Prepared by: Sai Lakshmi Nikitha Akarapu

ID: SXA210112 Date: 07/08/2022

BUAN 6341 Applied Machine Learning

ASSIGNMENT NO 4

Seoul Rental Bike prediction Part IV

Executive Summary

- **Sensible feature selection and preparation improve prediction accuracy.**
- **Converted the data set to binary classification using a median of the output parameter.**
- **Experimented with the data set by implementing K-means clustering and Expectation Maximization by using the elbow method to define a number of clusters.**
- **Optimizing parameters by implementing a decision tree, PCA, ICA, and Randomized projections for improvising where Dt has reduced feature set of 7, PCA has 8, ICA has made features mutually independent when observed data distributions and RCA has made the noise fewer distributions by multiplying with the random matrix.**
- **Performed K-Means and EM on the Reduced feature set and observed clustering is based on Temperature, visibility, and Humidity, so these are the main features.**
- **Applied clustering algorithms and implemented neural networks on the dimensionally reduced dataset to see the performance difference**
- **Overall, from this experiment, we can either consider DT, RP reduced feature set, or cluster results and apply ANN are good in terms of performance and time taken.**

Introduction

In this project, the objective is to use the dataset that is given as part of the previous assignment to convert into a classification problem statement, then apply clustering algorithms such as K-Means and EM, then perform dimensionality reduction using Decision Trees, PCA, ICA, and RCA, then performing clustering on dimensionality reduction features, then applying neural network on dimensionality reduction feature set and finally using cluster outputs and apply neural network for classification.

About the Data

The dataset consists of 14 features and 8760 records. The data is the rented bike count captured at each hour and weather conditions (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), along with date information. Dew point Temperature is removed as it is highly correlated with the Temperature. The date column is removed as it doesn't help us more in clustering. Seasons are hot encoded and one of the season variables is dropped to avoid multi-collinearity.

Project Outline

Part 1: Implement a clustering algorithm like K-means clustering and Expectation Maximization on the dataset using the optimized number of clusters.

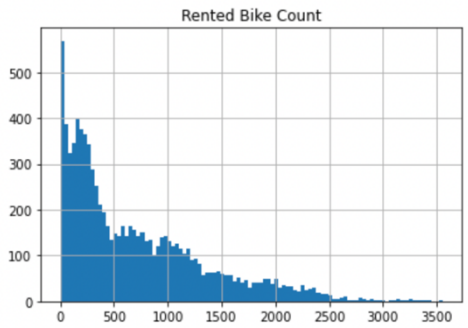
Part 2: Implement feature dimensionality reduction algorithms on the dataset using Decision Tree, Principal Component Analysis, Independent Component Analysis, and Randomized projections.

Part 3: Experimented by applying k-means clustering on data after dimensionality reduction.

Part 4: Experimented by applying neural networks to data after dimensionality reduction.

Part 5: Using the clustering results from task 1 as the new features and applying neural network learner on this new data.

Part 1: Clustering algorithms

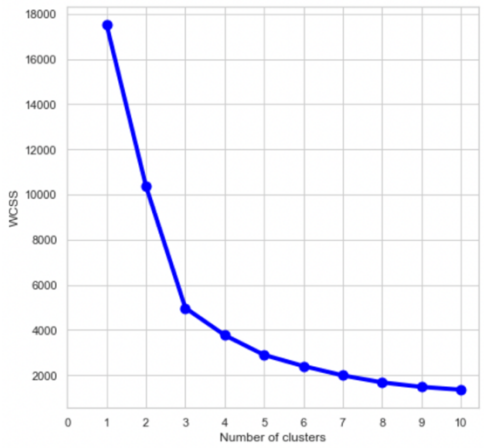


From the histogram, we can observe that the Rented Bike Count data is right-skewed. The median is the best technique to separate the data set into two distributions. The median of the output feature is 504.5. The rented bike count is greater than the median value, considered class 1, and less than or equal to the value is class 0.

K-means Clustering:

We will use sklearn.cluster.KMeans from a scikit-learn package for this experimentation.

1. In general, we can evaluate how well the models are performing based on different K clusters since clusters are used in the modeling. The 'Elbow method' gives us some intuition about k so experimented with the elbow method on the dataset to select the optimal number of clusters for clustering.



From the experimentation, we can observe that curve is forming a bend or elbow at the number of clusters is 3 and WCSS starts to flatten out. Hence, we can conclude that the optimal number of clusters for clustering algorithms is 3.

2. Experimented with the k-mean clustering on two columns named Visibility and Temperature to separate the dataset into different partitions.

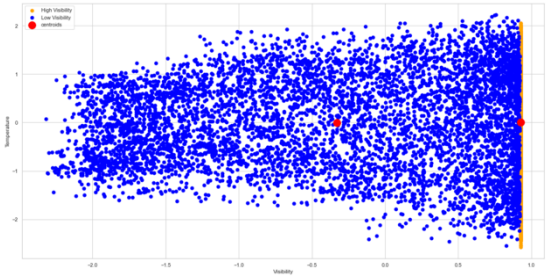
We used a number of clusters as 3 as it will be the ideal one to take and experiment with the dataset. From the colored scatter plot, we can clearly see the distinguishable clusters with centroids 1) Cluster 0(orange): Medium Temperature – Low Visibility,
2) Cluster 1(green): High Temperature – Low visibility, 3) Cluster 2(blue): Low Temperature – High Visibility.

It's observed that when the temperature is medium rental bike count is medium, when the temperature is high rental bike count is high when the temperature is low rental bike count is low irrespective of visibility.

Expectation Maximization:

1. Experimented with the model by application of the Expectation Maximization to a Gaussian Mixture. The main assumption of this model is that there are a certain number of Gaussian distributions, and each of these distributions represents a cluster. Gaussian mixture models are an approach to density estimation where the parameters of the distributions are fit.

   For this experiment, we will use sklearn.mixture.GaussianMixture from the scikit-learn package. We have experimented with hyperparameters like 'cvtype' and 'n_components'. The 'full' cvtype and '2' n_components would be ideal for expectation maximization.



From the visualization, we can observe the 2 clusters distinguish as Cluster 0(orange): High Visibility, and Cluster 1(blue): Low Visibility. Hence, a Gaussian Mixture model tries to group the observations belonging to a single distribution together.

It's observed that the EM technique used only visibility to perform clustering and can be seen that when High visibility rental bike count is more than the median value and in low visibility rental bike count is less than the median value.

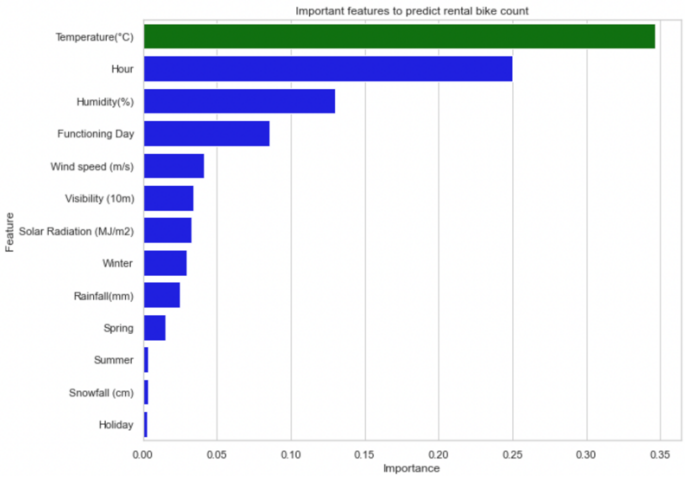
Conclusion:

1) K-Means uses temperature and visibility to perform clustering
2) EM uses only visibility to perform clustering.

Task 2: Dimension Reduction Techniques

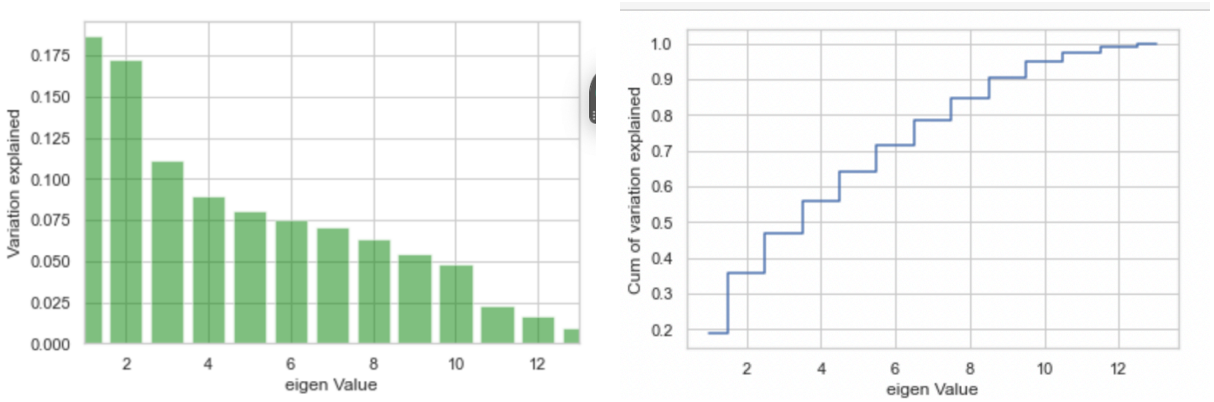
1. Experimented with **Decision Tree Classifier** to find the importance of the variables. The data was split into 70% train and 30% test for training and testing the model. We will use sklearn.tree.DecisionTreeClassifier.
   The higher the value, the more important the feature. The importance of a feature is computed as the total reduction of the criterion brought by that feature which is known as the Gini importance.

	Feature	Importance
0	Temperature(°C)	0.346672
1	Hour	0.250331
2	Humidity(%)	0.129736
3	Functioning Day	0.085628
4	Wind speed (m/s)	0.041563
5	Visibility (10m)	0.033861
6	Solar Radiation (MJ/m2)	0.032734
7	Winter	0.029776
8	Rainfall(mm)	0.025069
9	Spring	0.015089
10	Summer	0.003542
11	Snowfall (cm)	0.003461
12	Holiday	0.002538



From the above feature importance table, we can identify that 'Temperature(°C)' is an important factor in deciding the count of rental bikes. Hour, Humidity(%), and Functioning Day also have an influence on the count. On the contrary, Holiday, Snowfall(cm), and summer seem to have less importance for the rental bike counts.

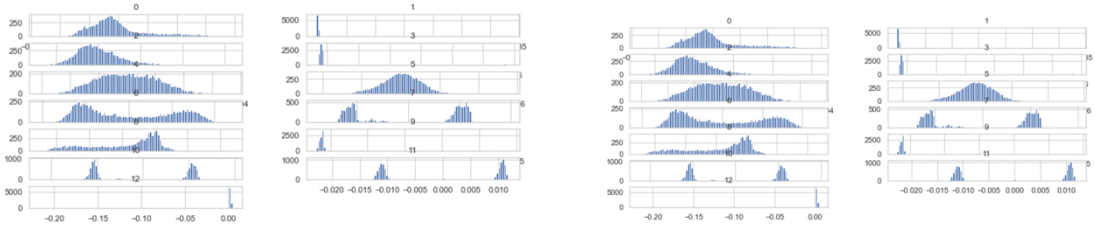
2. Experimented with the dataset to reduce dimensions using **Principal Component Analysis**. PCA is used to reduce the number of variables/features, avoid multicollinearity, and solve the issue of eigenvectors/values before applying neural networks. PCA assumes that the data is Gaussian distributed.



From this technique, we can observe that Principal Component 1 (PC1) contributed 18.7%, PC2 contributed 17.22%, PC3 contributed 11.07%, and so on. The further we go, the lesser the contribution to the total variance. If we look at the cumulative values of variance, we need 8 principal components that seem very reasonable and can explain over 80%-85% of the variation in the original data.

3. Experimented with the dataset to reduce dimensions using **Independent Component Analysis**. ICA perseveres the most 'independent' dimensions of the data which are different from the dimensions with the most variance by assuming those components have a non-Gaussian distribution. We will use sklearn.decomposition.FastICA from a scikit-learn package.

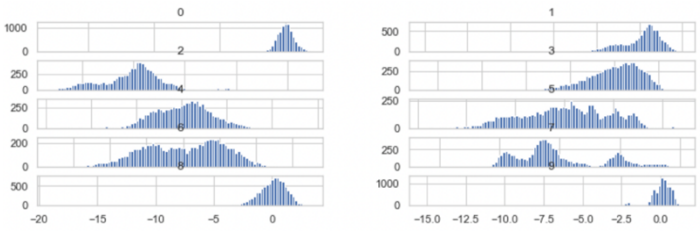
   ICA is different from a standard PCA because it looks for statistically independent and uncorrelated components.



Experimented with ICA using different algorithms, we see there is no change in the distribution of data which might have no difference in finding local features, hence either one can be used and observe that features are more mutually independent.

4. Experimented with the dataset using Randomized Projection for dimensionality reduction of high-dimensional data sets where the original high-dimensional data is projected onto a lower-dimensional subspace using a random matrix whose columns have unit lengths.

We will use sklearn.random_projection.GaussianRandomProjection from a scikit-learn package. This reduces dimensionality through Gaussian random projection, which has been found to be computationally efficient and accurate.



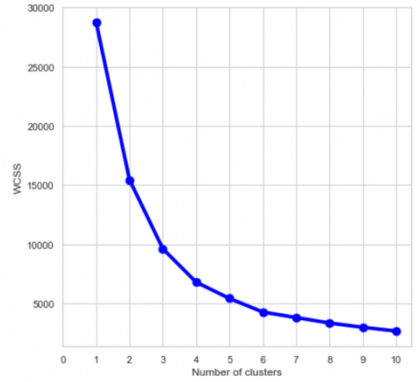
We will be taking 10 features as a random projection subset and might need to experiment with a number of components during model training. We must adjust the number of features to achieve good performance for now we are taking 10.

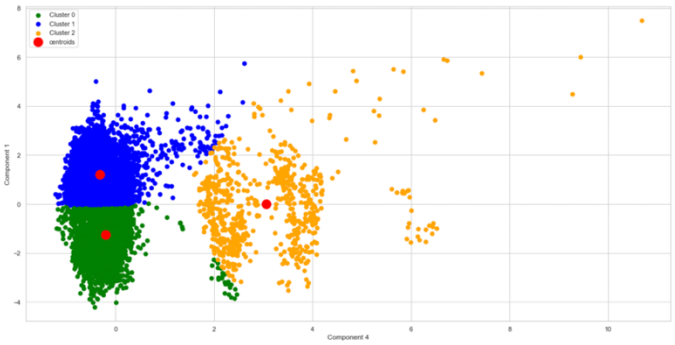
Conclusion:

1) Decision Tree has reduced feature set to 7 which are Temperature, Hour, Humidity(%), Functioning Day, Wind Speed, Visibility, and Solar Radiation.
2) PCA has reduced the feature set to 8 with 85% of variation explained using these 8 features.
3) ICA is performed but no difference in distribution for algorithm 'parallel' and 'deflation' but need to perform feature selection to input to the model.
4) RP is performed and baselined components to 11 after Task 4.

Task 3: Clustering on dimensionality reduced techniques

1. Experimenting with k-mean clustering on Principal Component Analysis by using two components



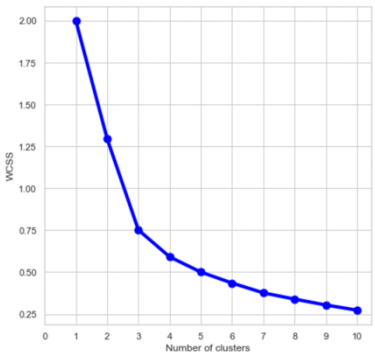


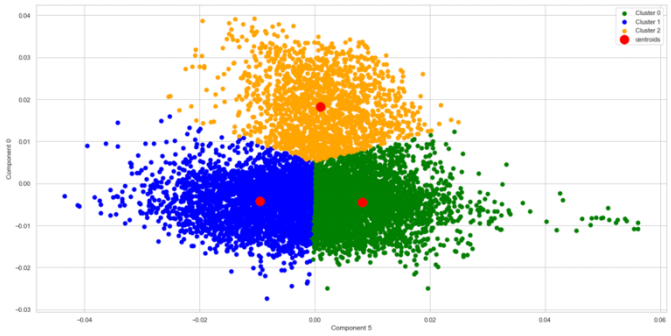
From the elbow method output, we can infer that the bend started at the number of clusters of value 3. The optimal number of clusters to be used in the technique is 3.

Cluster	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Holiday	Functioning Day	Spring	Summer
0	0.302833	14.797443	14.015693	43.544748	2.264026	1734.869140	1.041123	0.002156	0.033793	0.007270	1.000000	0.250940	0.263224
1	0.639882	8.234934	11.762126	72.395394	1.190642	1138.613180	0.118317	0.184174	0.123567	0.004410	1.000000	0.270211	0.267516
2	0.812772	11.753266	12.964877	59.280116	1.768940	1478.058055	0.507112	0.786792	0.026705	0.558781	0.571843	0.150943	0.095791

From the statistics of the features, we can infer that Cluster 0(Green): Low Humidity, Cluster 1(Blue): High Humidity, and Cluster 3(Orange): Medium Humidity.

2. Experimenting with k-mean clustering on Independent Component Analysis by using two components



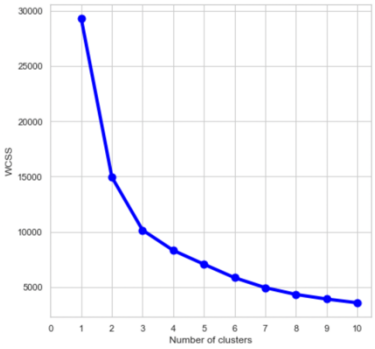


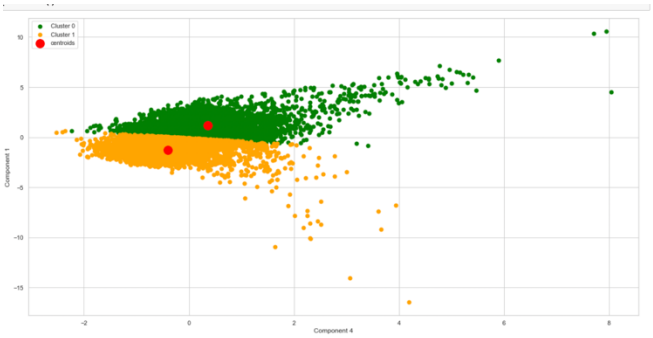
From the elbow method output, we can infer that the bend started at the number of clusters of value 3, which is the ideal number.

Cluster	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Holiday	Functioning Day	Spring	Summer
0	0.542240	10.149798	11.018003	51.155196	1.748907	1458.211606	0.261671	0.128340	0.077625	0.046964	0.967341	0.235897	0.240486
1	0.607164	12.496744	11.668561	73.617821	1.509562	1338.251924	0.143961	0.237655	0.096388	0.050918	0.967140	0.248076	0.223801
2	0.190817	12.475253	19.449195	42.844961	2.105665	1588.136553	2.104723	0.014431	0.026476	0.051282	0.962433	0.295766	0.334526

From the feature statistics, we can infer that Cluster 0(green): Low Temperature, Cluster 1(blue): Low Temperature, and Cluster 2(orange): High Temperature.

3. Experimenting with k-mean clustering on Random Projections Analysis by using two components.



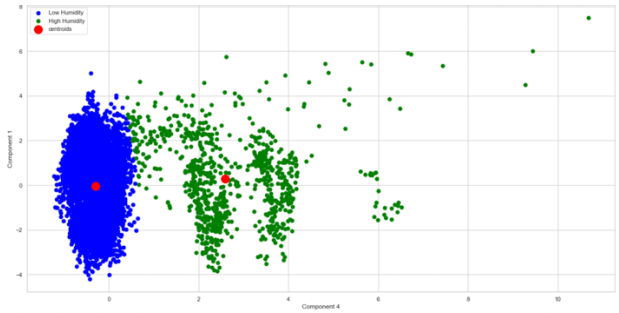


From the elbow method output, we can infer that the bend started at the number of clusters of value 2, which is the ideal number.

Cluster	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Holiday	Functioning Day	Spring	Summer
0	0.749022	8.175359	6.162299	59.114298	1.462538	1419.942634	0.202112	0.023381	0.142894	0.046502	0.971969	0.128857	0.138635
1	0.224387	15.179654	20.321188	57.243386	2.015296	1455.511785	0.975298	0.287374	0.000000	0.052429	0.960077	0.388408	0.377585

From the feature statistics, we can infer that Cluster 0(green): High Humidity, Cluster 1(orange): Low Humidity

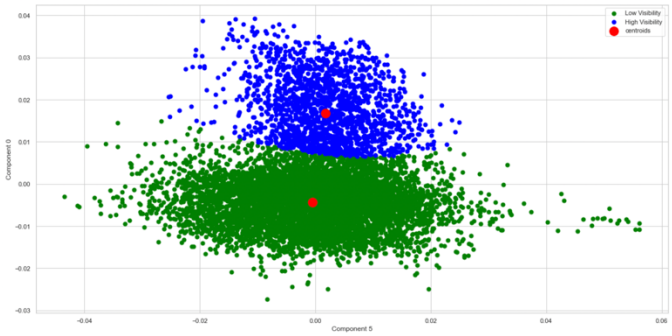
4. Experimenting with Expectation Maximization clustering on Principal Component Analysis by using two components



Cluster_EM	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Holiday	Functioning Day	Spring	Summ
0	0.463906	11.470238	12.831598	57.629559	1.718503	1447.723151	0.577865	0.034904	0.077786	0.0	1.000000	0.259119	0.2643
1	0.829861	11.771991	13.351968	63.679398	1.783449	1337.236111	0.489109	1.188542	0.050231	0.5	0.658565	0.187500	0.1400

From the feature statistics, we can infer that Cluster 0(blue): Low Humidity, Cluster 1(green): High Humidity.

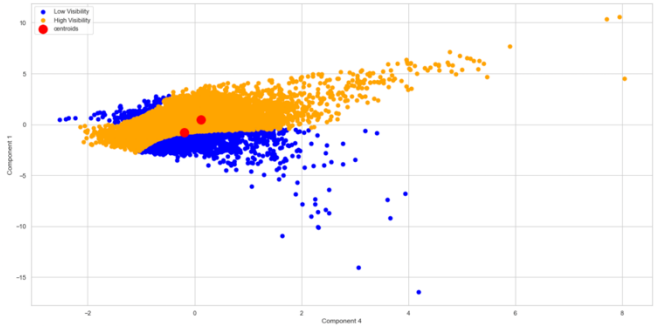
5. Experimenting with Expectation Maximization clustering on Independent Component Analysis by using two components



Cluster_EM	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Holiday	Functioning Day	Spring	Sumi
0	0.574664	11.247841	11.320821	62.121302	1.630587	1400.333192	0.199635	0.180934	0.087771	0.048691	0.967304	0.240623	0.233
1	0.188791	12.551032	19.393982	41.991150	2.118053	1588.932153	2.109139	0.014277	0.022124	0.051917	0.962242	0.299705	0.329

From the feature statistics, we can infer that Cluster 0(green): Low Visibility, and Cluster 1(blue): High Visibility.

6. Experimenting with Expectation Maximization clustering on Random Projections Analysis by using two components.



Cluster_EM	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Holiday	Functioning Day	Spring	Sumi
0	0.345935	14.984982	20.031642	63.733817	2.340186	1211.036769	0.864837	0.603832	0.001036	0.042983	0.954946	0.436044	0.379
1	0.543564	10.514570	10.861517	56.668912	1.550930	1500.670962	0.485490	0.019988	0.096002	0.051106	0.969542	0.200029	0.216

From the feature statistics, we can infer that Cluster 0(blue): High Humidity, and Cluster 1(orange): Low Humidity.

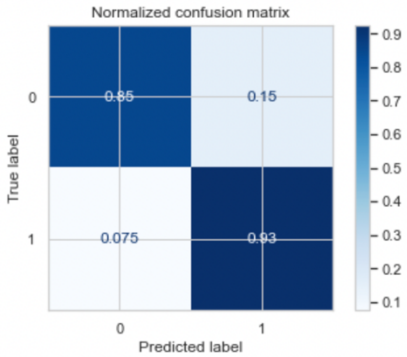
Conclusion:

1) After applying Dimensionality reduction techniques, it is observed there is a reduction in the number of clusters, reduction in noise is observed as we see the data points are closer to the centroid and each other.
2) It is observed that based on temperature, visibility, wind speed, and humidity clusters are being formed.

Task 4: Applying Neural Networks

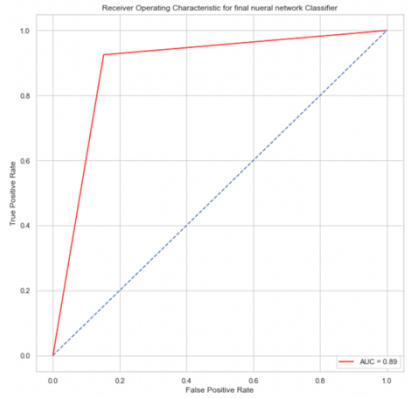
For the experimentation, we can use the optimized neural network produced in Assignment 03 using hidden_layer_sizes=(25,25),activation='relu',learning_rate_init=0.001,tol=1e-4,solver='sgd',max_iter=1000000,random_state=42

Neural networks on the complete dataset:



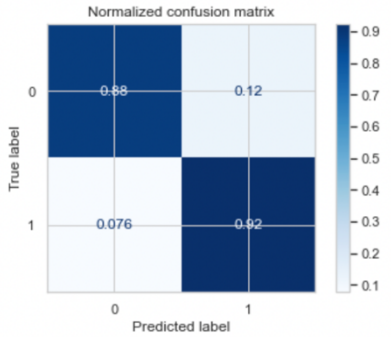
The accuracies of train and test datasets of optimized MLP classifier on complete feature set are:

Neural Network Model	Train Accuracy	Test Accuracy	AUC	Type I Error	Type II Error	Time Taken
All features	88.80	88.70	0.89	15.00	7.50	5.00



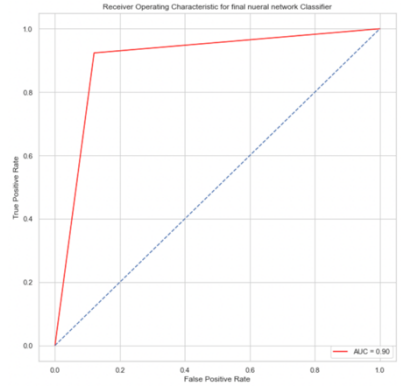
The optimized model with the MLP classifier has an AUC of 0.89. The model with the MLP classifier on a full-featured dataset is good at classifying.

Neural network on the Decision Tree Dimension Reduction:



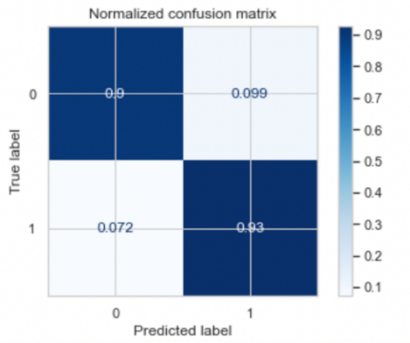
The accuracies of train and test data sets of optimized MLP classifier on Decision Tree Dimension reduction set are:

Neural Network Model	Train Accuracy	Test Accuracy	AUC	Type I Error	Type II Error	Time Taken
Decision Tree	91.00	90.18	0.90	12.00	7.60	17.00



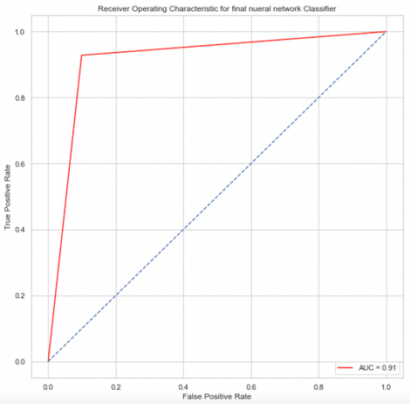
The optimized model with the MLP classifier has an AUC of 0.90. The model with the MLP classifier on a Decision Tree Dimension reduced dataset is good at classifying.

Neural networks on the PCA Dimension Reduction dataset:



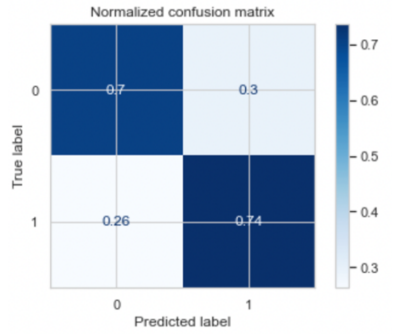
The accuracies of train and test data sets of optimized MLP classifier on PCA Dimension reduction set are:

Neural Network Model	Train Accuracy	Test Accuracy	AUC	Type I Error	Type II Error	Time Taken
PCA	91.78	91.48	0.91	9.90	7.20	17.56



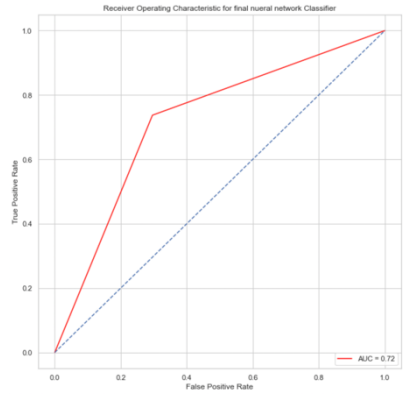
The optimized model with the MLP classifier has an AUC of 0.91. The model with the MLP classifier on a PCA Dimension reduced dataset is good at classifying.

Neural networks on the ICA Dimension Reduction dataset:



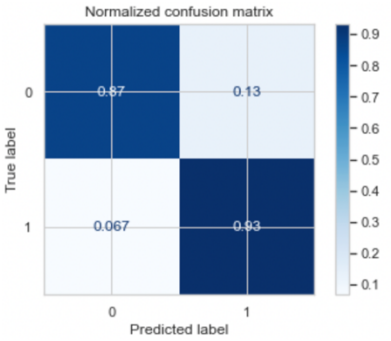
The accuracies of train and test data sets of optimized MLP classifier on ICA Dimension reduction set are:

Neural Network Model	Train Accuracy	Test Accuracy	AUC	Type I Error	Type II Error	Time Taken
ICA	70.14	72.07	0.72	30.00	26.00	0.5



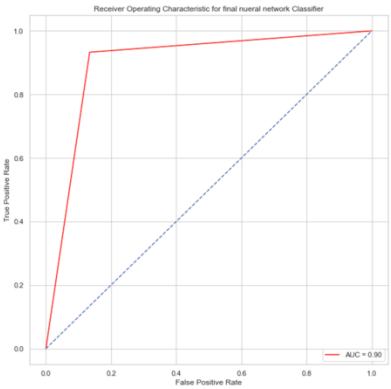
The optimized model with the MLP classifier has an AUC of 0.72. The model with the MLP classifier on an ICA Dimension reduced dataset is not great at classifying.

Neural networks on the Random Projection Dimension Reduction dataset:



Experimented on a different number of components, and best key performance indexes are projected when a number of components/ features is 11. The accuracies of train and test data sets of optimized MLP classifier on Random Projection Dimension reduction set are:

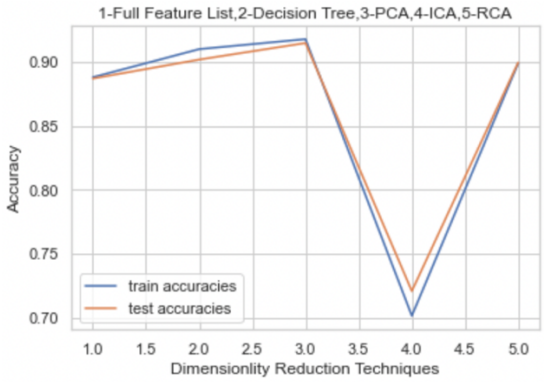
Neural Network Model	Train Accuracy	Test Accuracy	AUC	Type I Error	Type II Error	Time Taken
RP	88.89	89.95	0.90	13.00	6.70	4

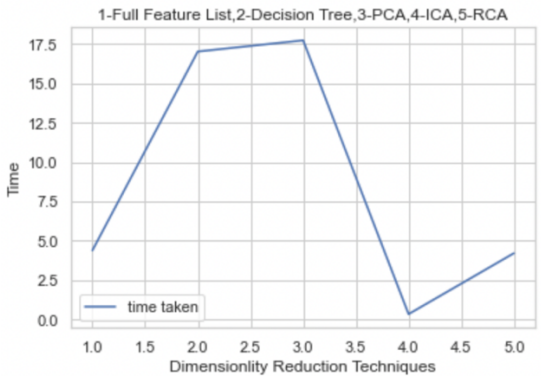


The optimized model with the MLP classifier has an AUC of 0.90. The model with the MLP classifier on a Random Projection Dimension reduced dataset is great at classifying.

Conclusion:

Neural Network Model	Train Accuracy	Test Accuracy	AUC	Type I Error	Type II Error	Time Taken
All features	88.80	88.70	0.89	15.00	7.50	5.00
Decision Tree	91.00	90.18	0.90	12.00	7.60	17.00
PCA	91.78	91.48	0.91	9.90	7.20	17.56
ICA	70.14	72.07	0.72	30.00	26.00	0.5
RP	88.89	89.95	0.90	13.00	6.70	4





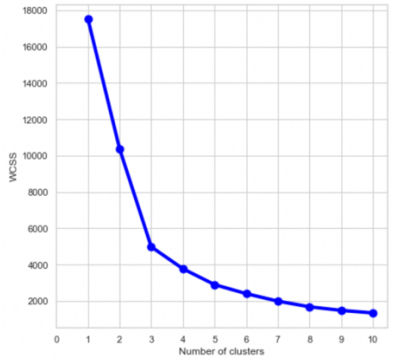
i) We can see that the Random Projection reduced feature is the best one in terms of computation and accuracy which is close to 90%.

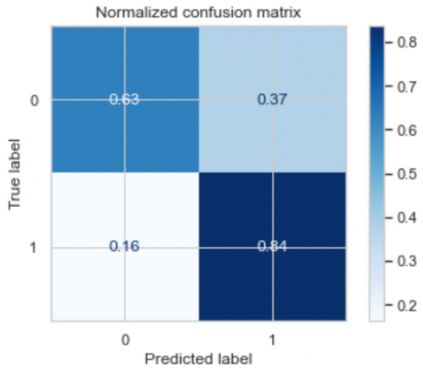
ii) We can see ICA is not performing well even after reducing the feature set where type 1 errors are more which is crucial for our problem and might work better for image processing or computer vision problems.

iii) PCA and DT have done good w.r.t performance and taken more time.

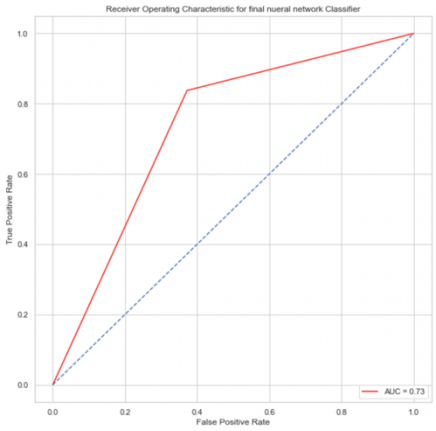
Task 5: Neural Networks on clustering algorithms

Experimenting with the neural networks on the clustering algorithms. We can use the optimized neural network produced in Assignment 03 using hyperparameters of hidden_layer_sizes=(25,25), activation='relu', learning_rate_init=0.001, tol=1e-4, solver='sgd', max_iter=1000000, random_state=42





The accuracies of train and test data sets of optimized MLP classifier on the set of clustering results as features and class label as the output are:



Neural Network Model	Train Accuracy	Test Accuracy	AUC	Type I Error	Type II Error
Cluster	72.11	73.34	0.73	37.00	16.00

The optimized model with the MLP classifier has an AUC of 0.73. The model with the MLP classifier on a clustering dataset is moderately good at classifying.

Conclusion:

Neural Network Model	Train Accuracy	Test Accuracy	AUC	Type I Error	Type II Error
Cluster	72.11	73.34	0.73	37.00	16.00
All feature set	88.80	88.70	0.89	15.00	7.50

We can see that with only cluster results accuracy is down by 17% but the time taken has reduced to 60% and type 1 errors are also comparatively better which is our primary problem statement.

Results:

1) When Applied K Means and EM on the entire feature set it is observed an ideal number of clusters for K-means is 3 and for EM is 2 which uses either temperature or visibility for clustering.
2) During the Dimensionality reduction technique, we observe DT has reduced the feature set to 7, PCA has reduced the feature set to 8, ICA has made the data set mutually independent, and RP has reduced the noise in the data set.
3) Applying K-Means and EM on reduced data set it is observed that the number of clusters has been reduced and used Temperature, visibility, and humidity for clustering and it concludes these 3 are the main features.
4) Applying Neural Network algorithm on reduced data set:

Neural Network Model	Train Accuracy	Test Accuracy	AUC	Type I Error	Type II Error	Time Taken
All features	88.80	88.70	0.89	15.00	7.50	5.00
Decision Tree	91.00	90.18	0.90	12.00	7.60	17.00
PCA	91.78	91.48	0.91	9.90	7.20	17.56
ICA	70.14	72.07	0.72	30.00	26.00	0.5
RP	88.89	89.95	0.90	13.00	6.70	4

5) When Clusters results alone are considered even though there is performance variation but improved time to 60% which is good.
6) Overall either I consider DT for dimensionality reduction and perform ANN on top of that or Cluster the dataset uses the cluster output and performs ANN.