Prepared by : DharmaTeja Chelamalasetty, id : DXC210022 Date : 08/07/2022

BUAN 6341 Applied Machine Learning

ASSIGNMENT NO 4

Seoul Rental Bike prediction Part IV

Executive Summary

- **Converted the data set to 2 classes based on Rented Bike Count median value, if <= Median it is treated as class 0, otherwise class 1.**
- **Performed K-Means and EM Clustering techniques and observed model uses temperature or visibility for clustering .**
- **Applied DT,PCA,ICA,RCA reduction techniques and observed DT has reduced feature set to 4, PCA to 8,ICA has made features mutually independent when observed data distributions and RCA has made the noise less distributions by multiplying with random matrix .**
- **Performed K-Means and EM on Reduced feature set and observed clustering is based on Temperature, visibility, Humidity and Windspeed, so these are the main features .**
- **Applied ANN on reduced feature set and observed DT reduced set is far better in terms of performance and time taken, though PCA improved the performance but increased the time, same with RCA, whereas ICA has reduced the performance but time taken has been improved close to 85~95% which can infer that ICA might not perform good on this dataset as it might have less localized features .**
- **Applied ANN on the results of the clustering and observed time taken has improved by 60% whereas model performance is down by 17% .**
- **Overall from this experiment we can either consider DT 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 previous assignment convert in to classification problem statement, then applying clustering algorithms such as K-Means and EM, then performing 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 dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information. We observe collinearity between Temperature and Dew Point temperature and dropped Dew point temperature , and observed Date is not much use to the classification problem and dropped

the same and performed hot encoding on seasons and dropped 1 season feature in order to avoid collinearity .

Project Outline

The Project is outlined to have 5 parts:

Part 1 : Run the clustering algorithms on your dataset and describe your observations (with plots) using K-Means and EM .

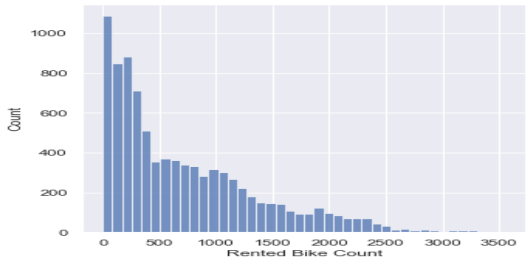
Part 2 : Apply the dimensionality reduction algorithms on your dataset and describe your observations(with plots) using Decision Trees, PCA, ICA and RCA.

Part 3: Run the clustering algorithms again, this time after applying dimensionality reduction. Describe the difference compared to previous experimentation (with plots).

Part 4: Run your neural network learner from assignment 3 on the data after dimensionality reduction (from task 2). Explain and plot your observations (error rates, etc.)

Part 5: Use the clustering results from task 1 as the new features and apply neural network learner on this new data consisting of only clustering results as features and class label as the output. Again, plot and explain your results

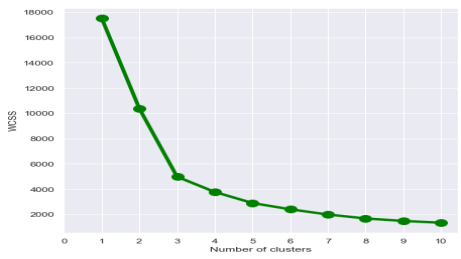
Part 1: Clustering



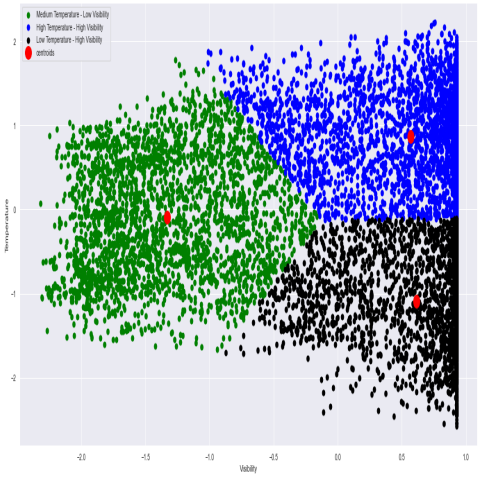
We can observe from the figure that the data of the rented bike count is right skewed. So, we apply a median technique to divide the data in 2 distributions and to handle outliers.
Since the Median value is 504.5., Less than or equal to the value is considered as class 0 and greater than the value is considered as class 1.

K-Means Clustering :

We will use elbow method to determine the nr of clusters for K-Means .



After conducting experiment using WCSS score it is observed that **nr of clusters = 3** can be baselined



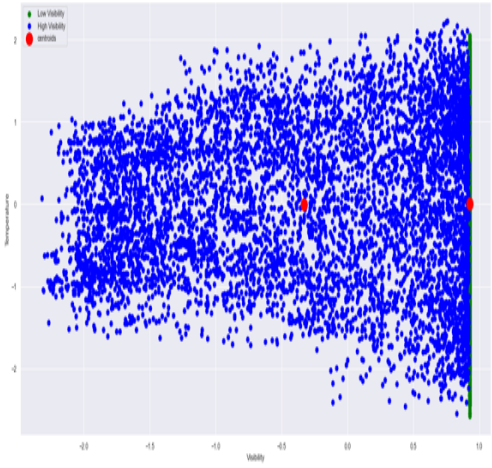
After Applying K-Means Clustering technique on the data set, we can observe that:

Cluster_KM	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall	Snowfall	Seasons_Spring	Seasons_Summer	Sea
0	0.385932	10.646818	11.755601	72.613323	1.498400	627.713063	0.358679	0.393822	0.14038	0.379978	0.175288	
1	0.780148	12.478953	23.201081	55.802332	1.770250	1782.863197	0.888393	0.058362	0.00000	0.201934	0.494027	
2	0.234650	11.050450	-0.120415	46.440751	1.900587	1811.256160	0.351212	0.015291	0.10966	0.186547	0.000000	

Cluster 0: Medium Temp – Low Visibility
Cluster 1: High Temp – High Visibility
Cluster 2: Low temp – High Visibility
It's observed that when **temp is medium rental bike count is medium, when temp is high rental bike count is high, when temp is low rental bike count is low** irrespective of visibility.

EM Clustering :

In order to derive the ideal nr of cluster for EM we have done experimenting on hyperparameters n_components and covariance type using BIC values and observed that covariance type = 'full' and n_components = 2, BIC value is low, hence we finalized these hyper parameters .



After Applying EM Clustering technique on the data set, we can observe that:

Cluster_EM	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall	Snowfall	Seasons_Spring	Seasons_Summer	Sea
0	0.537424	12.162750	12.911923	50.401654	1.793429	1999.966057	0.411989	0.009399	0.036292	0.166232	0.243255	
1	0.486691	11.264314	12.872609	61.008821	1.700542	1236.563293	0.624986	0.198220	0.088858	0.282575	0.255184	

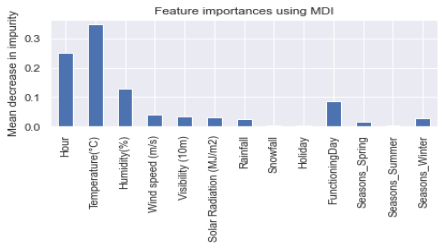
Cluster 0: High Visibility
Cluster 1: Low Visibility
It's observed that EM technique used only visibility to perform clustering and can be seen that when **High visibility rental bike count is more than median value and in low visibility rental bike count is less than median value.**

Conclusion :

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

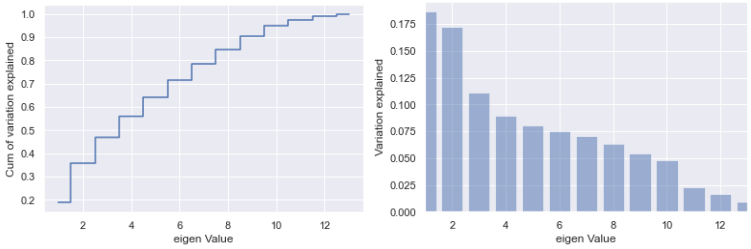
Part 2 : Dimensionality Reduction

i) Using Decision Trees :



Using Decision tree feature importance, it is observed that 4 features are much more important than compared to others and they are Hour, Temperature, Humidity and Functioning day.

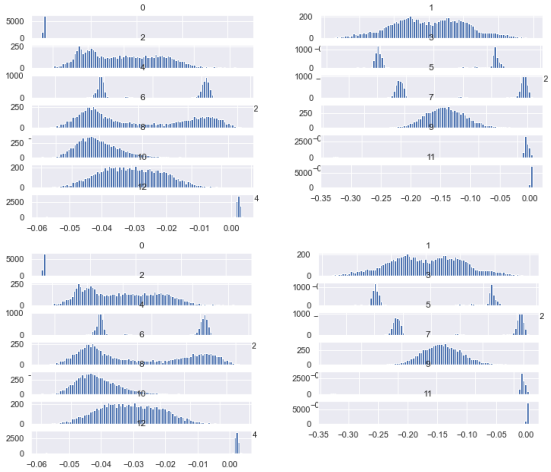
ii) PCA :



Using PCA we can observe that using 8 features 85% of the variation is explained, and hence we will use 8 features using PCA.

iii) ICA :

As part of experiment we performed ICA on the data that is transformed using PCA technique in order to reduction further noise in the data . Below is the projection of distribution of features using algorithms **'parallel' & 'deflation' .**

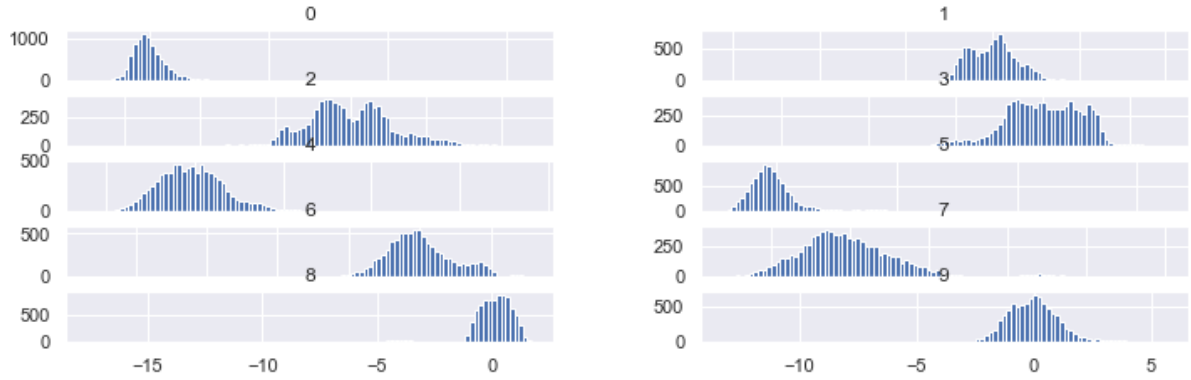


Using ICA using different algorithms we see there is no change in distribution of data which might have no difference in finding local features, hence either one can be used and observe that features are more mutually independent.

iv) RCA :

Gaussian random method projects the original input space on a randomly generated matrix to reduce dimensions. We'll define the model by using the Gaussian Random Projection class by setting the components numbers. Here, we'll shrink the feature data from 14 to [5,9,10,11,12].

Below is the distribution of n_components=10 .



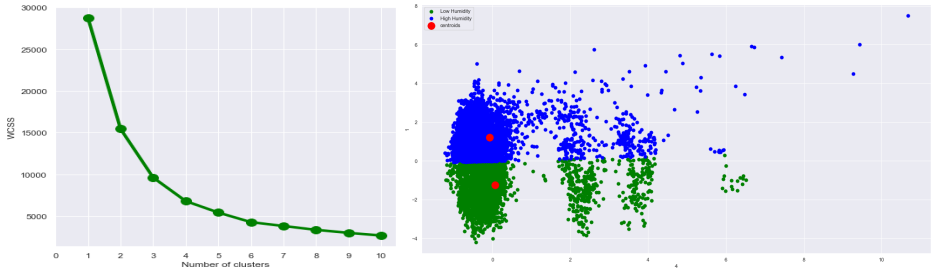
Conclusion :

1) DT has reduced feature set to 4 :- Hour, Temp, humidify and Functioning day .
2) PCA has reduced 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 model .
4) RCA is performed and baselined components to 10 after the Task 4 .

Part 3:

i) Applying K Means & EM on PCA :

K-Means :



After Applying KMeans Clustering technique on the PCA data set, we can observe that Clusters are 2 from elbow method and formed based on Humidity

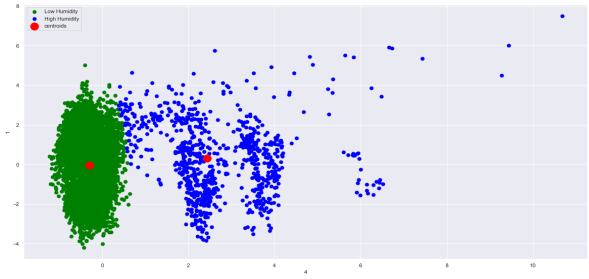
Cluster	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall	Snowfall	Seasons_Spring	Seasons_Summer	Seaso
0	0.661421	14.811155	13.998935	43.631335	2.263157	1738.422819	1.032236	0.001990	0.030386	0.240685	0.247628	
1	0.342870	8.276864	11.796576	72.433206	1.200969	1143.246001	0.118297	0.291485	0.118563	0.263122	0.256364	

Cluster 0: Low Humidity

Cluster 1: High Humidity

It's observed that KMeans technique used only Humidity to perform clustering and can be seen that when **Low Humidity rental bike count is more than median value and in Low Humidity rental bike count is less than median value.**

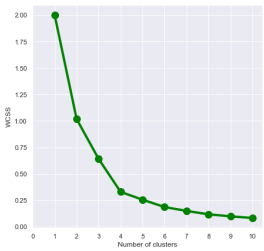
EM :

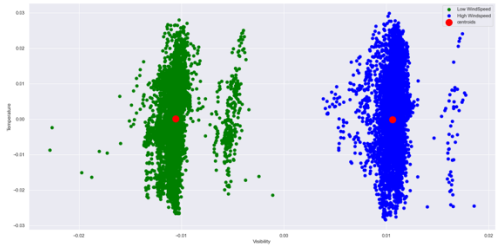


After Applying EM Clustering technique on the PCA data set, we can observe that Clusters are formed based on humidity.
Cluster 0: Low Humidity
Cluster 1: High Humidity
It's observed that EM technique used only Humidity to perform clustering and can be seen that when **Low humidity rental bike count is more than median value and in high Humidity rental bike count is less than median value.**

ii) Applying KMeans & EM on ICA Data set :

K Means :



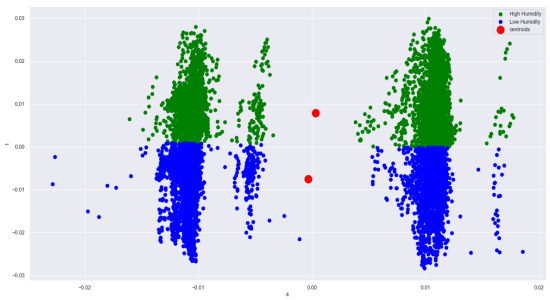


After Applying KMeans Clustering technique on the ICA data set, we can observe that Clusters are 2 from elbow method and formed based on Windspeed.

Cluster	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall	Snowfall	Seasons_Spring	Seasons_Summer	Seasons_Wir
0	0.695128	11.5	20.385861	62.120674	1.552391	1531.087659	0.64219	0.188479	0.028005	0.000000	0.502732	0.000
1	0.303800	11.5	5.338759	54.310440	1.898375	1342.046016	0.49563	0.108677	0.122390	0.505495	0.000000	0.494

Cluster 0: Low Windspeed
Cluster 1: High Windspeed
It's observed that KMeans technique used only windspeed to perform clustering and can be seen that when **Low Windspeed rental bike count is more than median value and in High Windspeed rental bike count is less than median value.**

EM :



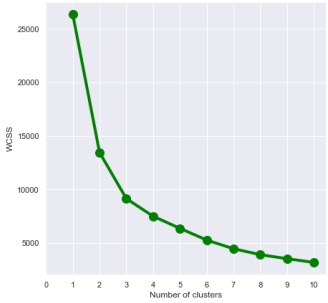
After Applying EM Clustering technique on the ICA data set, we can observe that Clusters are formed based on Humidity:

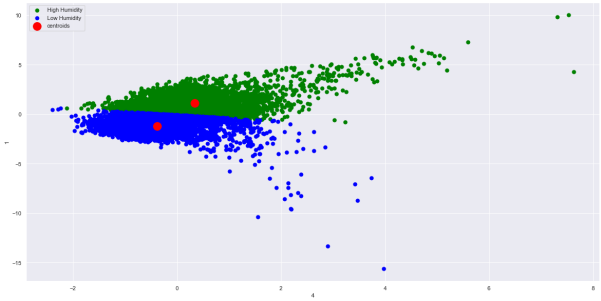
Cluster_EM	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall	Snowfall	Seasons_Spring	Seasons_Summer	Seasons
0	0.695128	11.5	20.385861	62.120674	1.552391	1531.087659	0.64219	0.188479	0.028005	0.000000	0.502732	0
1	0.303800	11.5	5.338759	54.310440	1.898375	1342.046016	0.49563	0.108677	0.122390	0.505495	0.000000	0

Cluster 0: High Humidity
Cluster 1: Low Humidity
It's observed that EM technique used only temperature to perform clustering and can be seen that when **High Humidity rental bike count is less than median value and in low Humidity rental bike count is more than median value.**

iii) Applying KMeans & EM on RCA Data set :

K Means :





After Applying KMeans Clustering technique on the RCA data set, we can observe that Clusters are 2 from elbow method and formed based on Humidity.

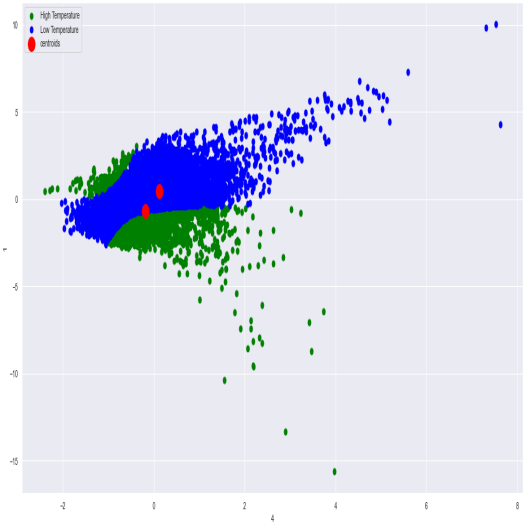
Cluster	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall	Snowfall	Seasons_Spring	Seasons_Summer	Seasons
0	0.250978	8.175359	6.162299	59.114298	1.462538	1419.942634	0.202112	0.023381	0.142894	0.128857	0.138635	
1	0.775613	15.179654	20.321188	57.243386	2.015296	1455.511785	0.975298	0.287374	0.000000	0.388408	0.377585	

Cluster 0: High Humidity

Cluster 1: Low Humidity

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

EM :



After Applying EM Clustering technique on the RCA data set, we can observe that Clusters are formed based on temperature:

Cluster_EM	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall	Snowfall	Seasons_Spring	Seasons_Summer	Sea
0	0.660617	14.791289	19.876225	62.671960	2.281760	1253.903358	0.866547	0.535390	0.000907	0.424229	0.374773	
1	0.446004	10.393533	10.531910	56.731696	1.537706	1498.320775	0.469118	0.018685	0.100000	0.194173	0.210799	

Cluster 0: High Temperature

Cluster 1: Low Temperature

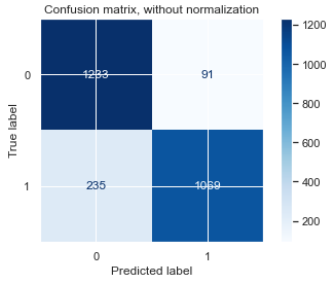
It's observed that EM technique used only temperature to perform clustering and can be seen that when **High Temperature rental bike count is more than median value and in low Temperature rental bike count is less than median value.**

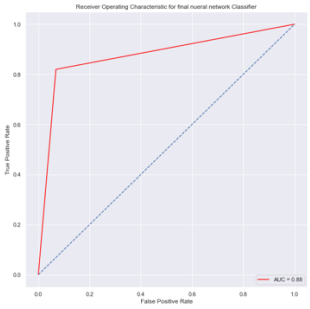
Conclusion :

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

Part 4 : Neural Network Implementation

i) Applying Neural Network on Full Feature set :

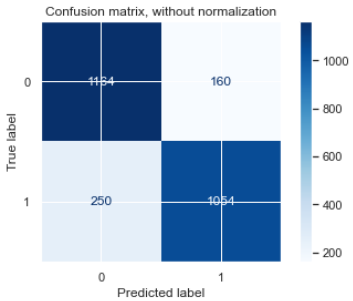


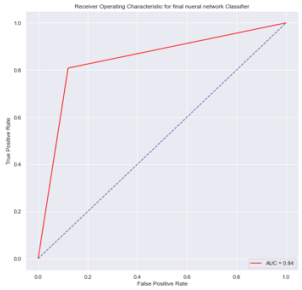


Implementing Neural Network on entire dataset model has achieved train and test accuracies close to 88 and AUC is also 88, Type I Errors being 91 and Type II Errors being 295 and time taken is 8.5 secs

Neural Network on fFeature Set	Train Accuracy	Test Accuracy	AUC	Time Taken
Full Feature set	87.9	87.5	88	8.5

ii) Applying Neural Network on Dimensionality Reduction set by Decision Trees :

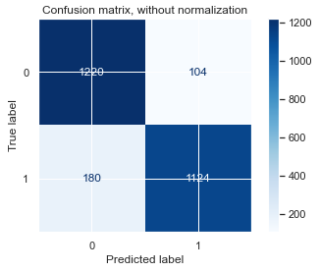


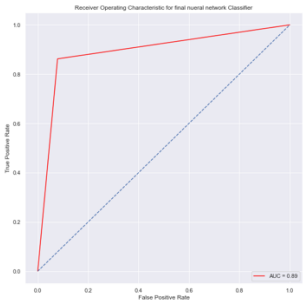


Implementing Neural Network on reduced data of DT model has achieved train and test accuracies close to 84 and AUC is also 84, Type I Errors being 160 and Type II Errors being 250 and time taken is 2.4 secs

Neural Network on fFeature Set	Train Accuracy	Test Accuracy	AUC	Time Taken
Dimensionlaity Reduction DT	84.7	84.3	84	2.4

iii) Applying Neural Network on Dimensionality Reduction set by PCA :

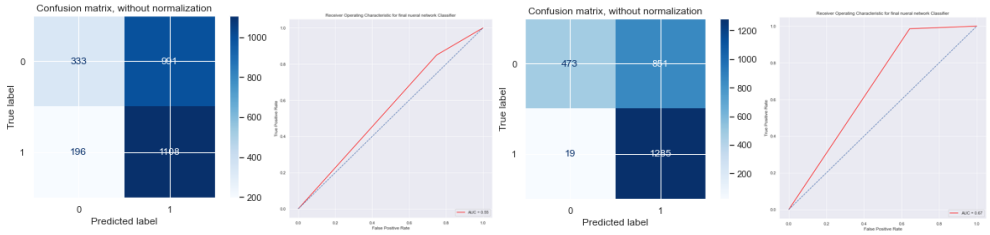




Implementing Neural Network on reduced data of PCA technique model has achieved train and test accuracies close to 89 and AUC is also 89, Type I Errors being 104 and Type II Errors being 180 and time taken is 13.03 secs

Neural Network on fFeature Set	Train Accuracy	Test Accuracy	AUC	Time Taken
Dimensionlaity Reduction PCA	88.7	89.1	89	13.03

iv) Applying Neural Network on Dimensionality Reduction set by ICA :

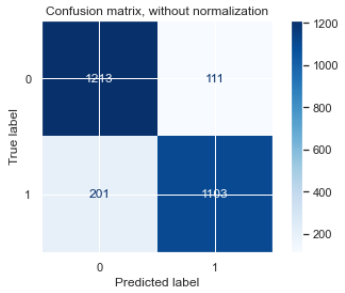


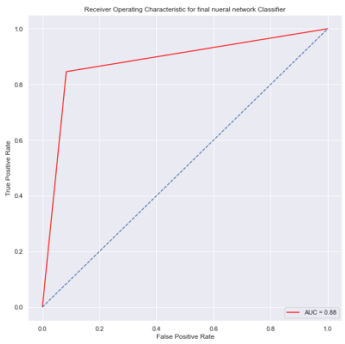
Implementing Neural Network on reduced data of ICA has 2 parts with full feature set and with reduced feature set after applying feature importance from DT model.
Below are the results :

Neural Network on fFeature Set	Train Accuracy	Test Accuracy	AUC	Time Taken
Dimensionlaity Reduction ICA	55.7	54.8	55	0.365
Dimensionlaity Reduction ICA after applying DT	65.8	66.8	67	0.46

With full feature set we see train and test are close to 55 even AUC is at 55 but time taken by the model is very low at 0.365 secs
With reduced feature ICA data set we see train and test are close to 66 with AUC being 67 and time taken is also low at 0.46 secs

v) Applying Neural Network on Dimensionality Reduction set by RCA :



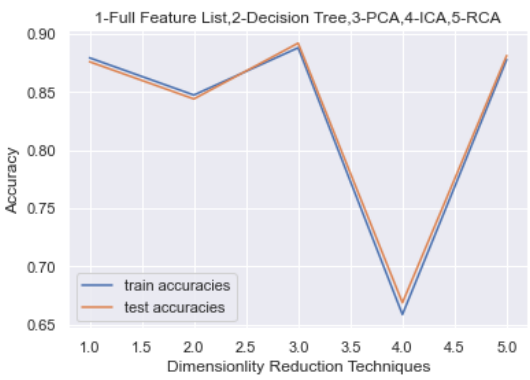


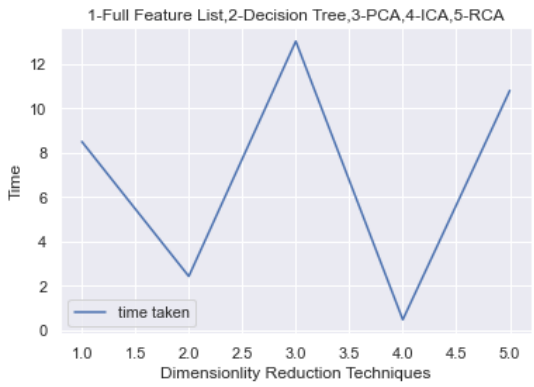
We ran the Neural Network with RCA data set with having n_components as 5,9,10,11,12 and observe that n_components have best time and accuracy. hence only showcasing the results of the best tuned model.
Train and Test Accuracies being 87 and 88 with AUC being at 88 and time taken is 10 secs .

Neural Network on fFeature Set	Train Accuracy	Test Accuracy	AUC	Time Taken
Dimensionlaity Reduction RCA	87.1	88.1	88	10.01

Conclusion :

Neural Network on fFeature Set	Train Accuracy	Test Accuracy	AUC	Time Taken	% Change in Accuracy	% Change in Time	Type 1 Erros	Type II Errors
Full Feature set	87.9	87.5	88	8.5	-	-	91	295
Dimensionlaity Reduction DT	84.7	84.3	84	2.4	-4.54545455	71.76470588	160	250
Dimensionlaity Reduction PCA	88.7	89.1	89	13.03	1.136363636	-53.29411765	104	180
Dimensionlaity Reduction ICA	55.7	54.8	55	0.365	-37.5	95.70588235	991	196
Dimensionlaity Reduction ICA after applying DT	65.8	66.8	67	0.46	-23.8636364	94.58823529	851	19
Dimensionlaity Reduction RCA	87.1	88.1	88	10.01	0	-17.76470588	111	201





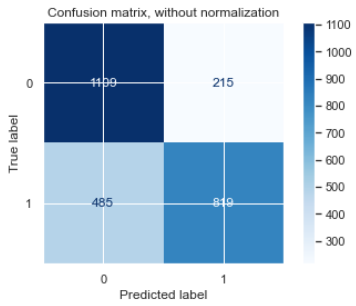
i) We can see that Decision Tree reduced feature is clearly the best one in terms of time taken and reduction of accuracy which is close to -4.5%. but increased time of 71% .

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

iii) PCA has done good w.r.t performance but it has increased time to 53% .

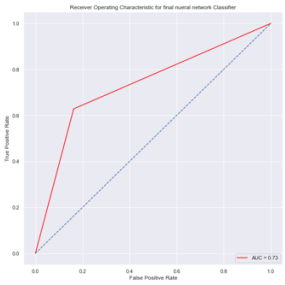
iv) RCA performed better w.r.t performance but has increased time to 18% .

Part 5: Applying Neural Network on Clustering Results :



Implementing Neural Network on only on cluster results model has achieved train and test accuracies close to 72 and AUC is also 73, Type I Errors being 215 and Type II Errors being 485 and time taken is 3.4 secs

Neural Network on fFeature Set	Train Accuracy	Test Accuracy	AUC	Time Taken
Clusrter Results	72.1	73.1	73	3.4



Conclusion :

Neural Network on fFeature Set	Train Accuracy	Test Accuracy	AUC	Time Taken	% Change in Accuracy	% Change in Time	Type 1 Erros	Type II Errors
Full Feature set	87.9	87.5	88	8.5	-	-	91	295
Dimensionlaity Reduction DT	84.7	84.3	84	2.4	-4.54545455	71.76470588	160	250
Dimensionlaity Reduction PCA	88.7	89.1	89	13.03	1.136363636	-53.29411765	104	180
Dimensionlaity Reduction ICA	55.7	54.8	55	0.365	-37.5	95.70588235	991	196
Dimensionlaity Reduction ICA after applying DT	65.8	66.8	67	0.46	-23.8636364	94.58823529	851	19
Dimensionlaity Reduction RCA	87.1	88.1	88	10.01	0	-17.76470588	111	201
Only Cluster results as features	72.1	73.1	73	3.4	-17.0454545	60	215	485

1) We can see that with only cluster results accuracy is down by 17% but 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 entire feature set it is observed ideal nr of clusters for K-means is 3 and for EM is 2 which uses either temperature or visibility for clustering .**
2) **During Dimensionality reduction technique we observe DT has reduced the feature set to 4, PCA has reduced feature set to 8, ICA has made the data set to mutually independent, RCA has reduced the noise in the data set .**
3) **Applying K-Means and EM on reduced data set it is observed that nr of clusters has been reduced and used Temperature, visibility, windspeed and humidity for clustering and it concludes these 4 are main features .**
4) **Applying Neural Network algorithm on reduced data set :**

Neural Network on fFeature Set	Train Accuracy	Test Accuracy	AUC	Time Taken	% Change in Accuracy	% Change in Time	Type 1 Erros	Type II Errors
Full Feature set	87.9	87.5	88	8.5	-	-	91	295
Dimensionlaity Reduction DT	84.7	84.3	84	2.4	-4.54545455	71.76470588	160	250
Dimensionlaity Reduction PCA	88.7	89.1	89	13.03	1.136363636	-53.29411765	104	180
Dimensionlaity Reduction ICA	55.7	54.8	55	0.365	-37.5	95.70588235	991	196
Dimensionlaity Reduction ICA after applying DT	65.8	66.8	67	0.46	-23.8636364	94.58823529	851	19
Dimensionlaity Reduction RCA	87.1	88.1	88	10.01	0	-17.76470588	111	201

It is observed that even though ICA takes very less time performance is not good, RCA data set is good in performance but increased time by 17%, PCA has improved the performance but increased time to 53% , whereas DT has performance variance of -4% but time has significantly improved by 71% which is so far better when compared to all techniques .

5) **When Clusters results alone are considered even though there is performance variation of -17% but increased 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 use the cluster output and perform ANN .**