Ramin Mehdizad Tekiyeh Date : March 7th, 2020

Clustering of GPU Speed Dataset

Introduction

In this project, the objective is to analyze capabilities of Clustering techniques on GPU dataset.

About Datasets

This dataset consists of 18 features and 241600 records. The first 14 features demonstrate specific parameters of a Graphic Processing Unit (GPU) recorded as integer numbers and the last 4 columns include run time of 4 specific program code in seconds on the unit with the mentioned specification. The objective is to cluster the units in two distinct clusters as high speed and low speed by choosing criterion of average of runtime values.

Preprocessing

Following steps or preprocessing are done on dataset:

- Four runtime features of each data record ([Run1]…[Run4]) are substituted with the average value in each record as the GPU speed under the feature name of [Runtime].
- Average value of [Runtime] feature is chosen as the criterion for classifying the records as low and high speed with labels 0 and 1.
- All the data is normalized using StandardScaler().

Algorithm Implementation

Clustering methods of KMeans and Expectation Maximization is implemented on both original dataset and reduced dimension dataset. Therefore, we have four dataset: original dataset, PCA reduced dataset, ICA reduced dataset and RP reduced dataset. The two clustering methods are applied on these datasets.

Project Outline

The report is outlined in 3 parts:

- Part 1: Feature Importance
- Part 2: KMeans Clustering
- Part 3: EM Clustering
- Part 4: Clustering Quality
- Part 5: Conclusion

Part 1: Feature Importance

In the first step, the importance of 14 features of dataset is analyzed by using decision tree technique. Moreover, feature importance of reduced dataset with 8 features is investigated. The following plots depict the results.

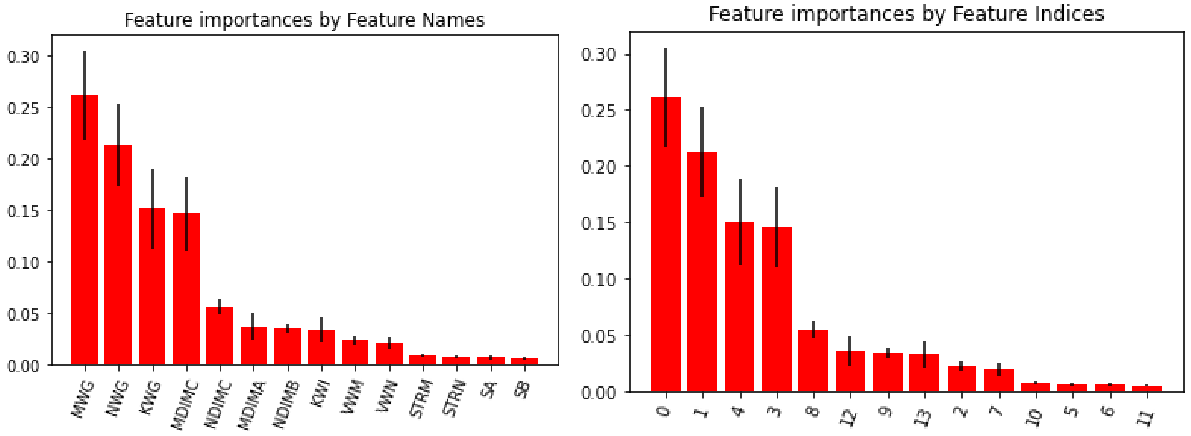

Fig. 9: Feature importance of original data plotted by their names and feature numbers.

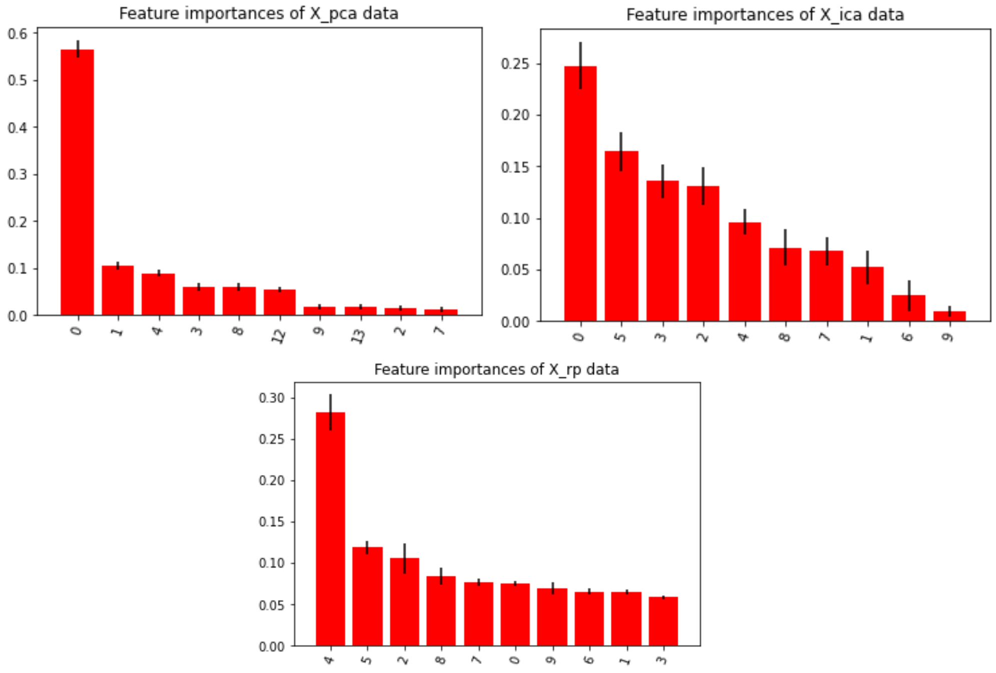

Fig. 9: Feature importance of dimensionally reduced dataset with 8 features for PCA, ICA and RP technique.

Results:

- Features MWG and NWG of the original dataset own the most importance among all 14 features.

- In dimensionally reduced datasets by PCA, ICA and RP, different features have the most importance because in each of them, data point are projected on different axes.

Part 2: KMeans Clustering

In this section, KMeans clustering is used on original and reduced datasets of 4 features. Following plots illustrate the clustered data. To illustrate the clusters, 2D and 3D plots of the 2 and 3 most important features are presented.

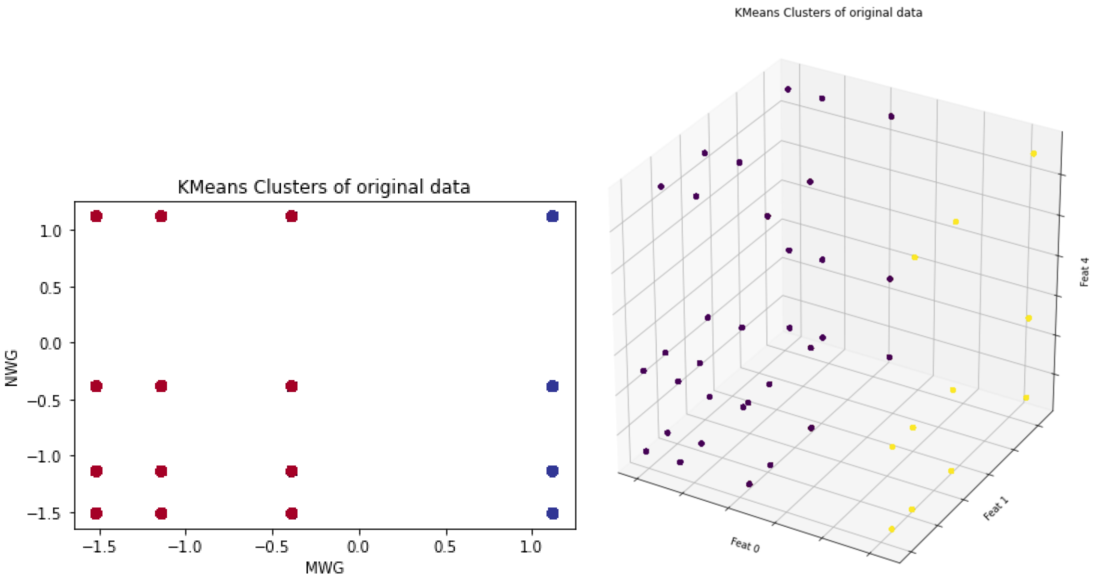

Fig. 9: Clusters by KMeans on original dataset.

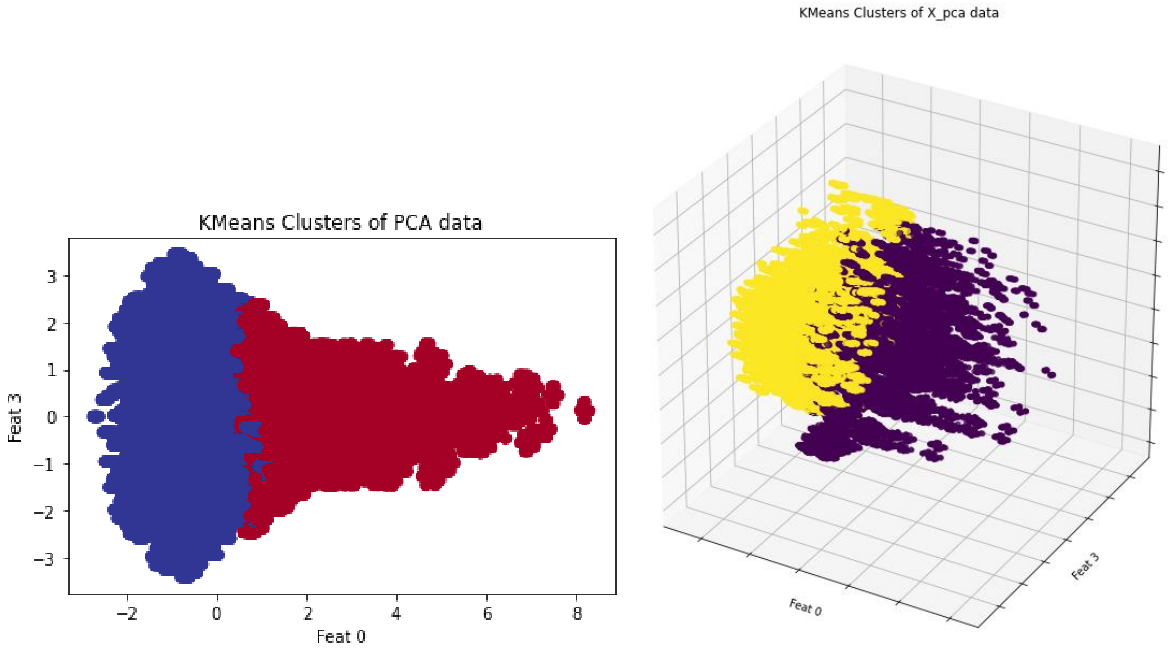

Fig. 9: Clusters by KMeans on PCA reduced dataset.

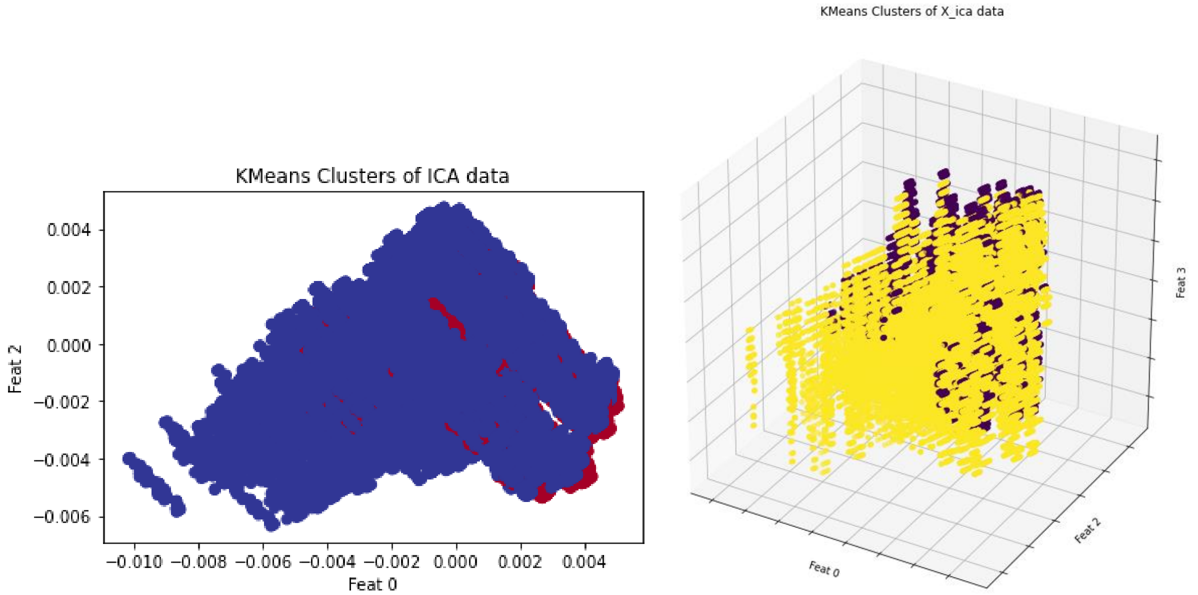

Fig. 9: Clusters by KMeans on ICA reduced dataset.

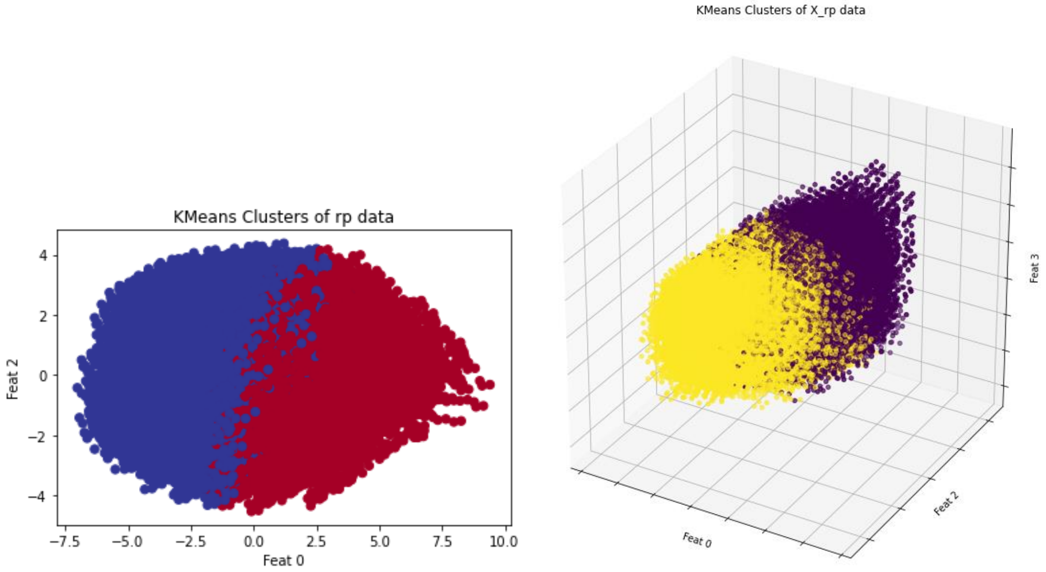

Fig. 9: Clusters by KMeans on RP reduced dataset.

Results:

The following results are observed:

- ICA gives the worst outcome.
- PCA works better than RP dimensionality reduction on our dataset.

Part 3: EM Clustering

In this section, Expectation Maximization (EM) clustering is used on original and reduced datasets of 4 features. Following plots illustrate the clustered data. To illustrate the clusters, 2D and 3D plots of the 2 and 3 most important features are presented.

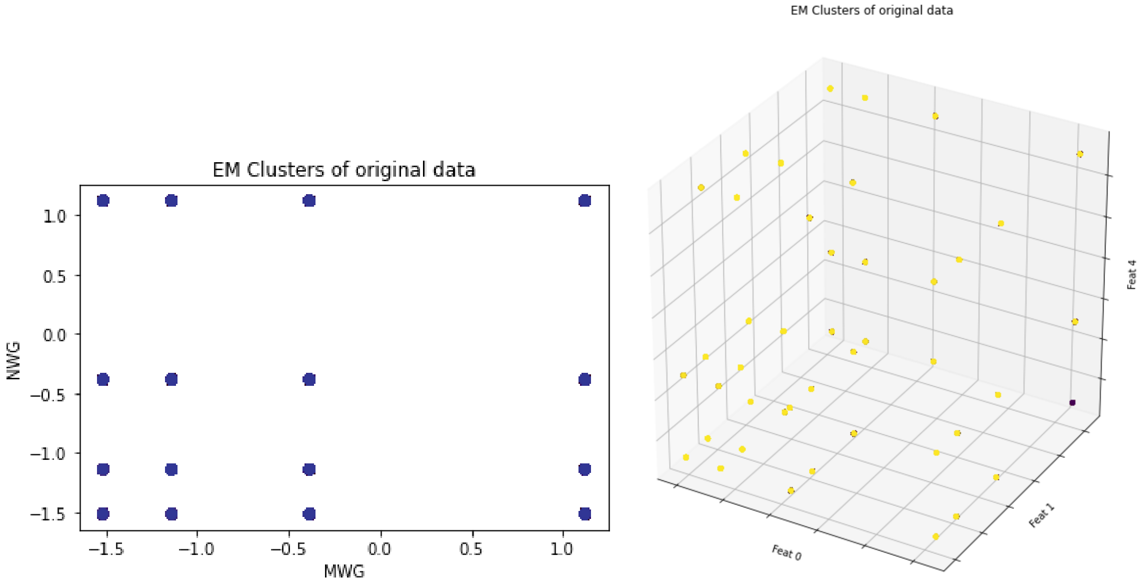

Fig. 9: Clusters by EM on original dataset.

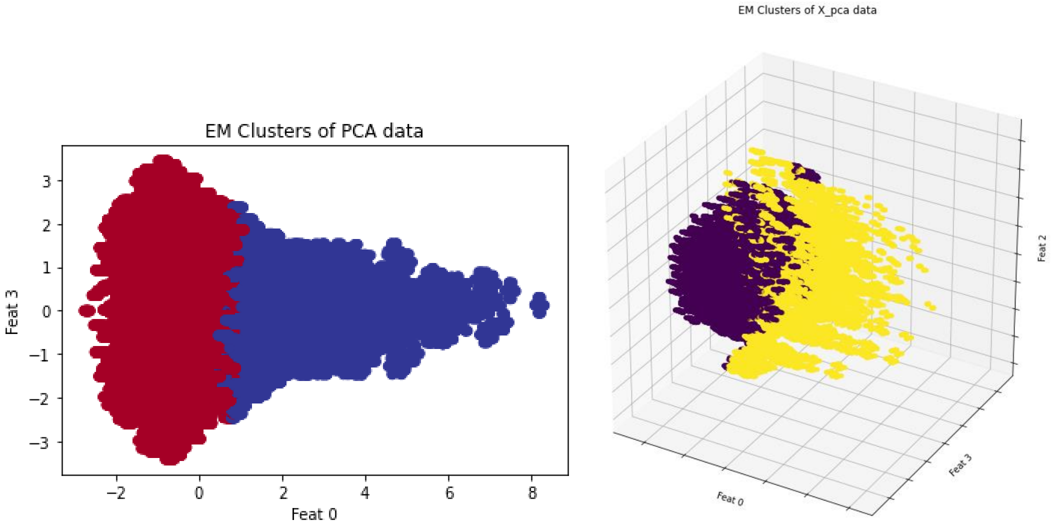

Fig. 9: Clusters by EM on PCA dimensionality reduced dataset.

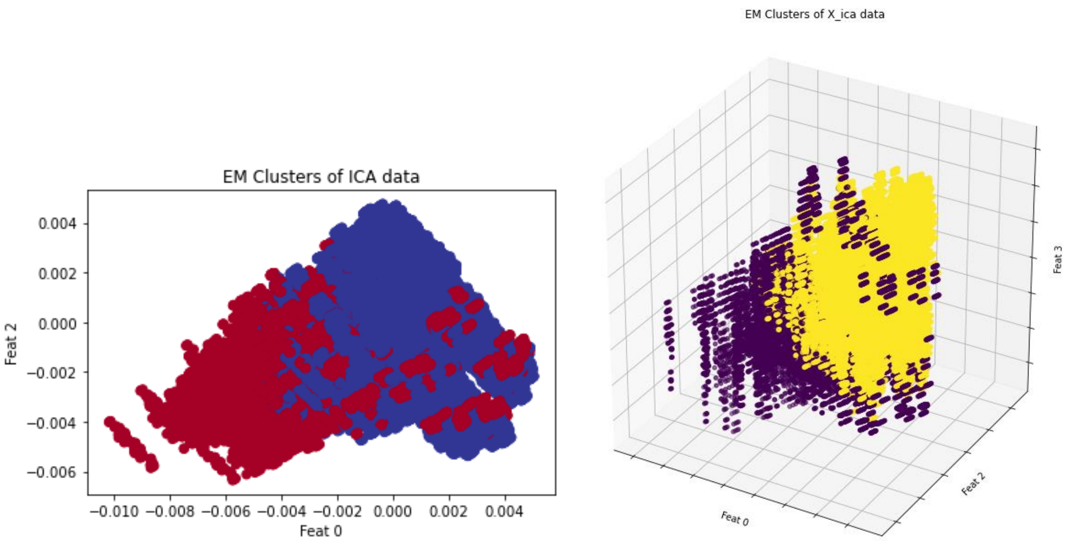

Fig. 9: Clusters by EM on ICA dimensionality reduced dataset.

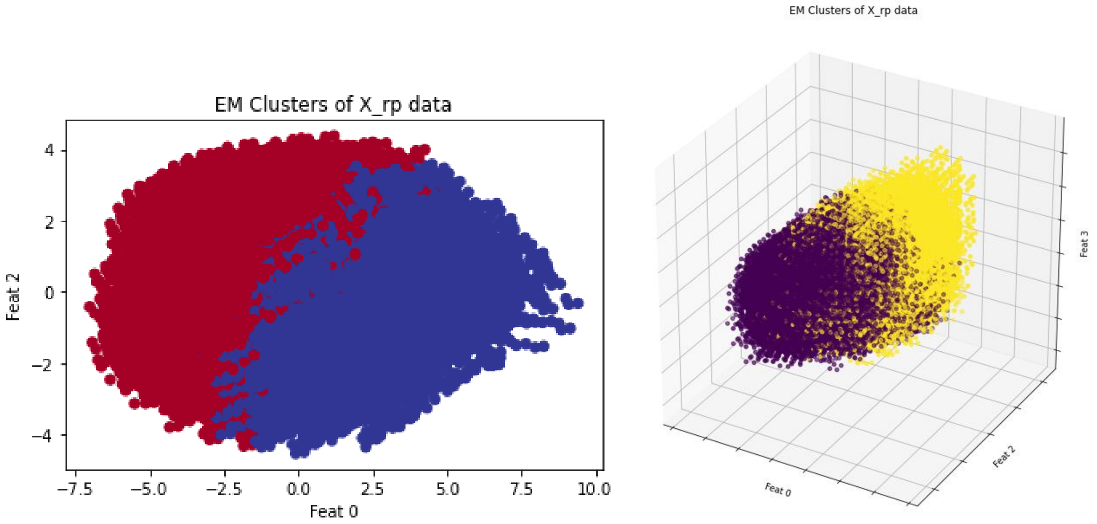

Fig. 9: Clusters by EM on RP dimensionality reduced dataset.

Results:

Just as in KMeans:

- ICA gives the worst outcome.
- PCA works better than RP dimensionality reduction on our dataset.

Part 3: Clustering Quality

To analyze the quality of clustering models, the number of dimensionality reduced dataset is varied from 4 to 14 features by PCA, ICA and RP method. Then KMeans and EM are applied on them and accuracy of the predictions is calculated. Also, SSE for KMeans and MSE for EM are calculated as the indication of quality of clustering. Following plots show the clustering quality of different models.

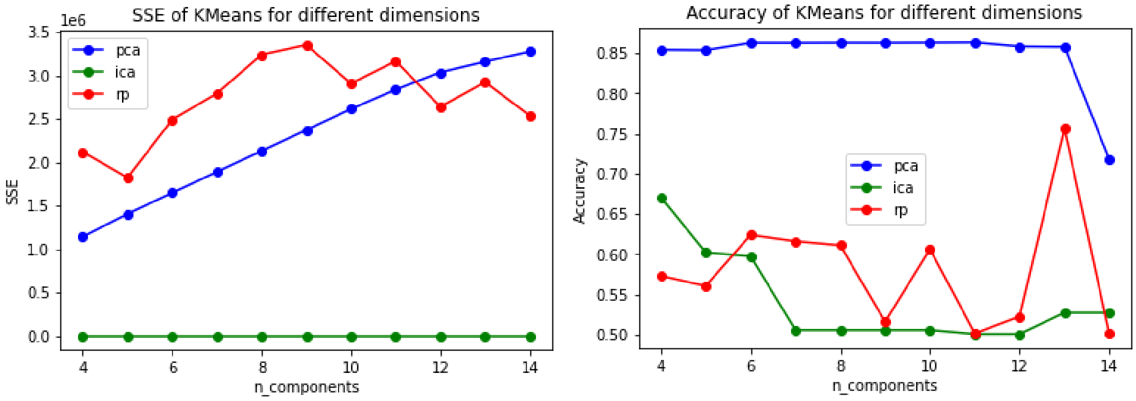

Fig. 9: Accuracy and SSE of KMeans models with different number of features.

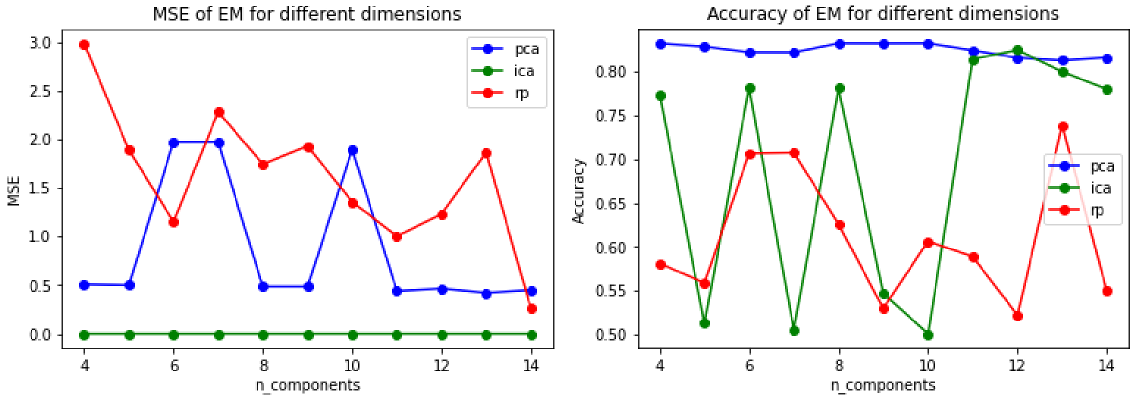

Fig. 9: Accuracy and MSE of EM models with different number of features.

Part 5: Conclusion

By analyzing the results of KMeans and EM clustering models, the following conclusions can be made.

- ✓ Accuracy of KMeans model on dimensionally reduced dataset of 8 features by PCA method is the highest on our dataset.
- ✓ By making a compromise between SSE and Accuracy, it is observed that KMeans model on dimensionally reduced dataset of 4 features by PCA method is the most effective model for GPU runtime dataset.
- ✓ ICA yields the worst outcome on both KMeans and EM models.