**Implementation of Clustering and Dimensionality Reduction Algorithms**

**Dataset description and goal:**

The goal of the project is to:

1. Implement clustering algorithms: K-means, Expectation Maximization
2. Implement dimensionality reduction algorithms: Feature selection, PCA, ICA, Randomized Projections
3. Evaluate performance of above algorithms and find optimal parameters by plotting learning curves.
4. Evaluate performance of Artificial Neural Networks with and without clustering inputs.

Two different datasets are chosen to evaluate and compare model performances.

**Dataset 1:**

Contains information on SGEMM GPU Kernel Performance that captures the running time of a matrix-matrix product operation. The dataset information can be viewed and downloaded at UCI Machine learning repository.

<https://archive.ics.uci.edu/ml/datasets/SGEMM+GPU+kernel+performance>.

On quick examination, the following characteristics are observed from the dataset.

* There are 18 columns and 241600 rows
* There are no missing values
* The target variable for the model is average of the four GPU run times
* The independent variables are not highly correlated. Hence there no variables are dropped from the model.
* There are few outliers present, and they need to be removed.

**Dataset 2:**

Contains information on 23 different species of gilled mushrooms that are classified as edible or poisonous based on given features. This is a very interesting dataset that allows to experiment with multiple algorithms like PCA, Logistic Regression, Classification Trees, Ensembles and Neural Networks that readily translate to real-world observations. Also, the dataset has good volume, quality and does not have any aggregations.

In modern world, Plant Biologists classify the different plant species by observing distinctive features and appearances. Drug Specialists, Nutritionists and Agencies like FDA further enhance the models by adding decision variables like adult & child health, interactions of food constituents with other chemicals in body, allergic reactions, animal testing, clinical trials, etc. to predict the positive and negative effects of the plant food item before approving and releasing to the consumer market. A similar simulation of the classification algorithm on multiple predictor variables is carried out in this dataset.

The dataset information can be viewed and downloaded at Kaggle or UCI Machine learning repository.

<https://www.kaggle.com/uciml/mushroom-classification>

<https://archive.ics.uci.edu/ml/datasets/mushroom>

On quick examination, the following characteristics are observed from the dataset.

* There are 23 columns and 8124 rows
* The target variable for the model is ‘class’
* There are 2480 missing values for the ‘stalk-root’ column.
* There are more than 2 categories for most of the predictor variables.

**Assumptions and initial dataset preparation:**

The following assumptions and initial operations are made on the datasets for the implementation of algorithms.

* Cross validation is used with Artificial Neural Network and is written from scratch.
* For building Artificial Neural Network, Keras with Tensorflow backend is used.
* For Neural networks, the number of neurons in the subsequent hidden layers is assumed from the formula:

***((2/3)\*no. of input layer neurons) + no. of output layer neurons***

* Randomized projection is carried out using GaussianMixture.

**Dataset 1:**

* Columns 1-10 and 15-18 are continuous and columns 11-14 are categorical variables.
* Scaling has been performed for columns 1-10 using Standard Scaler.
* For binary classification, column 15 is converted to either 1 or 0 using its mean value as cut off.
* The number of outliers (above or below 3 standard deviations from mean) is 5251 (accounts for 2.17 % of data) and they have been removed from the given dataset.

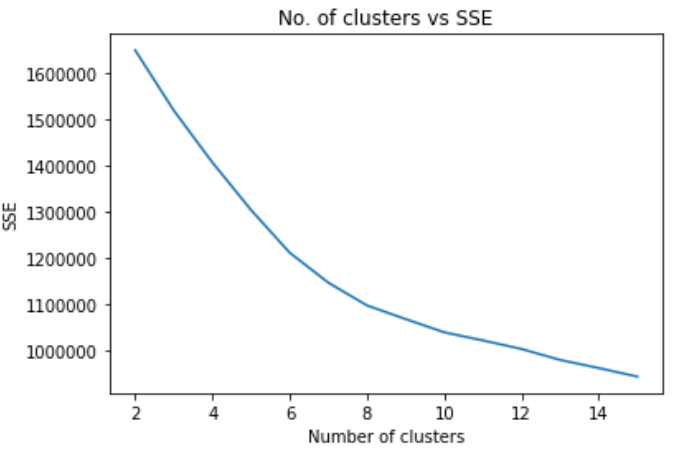
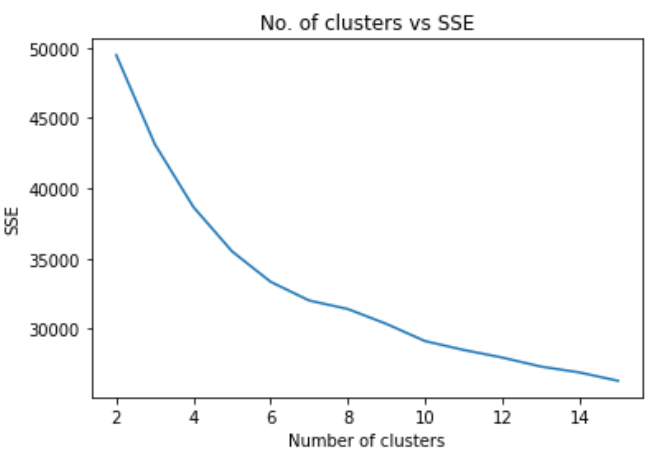
**Dataset 2:**

* For binary classification, column ‘class’ is converted to 1 if species is poisonous, and 0 if edible.
* Missing values for the ‘stalk-root’ column is imputed with the frequent value ‘stalk-root’ = ‘b’
* There are more than 2 categories for most predictors, and ordering them (1, 2, 3...) doesn't make any sense because they are not ordinal. Hence pivot is done on categories. This increases the number of columns.

**Task 1:**

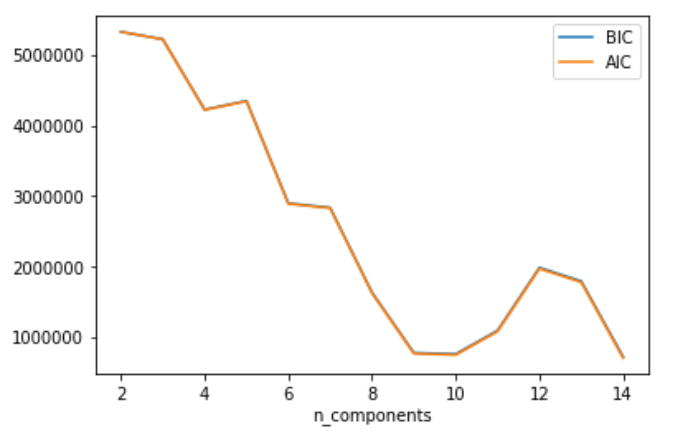
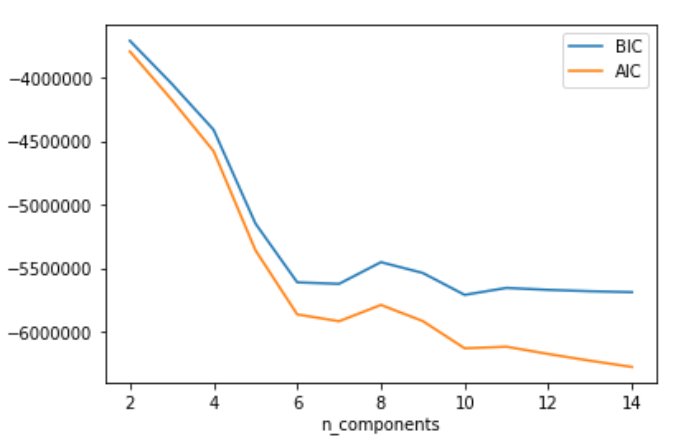
The dataset is scaled and partitioned into train set with 70% data, and test set with 30% data. K means clustering and Expectation Maximization algorithms are run on the dataset. For k means, SSE is plotted against number of clusters. For EM, AIC, BIC and Log likelihood values are plotted.

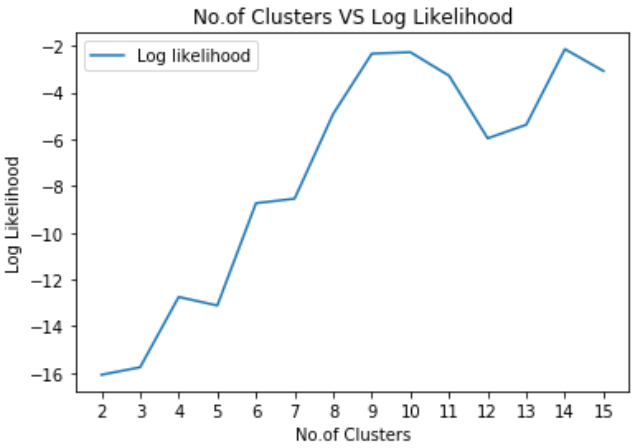
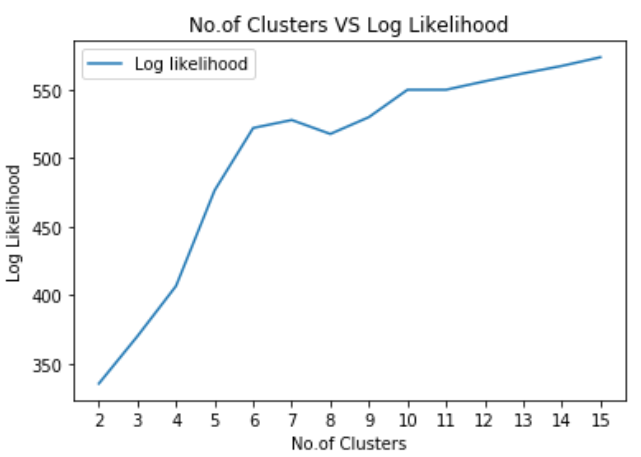
***K means - Sgemm Product* *K means – Mushroom***

From the elbow plots, it is observed that number of clusters N=8 is best for sgemm product dataset, and N=6 is better for Mushroom dataset.

***EM - Sgemm Product* *EM – Mushroom***

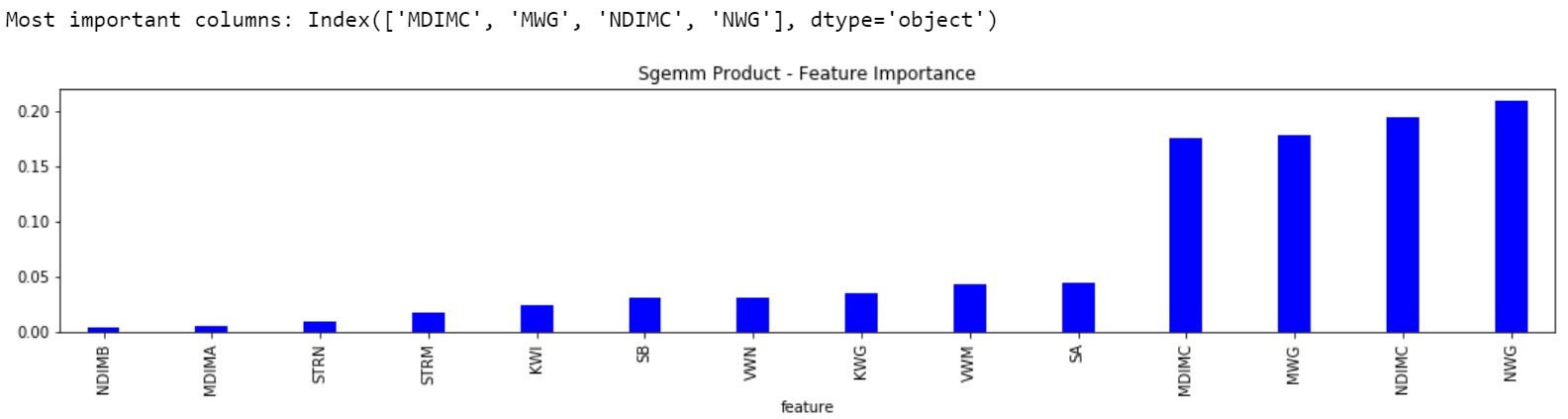
For AIC and BIC, we consider the least values and for log likelihood we consider the maximum value for best performance. Among AIC and BIC, we give more preference to BIC because it penalizes the model more in adding redundant variables. From the above plots, we can see that when No. of features N=9, we get best performance for sgemm product dataset, and N = 10 gives best performance for Mushroom dataset.

**Task 2:**

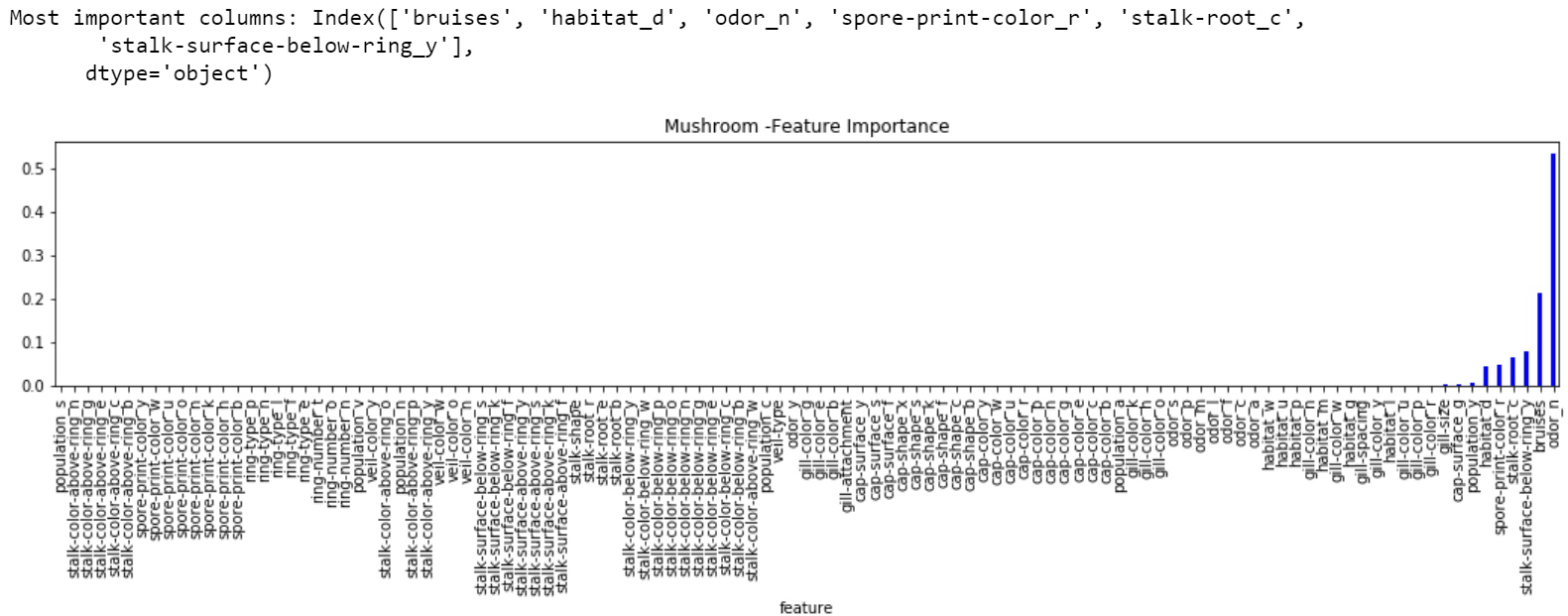
The below dimensionality reduction algorithms are run:

**Feature Selection:** Decision Tree and Recursive Feature Elimination (RFE) algorithms are run to find the best possible features for this model. It was found that both algorithms gave almost similar features.

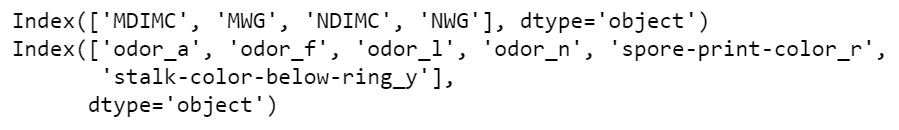
***Decision Tree - Sgemm Product***



***Decision Tree - Mushroom***



***Recursive Feature Selection:***

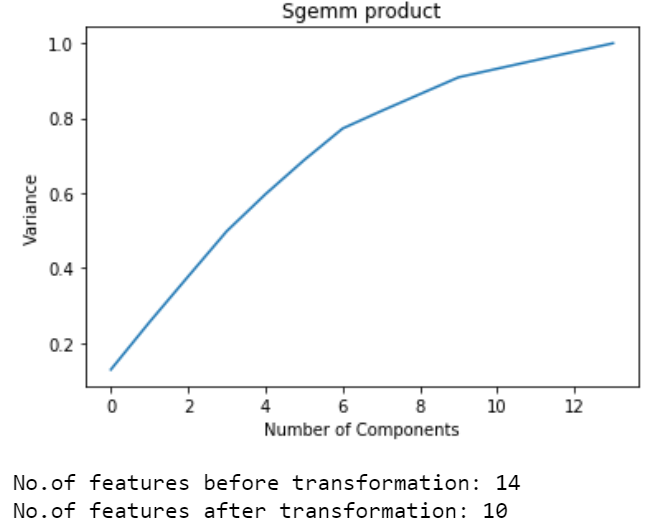
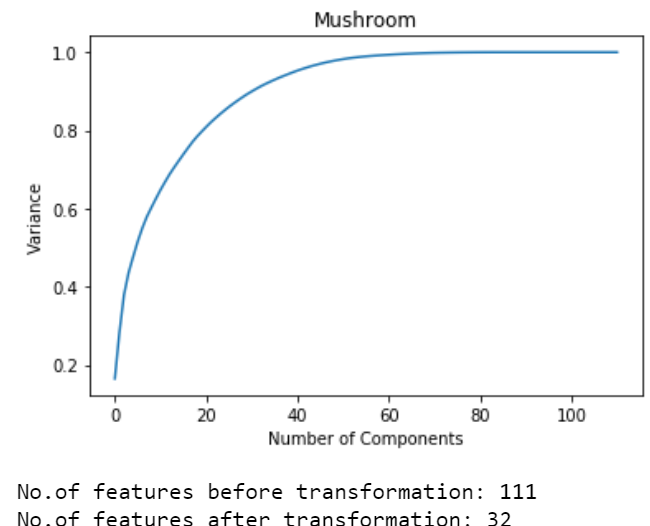


It is observed the most important features from Decision tree feature selection algorithm for sgemm product dataset are – MDIMC, MWG, NDIMC, NWG. We get the same features from recursive feature selection.

For mushroom dataset, the most important features from Decision tree feature selection algorithm are – Bruises, habitat\_d, odor\_n, spore-print-color\_r, stalk-root\_c, stalk-surface-below-ring\_y. From recursive feature selection, we get odor\_a, odor\_f, odor\_l, odor\_n, spore-print-color\_r and stalk-color-below-ring\_y.

**PCA:** Principal Component Analysis was run to find the best possible features for this model. A plot of Variance vs number of components was plotted, and the features contributing to most variance are selected.

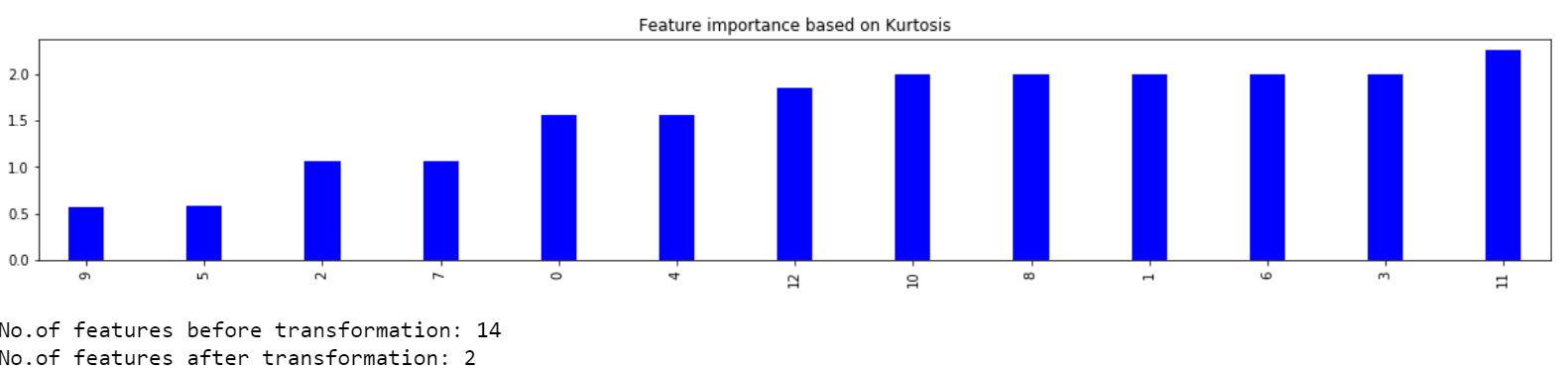
***Sgemm Product* *Mushroom***

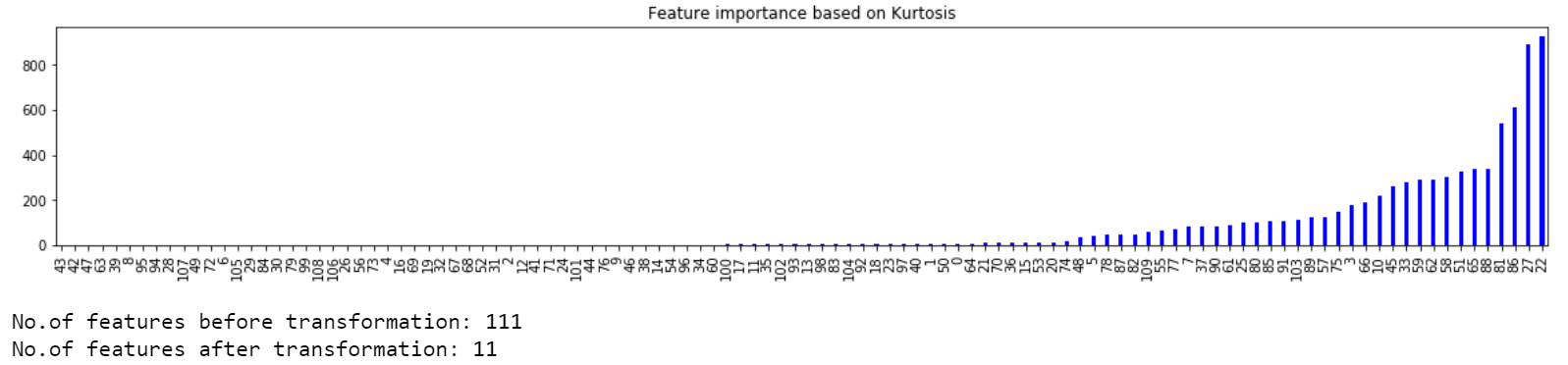
It is seen that for sgemm product dataset, after dimension reduction, PCA returned 10 components. For mushroom dataset, after dimension reduction, PCA returned 32 components

**ICA:** Independent Component Analysis was run to find the best possible features for this model based on kurtosis.

***Sgemm Product***

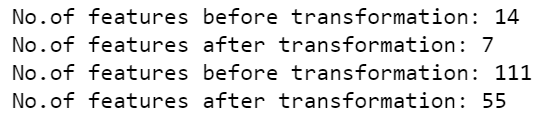


***Mushroom***



It is seen that for sgemm product dataset, after dimension reduction, ICA returned 2 components. For mushroom dataset, after dimension reduction, ICA returned 11 components.

**Randomized Projections:** Randomized Projections was run using GaussianMixture to find the best possible features for this model. For sgemm product dataset, after dimension reduction, RCA returned 7 components. For mushroom dataset, after dimension reduction, RCA returned 55 components.

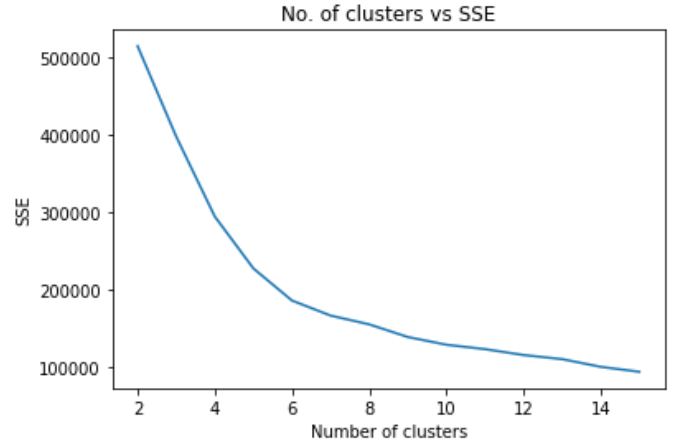
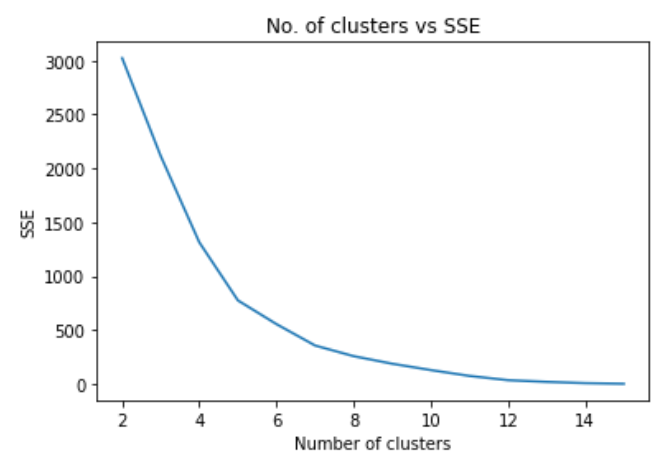


**Task 3:**

Clustering algorithms are run again after dimensionality reduction with reduced features and the performance is found.

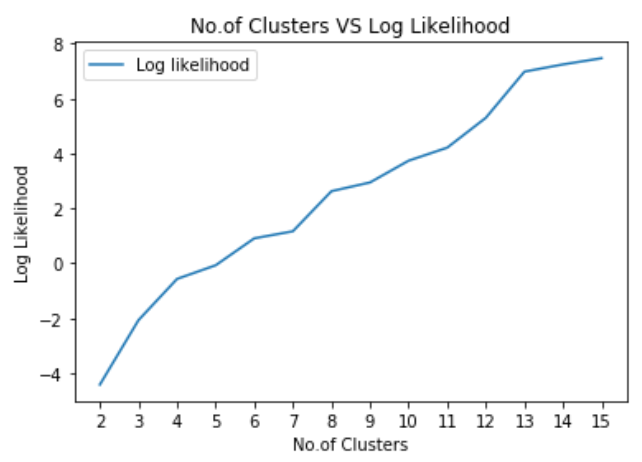
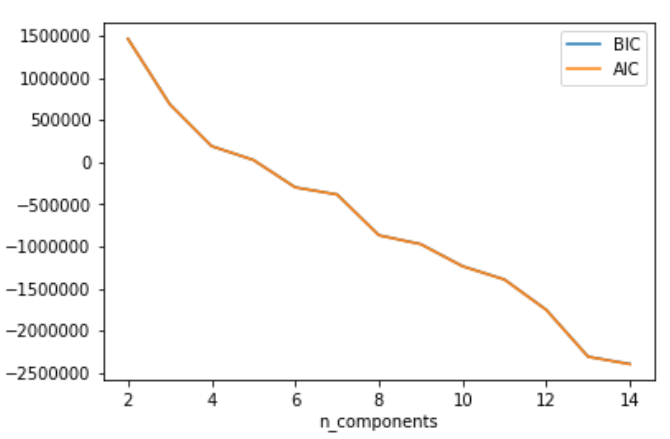
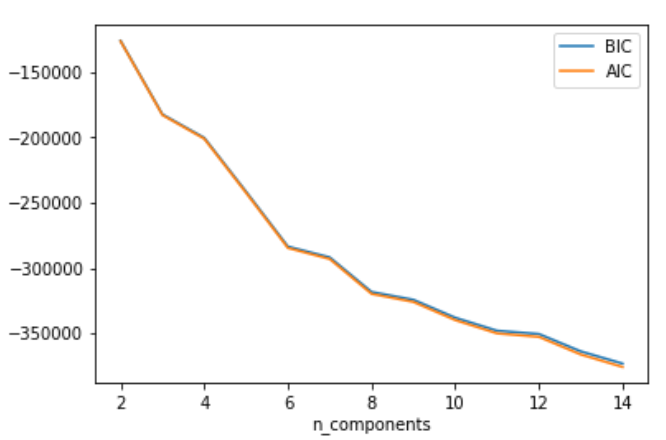
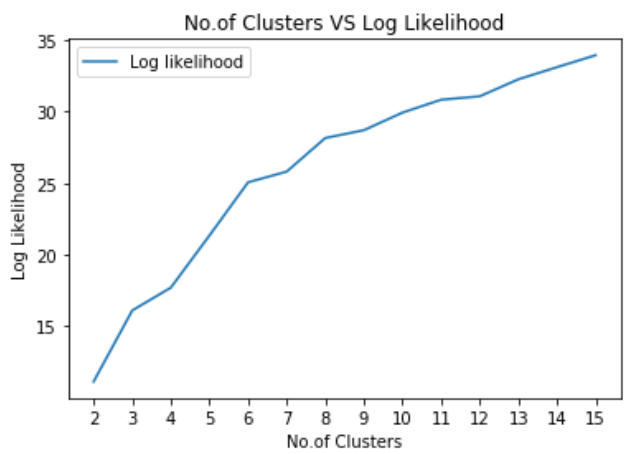
**Feature Selection – Decision Tree**

***K means - Sgemm Product* *K means -* *Mushroom***

Optimal clusters = 6 Optimal clusters = 5

***EM - Sgemm Product* *EM –* *Mushroom***

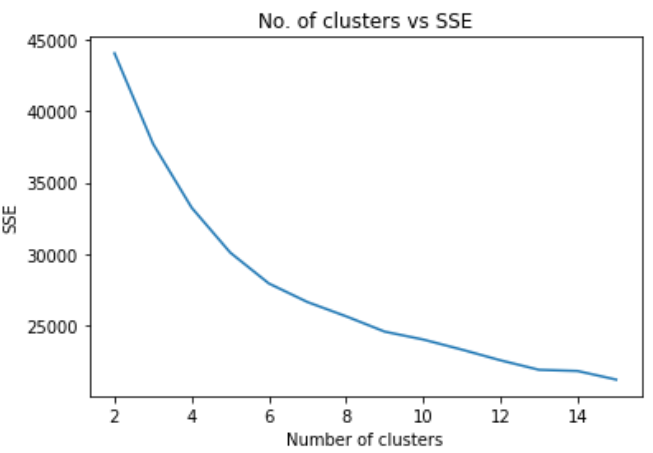
  

From the above plots, we can see that when No. of features N=13, we get best performance for sgemm product dataset, and N = 8 gives best performance for Mushroom dataset.

On running the clustering algorithms after dimension reduction, it is expected that the clusters are far more compact because the data points are projected on a lower dimension and the points come closer to each other. When the dimension is increased, it will be the opposite where distance between data points increases and the intra cluster distance increases, making elongated clusters.

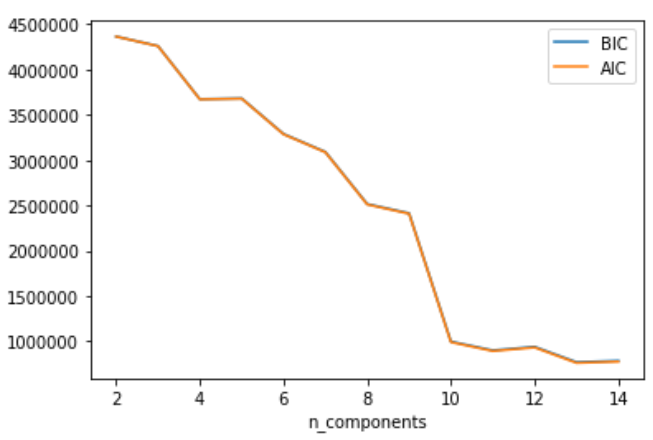
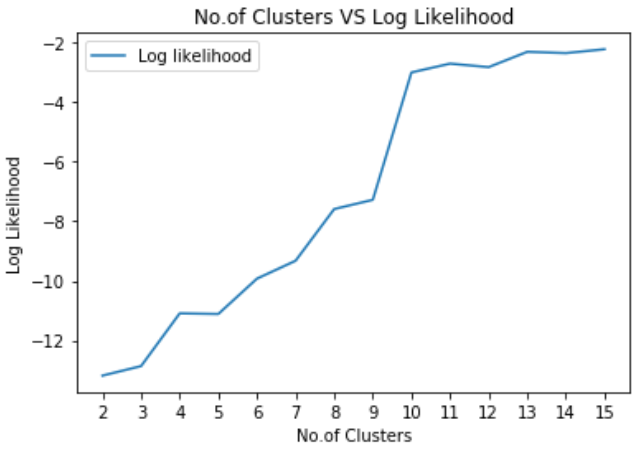
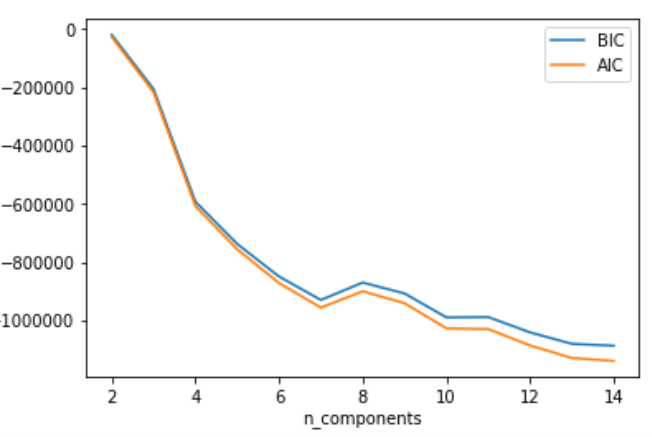
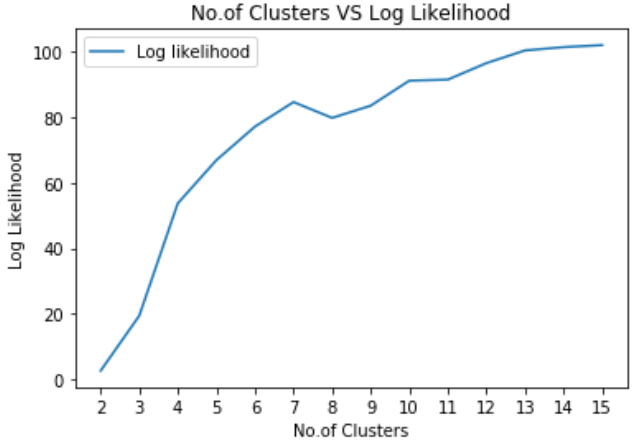
**PCA**

***K means - Sgemm Product* *K means –* *Mushroom***

Optimal clusters = 7 Optimal clusters = 6

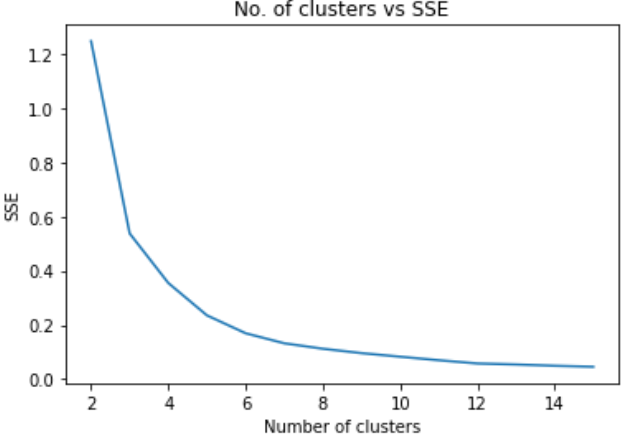
***EM - Sgemm Product* *EM –* *Mushroom***

From the above plots, we can see that when No. of features N=10, we get best performance for sgemm product dataset, and N = 7 gives best performance for Mushroom dataset.

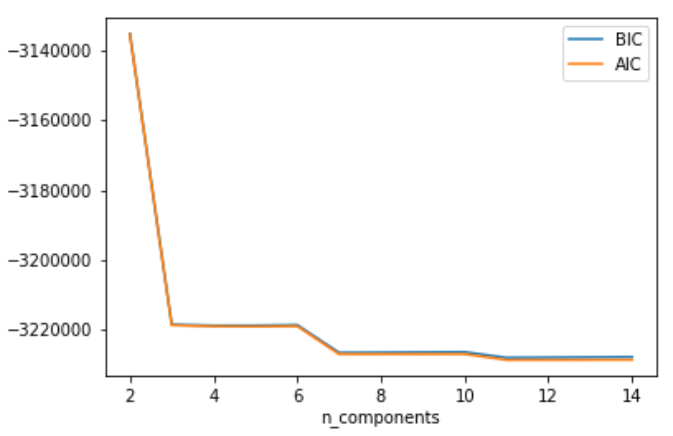
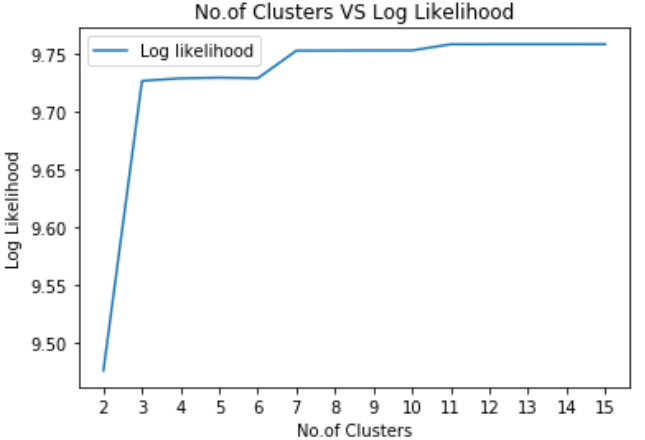
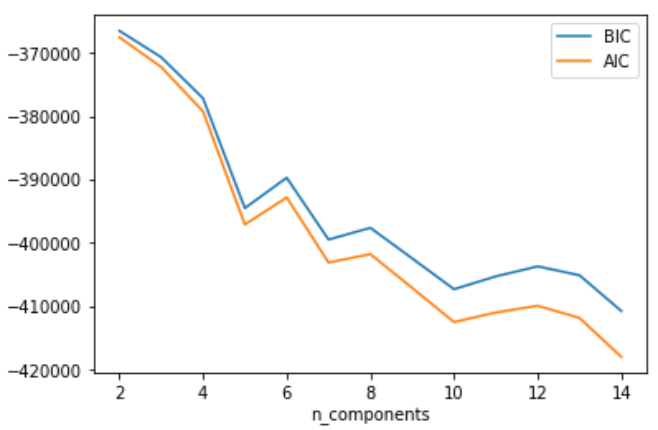
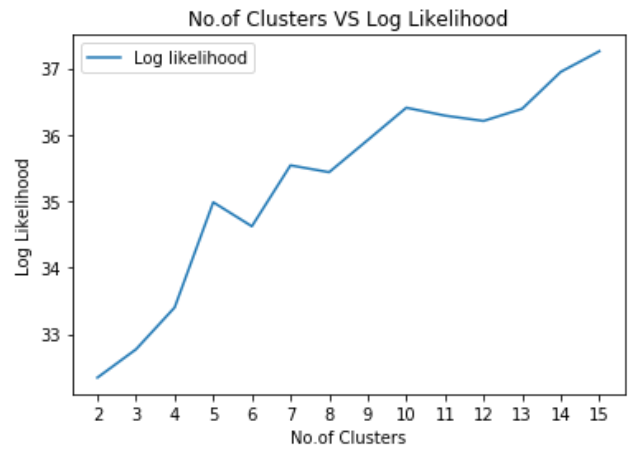
**ICA**

***K means - Sgemm Product* *K means –* *Mushroom***

Optimal clusters = 3 Optimal clusters = 5

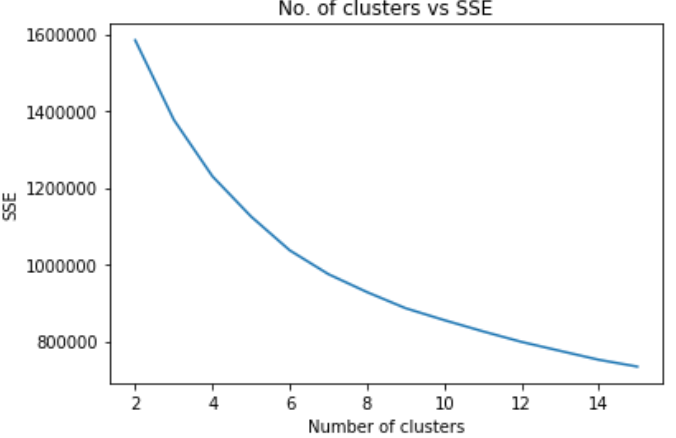
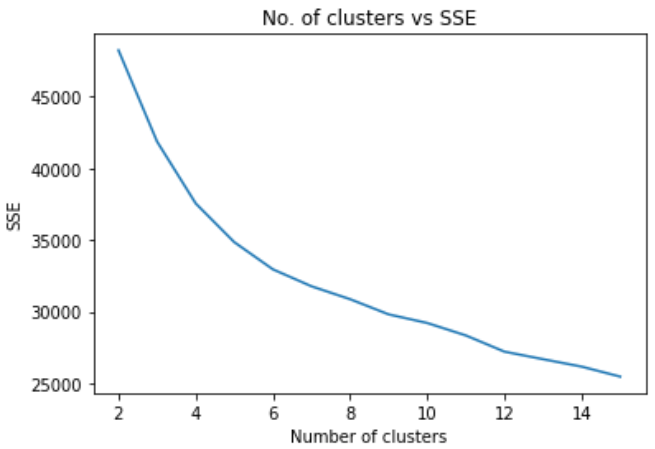
***EM - Sgemm Product* *EM –* *Mushroom***

From the above plots, we can see that when No. of features N=3, we get best performance for sgemm product dataset, and N = 5 gives best performance for Mushroom dataset.

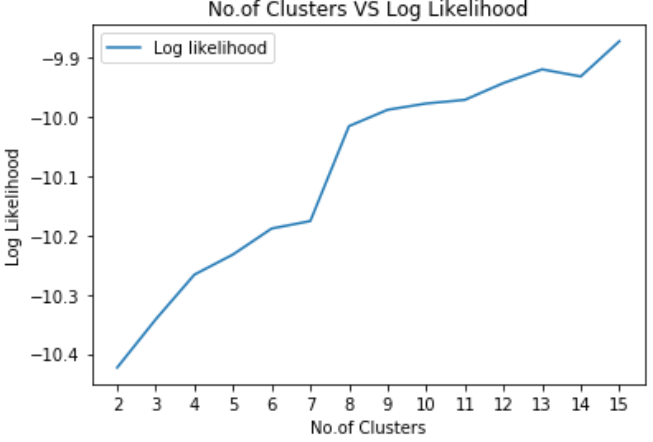
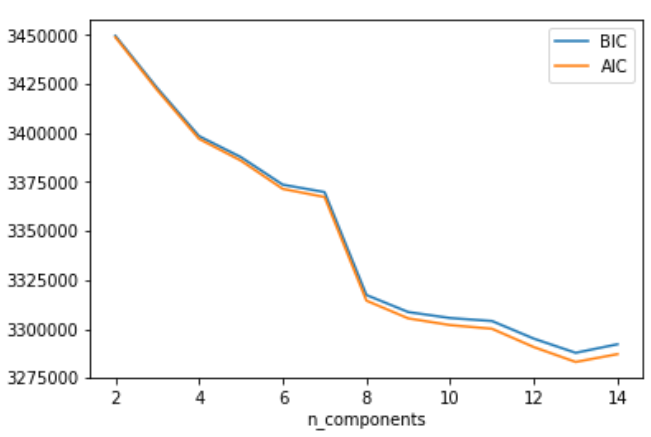
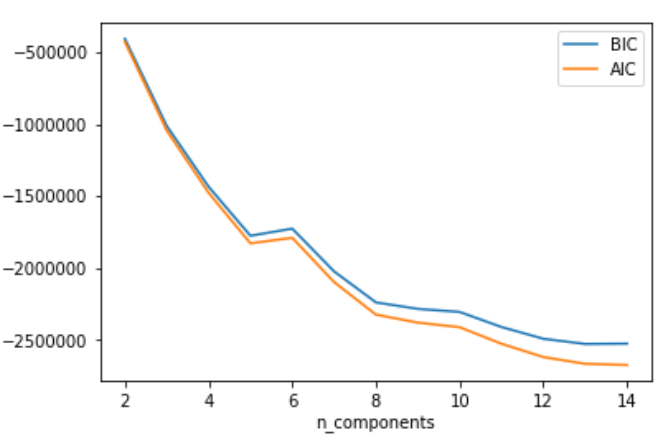
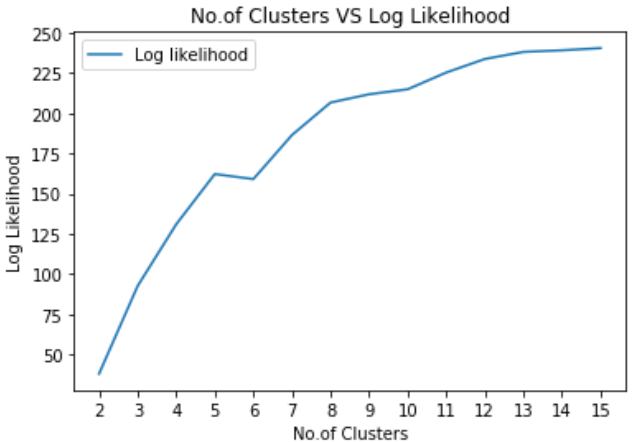
**Randomized Projections**

***K means - Sgemm Product* *K means -* *Mushroom***

Optimal clusters = 6 Optimal clusters = 5

***EM - Sgemm Product* *EM -* *Mushroom***

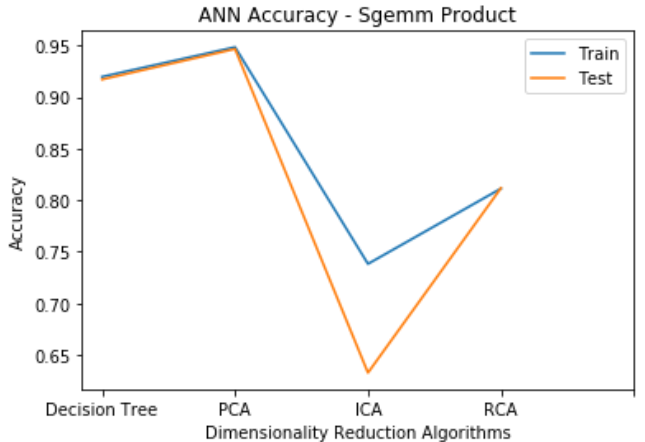
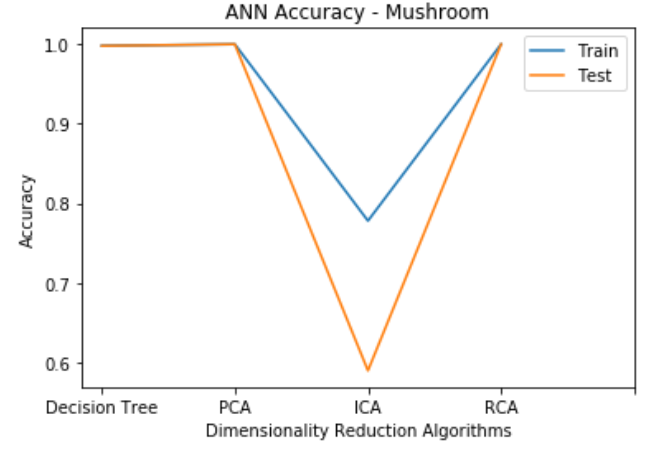
  

From the above plots, we can see that when No. of features N=8, we get best performance for sgemm product dataset, and N = 5 gives best performance for Mushroom dataset.

**Task 4:**

ANN algorithm is run after dimensionality reduction with reduced features and the performance is evaluated.

***ANN - Sgemm Product* *ANN -* *Mushroom***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Sgemm Product Dataset** | | **Mushroom Dataset** | |
| **Train accuracy** | **Test accuracy** | **Train accuracy** | **Test accuracy** |
| **Feature Selection** | 0.9197 | 0.9171 | 0.9980 | 0.9979 |
| **PCA** | 0.9484 | 0.9466 | 1.0 | 1.0 |
| **ICA** | 0.7383 | 0.6327 | 0.7778 | 0.5894 |
| **Randomized Projections** | 0.8113 | 0.8117 | 1.0 | 1.0 |

From the above plots, we can see that for sgemm product dataset, PCA gives best accuracy and ICA gives the least accuracy. For Mushroom dataset, Decision Tree feature selection, PCA and RCA give best performance, but ICA gives least accuracy.

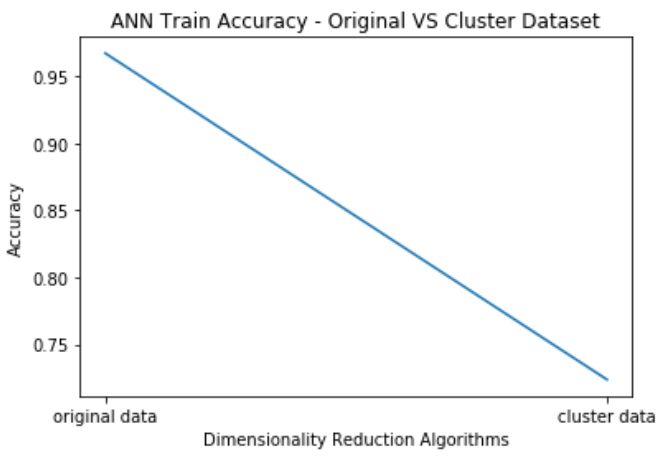
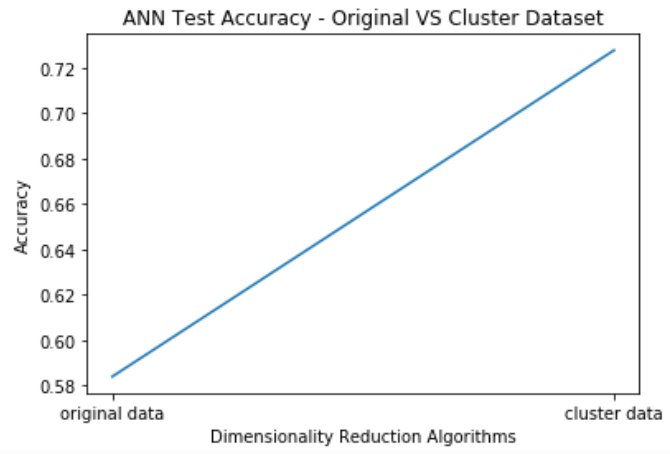
Hence PCA, RCA and Decision Tree Feature selection are not much affected from curse of dimensionality, but ICA suffers. Also, when evaluating the number of clusters after dimension reduction, each algorithm almost always gives same number of clusters.

**Task 5:**

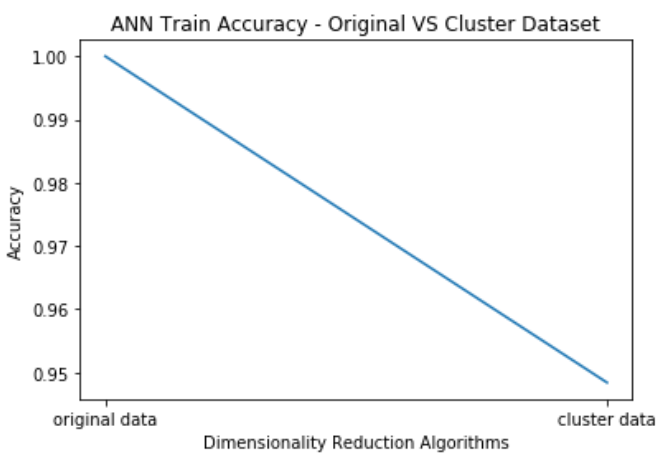
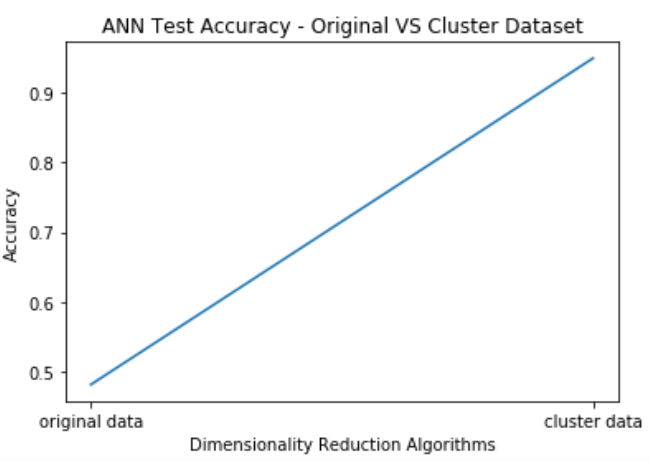
ANN algorithm is run using inputs from the cluster results from task 1. From task 1, we found that for K means, number of clusters N=8 is best for sgemm product dataset, and N=6 is better for Mushroom dataset. For EM, when No. of features N=9, we get best performance for sgemm product dataset, and N = 10 gives best performance for Mushroom dataset.

Now K means is run for K=8 and K=6. EM is run for N=9 and N=10. The output clusters and probabilities are fed for the ANN model. This new performance is compared against the performance of the original model and graphs are plotted.

***Sgemm Product***

***Mushroom***

From the above plots, we see that cluster input slightly decreases the performance of the train dataset, but significantly increases the performance of the test dataset. Hence, clustering helps the model to generalize well. Best k value gives a perfect tradeoff between the train accuracy and test accuracy.

We now conclude that for sgemm product dataset, number of clusters with k-means = 8 and EM = 9 is optimal. 10 columns are created, leaving 1 for k-means, since it is the hard cluster and rest are probability values. For Mushroom dataset, optimal number of clusters with k-means = 6 and EM = 10. 11 columns are created, leaving 1 for k-means, and rest are probability values.