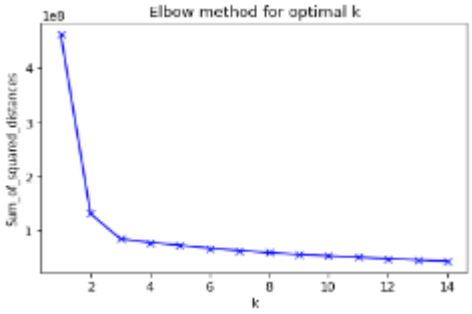
ML ASSIGNMENT 4

In this assignment we will implement the clustering algorithms like K-means and Expectation Maximization on the GPU Runtime data set and Brisbane weather data set. We have applied various dimensionality reduction techniques like PCA, Feature Selection, ICA and RCA.

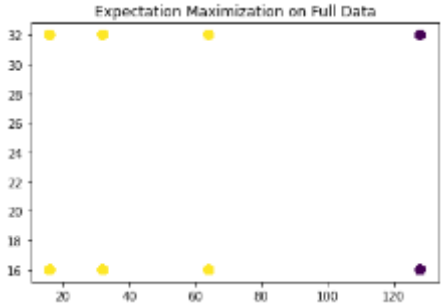
Dataset Description

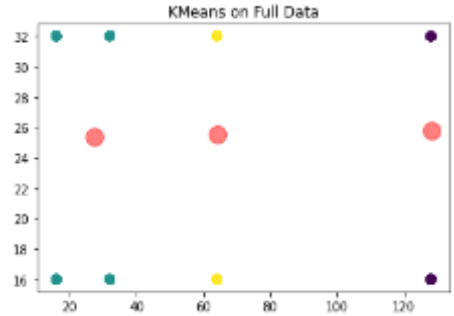
GPU data set consist of 14 features and 241600 rows. First 10 features are ordinal and last 4 as binary variables. In addition to this there are 4 run time variables. We have made a new column named average runtime and have dropped the remaining runtime variables. We have checked for outliers and have split the dataset into training set and testing set with the ratio of 30% testing set and 70% training set.

Experiment 1(Run the clustering algorithms)



Starting with K means algorithms, first of all we will find the best value of k. We will find it by using the Elbow curve.From this elbow curve in the left , we can see that the optimal value of K will be 3 as the SSE value after that starts decreasing at the rapid rate.





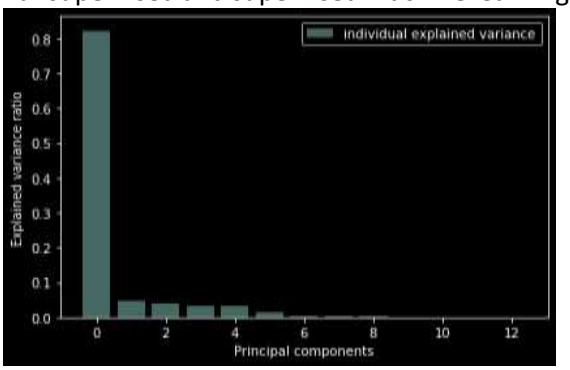
Now, taking the value of K as 3 , we ran K-means on full data set by dropping the target variable. From the figure on the right, we can see that there are 3 clusters and data is separated widely. Similarly we ran Expectation Maximization algorithm and drew the graph of it. We can see that on the left side. Here also we can see that data is widely spread.

EXPERIMENT 2(Dimensionality Techniques)

After performing the above two results, we will apply dimensional reduction techniques such as PCA, ICA and RCA.

PCA

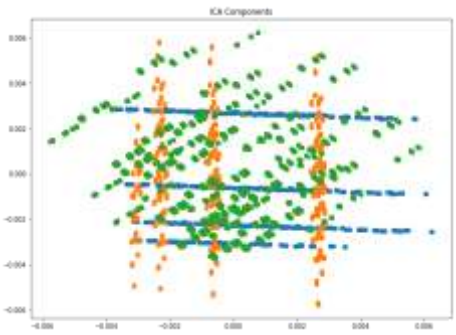
PCA is a dimensionality reduction technique used to transform high-dimensional datasets into a dataset with fewer variables, where the set of resulting variables explains the maximum variance within the dataset. PCA is used prior to unsupervised and supervised machine learning steps to reduce the number of features used in the analysis, thereby reducing the likelihood of error.



After performing PCA on the dataset, I retrieved the explained variance ratios to get a better idea of how principal components describe the variance in the data. Here we can see that the first component PC1, has the largest amount of variance which is about 90 percent.

ICA

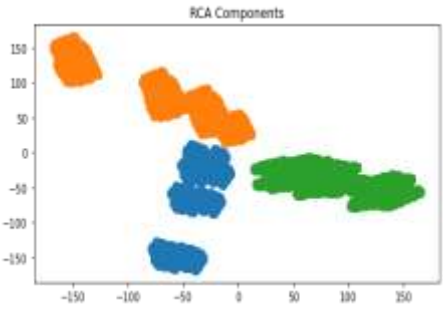
ICA is another dimensionality technique known as Independent component Analysis. Independent Component Analysis (ICA) is based on information-theory and is also one of the most widely used dimensionality reduction techniques. The major difference between PCA and ICA is that PCA looks for uncorrelated factors while ICA looks for independent factors.



Here, n_components will decide the number of components in the transformed data. We have transformed the data into 3 components using ICA.

RCA

Random projection is a technique used to reduce the dimensionality of a set of points The sklearn.random_projection.GaussianRandomProjection reduces the dimensionality by projecting the original input space on a randomly generated matrix where components are drawn from the following distribution N(0,1/ncomponents). We have transformed the data into 3 components.



FEATURE SELECTION

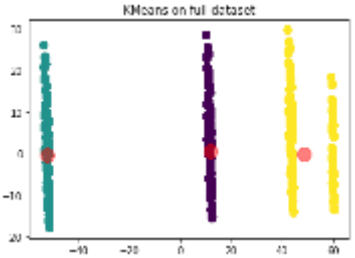
Step forward feature selection starts with the evaluation of each individual feature and selects that which results in the best performing selected algorithm model. Next, all possible combinations of the that selected feature and a subsequent feature are evaluated, and a second feature is selected, and so on, until the required predefined number of features is selected.

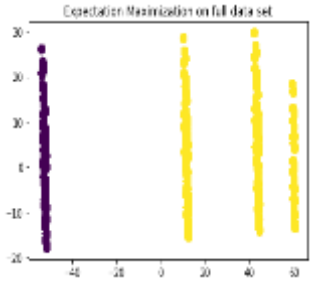
After performing feature selection on the original dataset, I obtained the best 7 features: NWG, KWG,NDIMC,NDIMB,VWN,STRM,SA.

EXPERIMENT 3(Clustering after Dimensionality reduction)

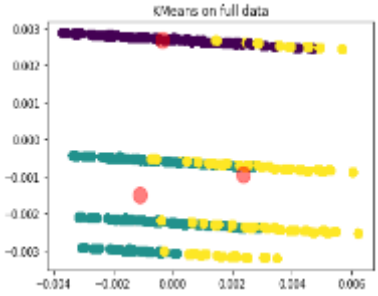
AFTER PCA

Taking the first two PCs that is PC1 and PC2, we will again perform K-means and Expectation maximization. Again going with the elbow curve methodology as shown, we find that the best value of k will be 3. Applying K-Means Algorithm and Expectation Maximization Algorithm, we get the below graphs . In this we see that data is more compact and separated. Hence after applying PCA, the above algorithms performed better and clustered the data more appropriately.

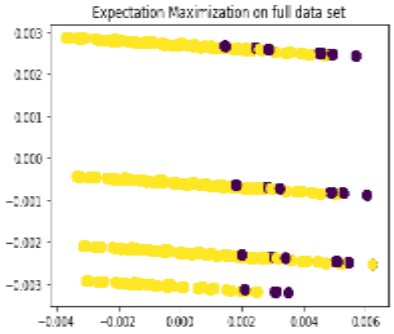




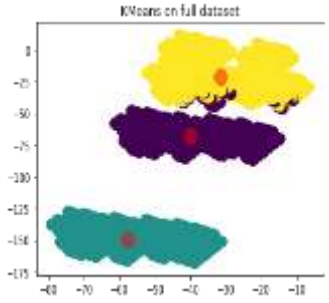
AFTER ICA



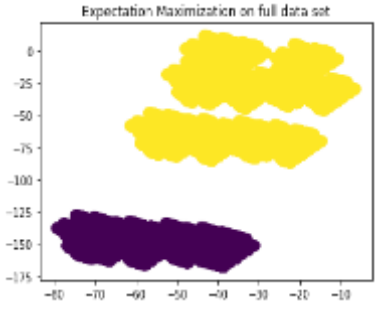
After performing ICA on the dataset, we performed K-means and expectation maximization again. The dataset was transformed into 3 components.In K-means graph we can see that we got three clusters in which one cluster (yellow) is spread across the data where as in Expectation Maximization, I used only 2 clusters for separating the data. In this most of the data belongs to cluster 1(Yellow). In this centers are not that appropriately visible as it was in PCA.



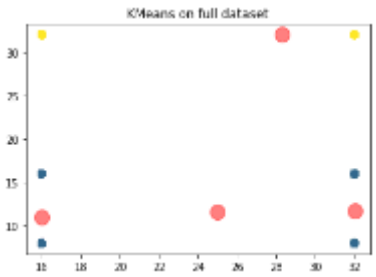
AFTER RCA



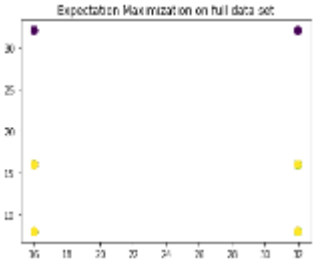
After performing RCA, the K-means graph and Expectation Maximization came out to be as shown. For K-means we used 3 clusters and for Expectation Maximization , we used 2 clusters. For both the algorithms, the clusters are well separated and centers are also properly visible. So far, this is the best result after dimensionality reduction techniques.



AFTER FEATURE SELECTION



After feature selection, we seleted 7 best features and ran K means and Expectation Maximization.Here the data is well separated for both the data sets but the data is very widly separated. After applying feature selection, the data became more scarce and well separated. But the compactness of the data is reduced as the number of features have reduced big time.



COMPARISON

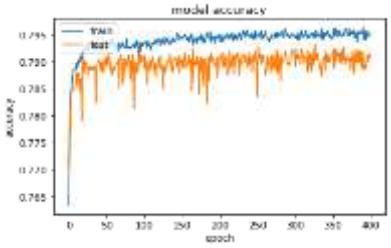
After performing the four dimensionality reduction techniques we can see that for the above dataset Randomized projection did the best job in separating the clusters. As in this data was compact for the given clusters and centers were well separated.

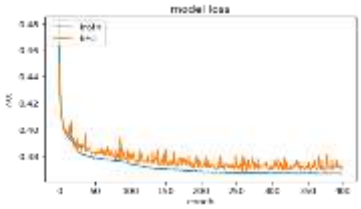
EXPERIMENT 4 (ARTIFICIAL NEURAL NETWORK)

In this dataset we were getting the best accuracy for 400 epochs , relu function and with 16 epochs.The accuracy which we recived for ANN previously was 88 percent.

We applied PCA, for the data set and we got the following results. The accuracy which we received for the same combination after applying PCA was 53 percent. There is a steep fall in accuracy for this dataset. Hence we concludes that PCA is not the apprporiate dimension reduction technique for this dataset. We can see the number of correctly classified datapoints are 16379. The model accuracy and model loss graph is shown below:

Confusion Matrix	Positive (1)	Negative (0)
Positive (1)	16379	26805
Negative (0)	3091	18175





In the accuracy graph, we can see that training and testing accuracy is increasing with the increase in the epochs. Similarly in the model loss, we can see that the model loss is reducing with the number of epochs.

EXPERIMENT 5 (CLUSTERING RESULTS AS NEW FEATURES)

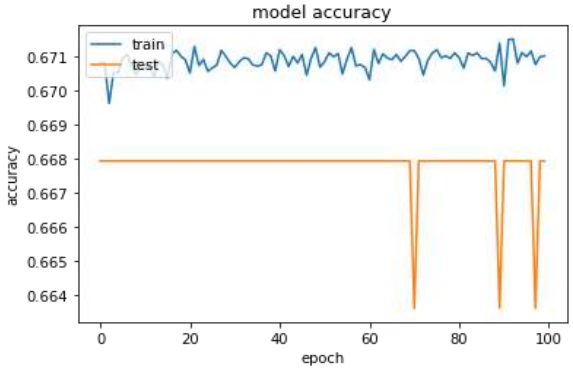
In this task, I have added a column of cluster labels to the dataset as a feature. I will be using the class labels to perform Neural Network Classification. In this we have taken the results from k- means and expectation maximization from task 1 as my x variables and original y variable as my target variable.

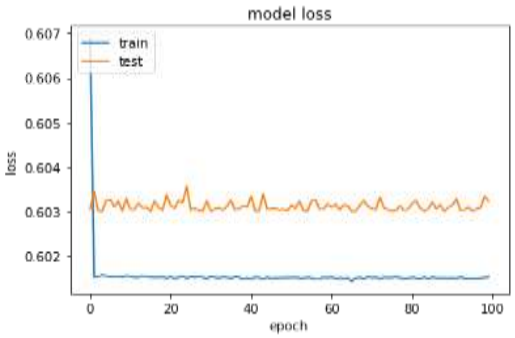
After applying artificial neural network algorithm to the newly created dataset and found out the accuracy. I have used the softmax activation function with 100 epochs. The confusion matrix for the above performed experiment is shown here.

From the confusion matrix here we can see that both true positives and false positives are high but the number of true positive is still greater. The overall accuracy we got for this model is 67 percent.

Confusion Matrix	Positive (1)	Negative (0)
Positive (1)	43184	0
Negative (0)	21266	0

From this we can infer that our alorithms k- means and Expectation Maximization did a good job. Here are the learning curves which we got after running ANN.





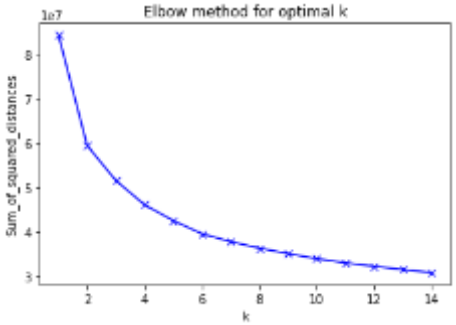
In the accuracy graph, we can see that training accuracy is almost same as the number of epochs are increased where are testing accuracy is almost same but have steep dip at some points. Similarly in the model loss graph, we can see that the model loss has a deep steep and remains constant throughout where as in testing the model is around.603 but then remains constant for the increasing number of epochs.

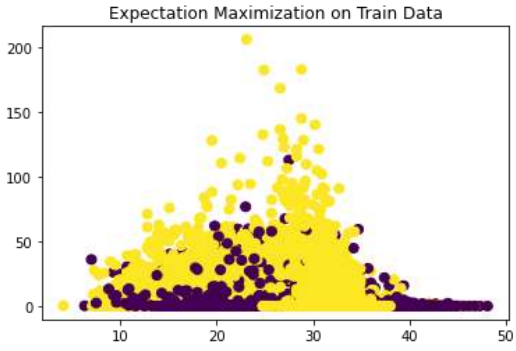
DATASET 2

Brisbane weather dataset contains 24 variables. It contains 56240 rows. In this I have applied classification algorithm to decide whether it will rain tomorrow or not. There are total 7 categorical variables including date column. We will create dummy variables for categorical variables. We have checked for outliers and have split the dataset into training set and testing set with the ratio of 30% testing set and 70% training set.

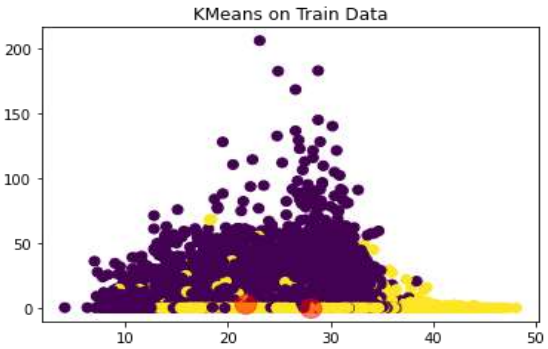
Experiment 1(Run the clustering algorithms)

Starting with K means algorithms, first of all we will find the best value of k. We will find it by using the Elbow curve.From this elbow curve in the left , we can see that the optimal value of K will be 2 as the SSE value after that starts decreasing at the rapid rate.





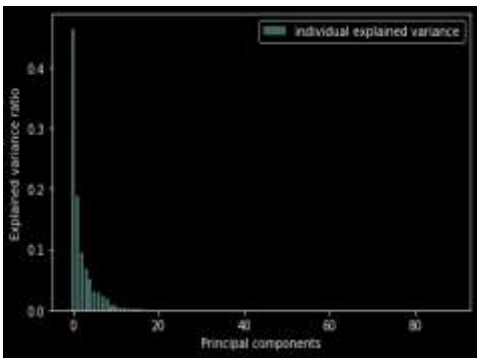
Now, taking the value of K as 2, we ran K-means on full data set by dropping the target variable. From the figure on the right, we can see that there are 2 clusters and data is separated. Similarly we ran Expectation Maximization algorithm and drew the graph of it. We can see that on the left side. Here also we can see that data is separated. In this we got the clusters but the clusters are merging into each other and the centers are not clearly defined.



EXPERIMENT 2(Dimensionality Techniques)

After performing the above two results, we will apply dimensional reduction techniques such as PCA, ICA and RCA.

PCA

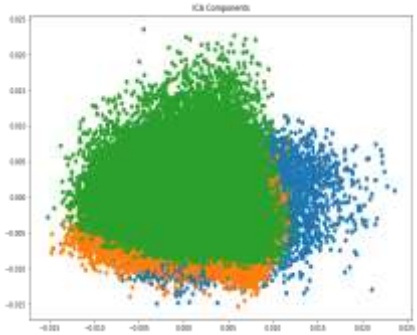


PCA is a dimensionality reduction technique used to transform high-dimensional datasets into a dataset with fewer variables, where the set of resulting variables explains the maximum variance within the dataset. PCA is used prior to unsupervised and supervised machine learning steps to reduce the number of features used in the analysis, thereby reducing the likelihood of error.

After performing PCA on the dataset, I retrieved the explained variance ratios to get a better idea of how principal components describe the variance in the data. Here we can see that the first component PC1, has the largest amount of variance which is about 90 percent.

ICA

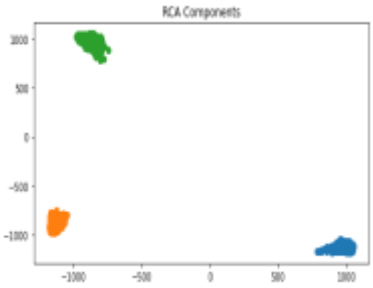
ICA is another dimensionality technique known as Independent component Analysis. Independent Component Analysis (ICA) is based on information-theory and is also one of the most widely used dimensionality reduction techniques. The major difference between PCA and ICA is that PCA looks for uncorrelated factors while ICA looks for independent factors.



Here, n_components will decide the number of components in the transformed data. We have transformed the data into 3 components using ICA. Left side is the graph for it.

RCA

Random projection is a technique used to reduce the dimensionality of a set of points The sklearn.random_projection.GaussianRandomProjection reduces the dimensionality by projecting the original input space on a randomly generated matrix where components are drawn from the following distribution N(0,1/ncomponents). We have transformed the data into 3 components. To the right side is the graph, we can see the three components separated widely.



FEATURE SELECTION

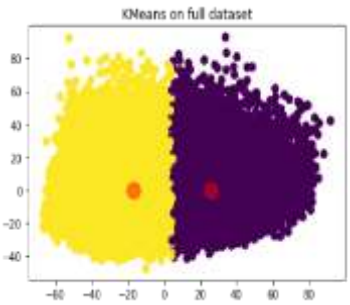
Step forward feature selection starts with the evaluation of each individual feature and selects that which results in the best performing selected algorithm model. Next, all possible combinations of the that selected feature and a subsequent feature are evaluated, and a second feature is selected, and so on, until the required predefined number of features is selected.

After performing feature selection on the original dataset, I obtained the best 20 features.

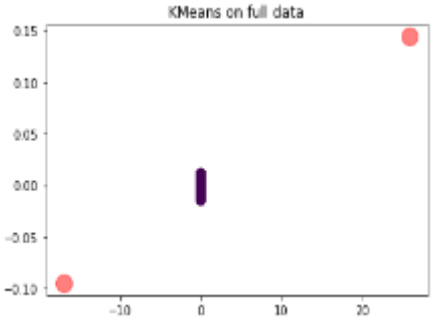
EXPERIMENT 3(Clustering after Dimensionality reduction)

AFTER PCA

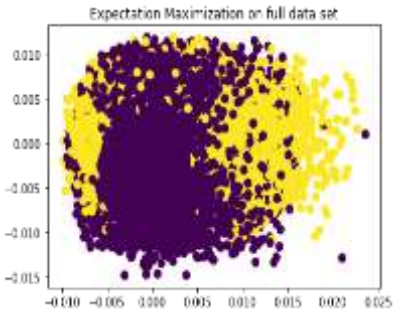
Taking the first two PCs that is PC1 and PC2, we will again perform K-means and Expectation maximization. Again going with the elbow curve methodology as shown, we find that the best value of k will be 2. Applying K-Means Algorithm and Expectation Maximization Algorithm, we get the below graphs . In this we see that data is more compact and separated. Hence after applying PCA, the above algorithms performed better and clustered the data perfectly.



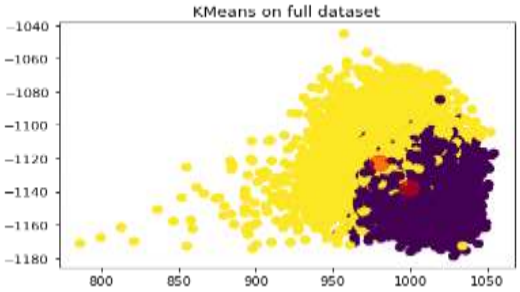
AFTER ICA



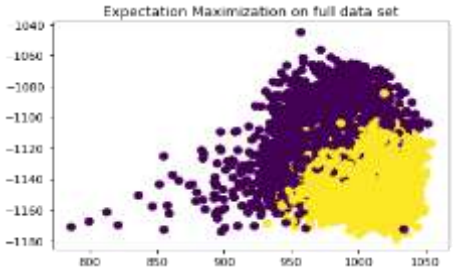
After performing ICA on the dataset, we performed K-means and expectation maximization again. The dataset was transformed into 2 components.In K-means graph we can see that we got 2clusters because we can see two centers but the data got widely separated and did not gave good presentation where as in Expectation Maximization we got 2 clusters for separating the data. In this however soft clustering is done, therefore the two clusters are merged into one another and there are no well separated centers.



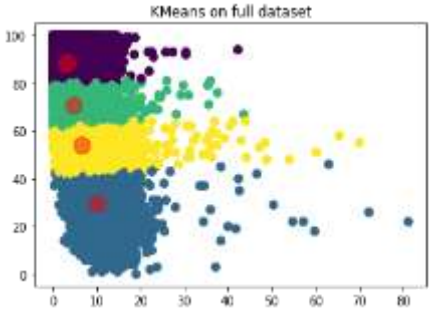
AFTER RCA



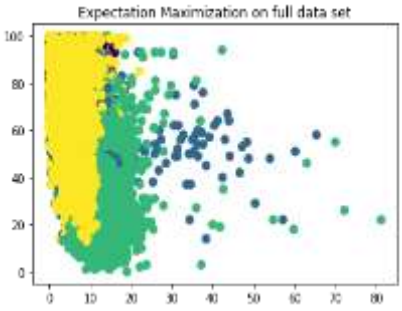
After performing RCA, the K-means graph and Expectation Maximization came out to be as shown. For K-means we used 2 clusters and for Expectation Maximization , we used 2 clusters. For both the algorithms, the clusters are well separated but the centers are very near to each other.



AFTER FEATURE SELECTION



After feature selection, we seleted 7 best features and ran K means and Expectation Maximization.Here the data is well separated for both the algorithms but the data is very widly separated. After applying feature selection, the data became more compact and well separated. In K-means we got 4 centers and in expectation maximization we got 4 clusters but the clusters are overlapping



COMPARISON

After performing the four dimensionality reduction techniques we can see that for the above dataset Prinicipal component Analysis did the best job in separating the clusters. As in this data was compact for the given clusters and centers were well separated. We can see that after applying dimensionality reduction techniques, our data got well separated . This may be because dimension reductions helps the clustering algorithms perform better and reduce the curse of dimensionality overall. But not all the techniques perform better on all the datasets , for this dataset PCA performed best.

EXPERIMENT 4 (ARTIFICIAL NEURAL NETWORK)

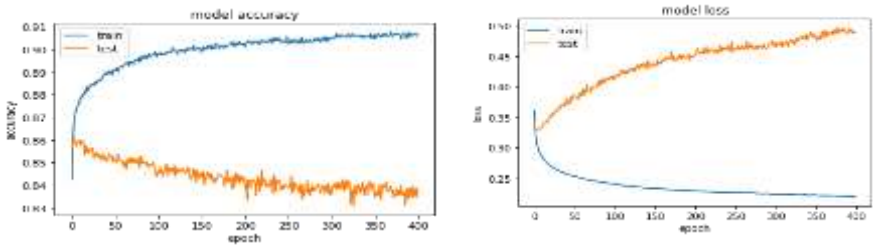
In this dataset we were getting the best accuracy for 400 epochs , relu function and with 16 epochs.The accuracy which we recived for ANN previously was 85 percent.

Confusion Matrix	Positive (1)	Negative (0)

Positive (1)	12108	1050
Negative (0)	1718	2050

We applied PCA, for the data set and we got the following results. The accuracy which we received for the same combination after applying PCA was 83 percent. There is a slight fall in accuracy for this dataset. Hence we concludes that PCA is not the apprporiate dimension reduction technique for this dataset. We can see the number of correctly classified datapoints are 12108. The model accuracy and model loss graph is shown below:

In the accuracy graph, we can see that training and testing accuracy is increasing with the increase in the epochs. Similarly in the model loss, we can see that the model loss is reducing with the number of epochs.



EXPERIMENT 5 (CLUSTERING RESULTS AS NEW FEATURES)

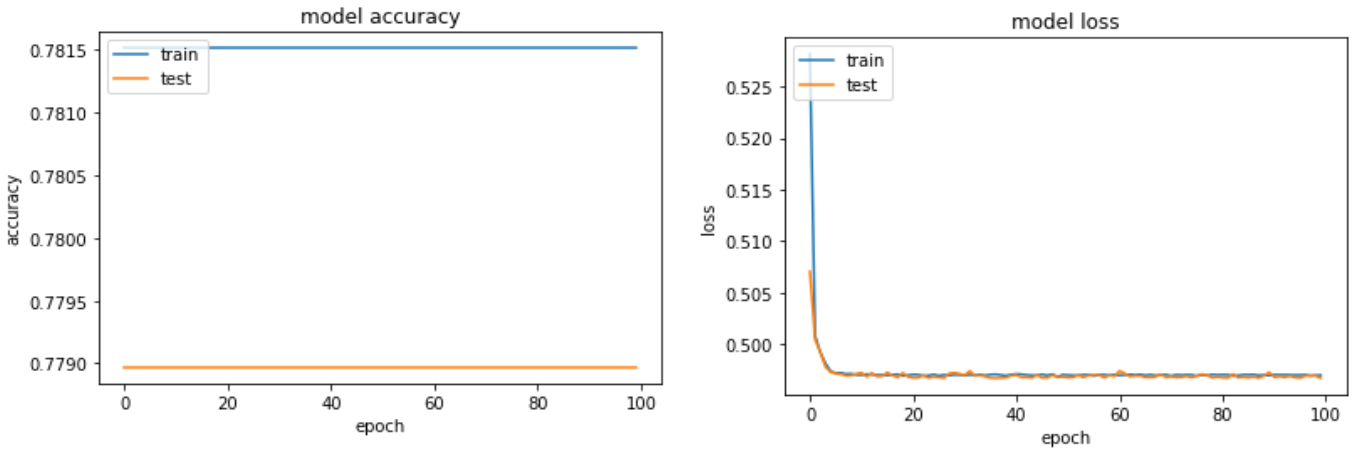
In this task, I have added a column of cluster labels to the dataset as a feature. I will be using the class labels to perform Neural Network Classification. In this we have taken the results from k- means and expectation maximization from task 1 as my x variables and original y variable as my target variable.

After applying artificial neural network algorithm to the newly created dataset and found out the accuracy. I have used the softmax activation function with 100 epochs. The confusion matrix for the above performed experiment is shown here.

From the confusion matrix here we can see that both true positives and false positives are high but the number of true positive is still greater. The overall accuracy we got for this model is 77 percent.

Confusion Matrix	Positive (1)	Negative (0)
Positive (1)	13158	0
Negative (0)	3768	0

From this we can infer that our alorithms k- means and Expectation Maximization did a good job. Here are the learning curves which we got after running ANN.



In the accuracy graph, we can see that training accuracy is constant as the number of epochs are increased where are testing accuracy is also same and is very high. Similarly in the model loss graph, we can see that the model loss remains constant throughout and is very low for both training and testing data.