement in the performance.








```
Accuracy: 0.69
Sensitivity: 0.68
Specificity: 0.66
```

C5. Comparison of Models

Model	# of Neurons in 1st hidden layer	5-fold CV Accuracy	Test Accuracy	Test Sensitivity	Test Specificity
k-Means Clustering (8 clusters with 24 features)	5	0.25	0.69	0.68	0.69
Expectation Maximization (8 clusters with 24 features)	25	0.26	0.65	0.72	0.64
Feature Importance using Random Forest (8 features)	15	0.23	0.68	0.68	0.68
Principal Component Analysis (10 features)	10	0.24	0.66	0.68	0.66
Independent Component Analysis (17 features)	-	-	-	-	-
Random Projections (10 features)	10	0.24	0.67	0.67	0.67
k-Means on Features Selected using Random Forest (8 clusters with 15 features)	20	0.24	0.65	0.64	0.65
k-Means on Features Transformed using PCA (10 clusters with 17 features)	1	0.24	0.69	0.68	0.66

D. Conclusion:

The bike sharing dataset is linearly separable, hence most of the models with original features perform similarly. The transformations on features do not make much of difference. The feature transformation done for the purpose of dimensionality reduction has led to poor performance. However, selecting all the features in PCA will have same results as that of models with original features. Expectation maximization model performs the best in terms of overall metrics. The clusters created with and without feature transformation resulted in similar sized clusters which are round and compact.

The heart disease dataset is not linearly separable and has unbalanced classes. Models with features transformed to other dimensions have not performed worse as compared to the similar models run on the bike sharing dataset. k-Means clustering resulted in superior performance with a neural network on test set. The clusters created with and without feature transformation resulted in similar sized clusters which are round and compact. However, none of the clusters do not clearly attribute to the class with lower occurrences.