Jeffrey Chang
jlc190004

BUAN 6341 APPLIED MACHINE LEARNING ASSIGNMENT 4

The purpose of this assignment is to implement three learning algorithms, K-Means, Expectation-Maximization, and neural network algorithms, on the two datasets, "Seoul Bike-sharing Demand Data" and "IBM HR-Employee-Attrition & Performance. Besides, I will implement four feature dimensionality reduction methods on each learning algorithms. This report details the various experiments conducted using the two datasets to understand the effect of feature selection and dimensionality reduction algorithms. It also shows the result of dimensionality reduction on Artificial Neural Network. Both datasets will use training set (70% of data) and test set (30% of data) if need.

1. Seoul Bike-Sharing Demand Dataset

1.1 Dataset Explore

(Table 1. Represented Data table of the Seoul Bike-Sharing Demand Dataset)

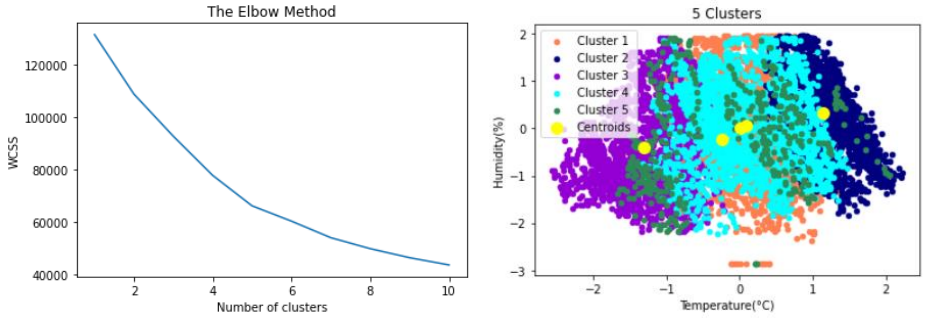
	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes

By reviewing the dataset, three variables are not appropriate to use to cluster or classify the dataset. And on this step, it is the first feature selection by using the understanding of the dataset.

1. Ignored the "Date" column since the model is a 24 hours model with hourly prediction and not considering a specific year and date.
2. Ignored the "Rented Bike Count" column since the "Rented Bike Count" information is positively correlated with the output "Function Day." The bike renting status will be "with bike renting" for count value other than zero; otherwise, the renting status will be "no bike renting." No further information is needed if considering the "Rented Bike Count" as an input attribute. After modification, I encoded the dataset by OneHotEncoder Method. The following is the shape of the dataset: [X (8760, 15), Y (9760)]

1.2 K-Means

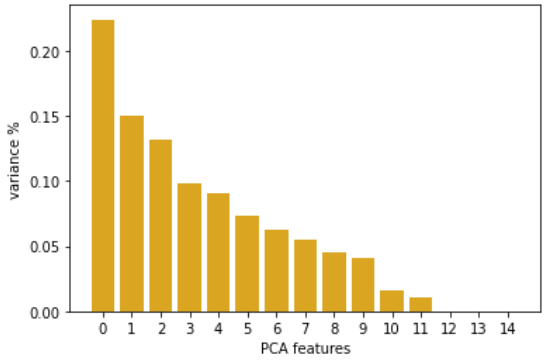
The Elbow method is used to determine the optimal number of clusters. In this method, we run the K-Means clustering on the dataset for a range of values of k and calculate the WCSS (within-cluster sum of squares) for each value of k. Our goal is to choose a small value of k with a low WCSS and mark the 'elbow' of an arm. The optimum value of k is 5.

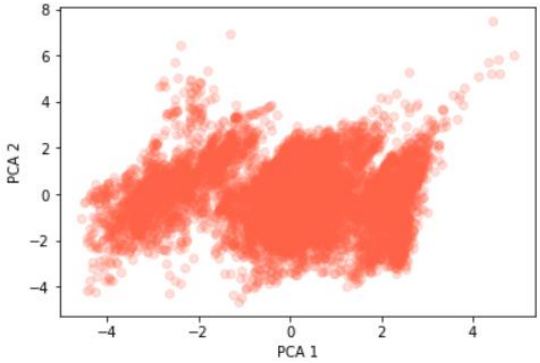


1.2.1 Feature Selection & Dimension reduction

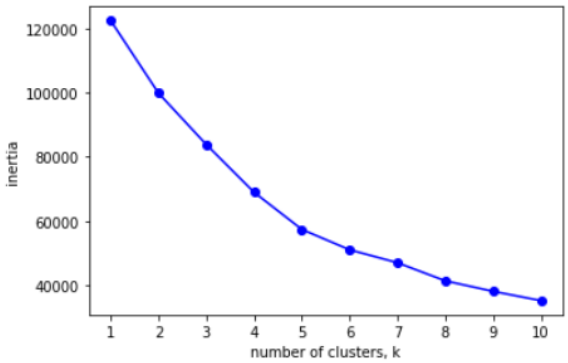
- **PCA**

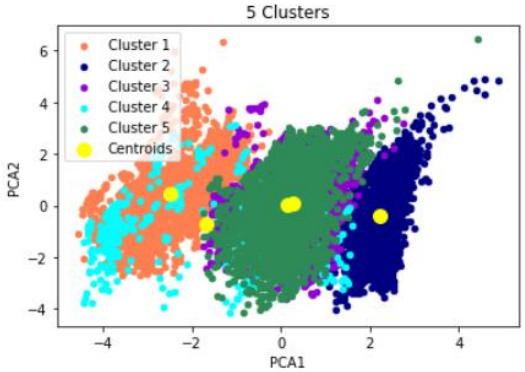
The PCA was applied to the modified dataset of Seoul Bike-sharing Demand Data, and the variance ratio was plotted. The first ten components contribute to more than 90% variance. So, it is an excellent reason to drop the last five PCA. With these first 2 components, we can jump to one of the most critical PCA applications, data visualization.





The Elbow method was used to determine the optimal number of clusters. As from the figure below, the optimum value of k is 5.



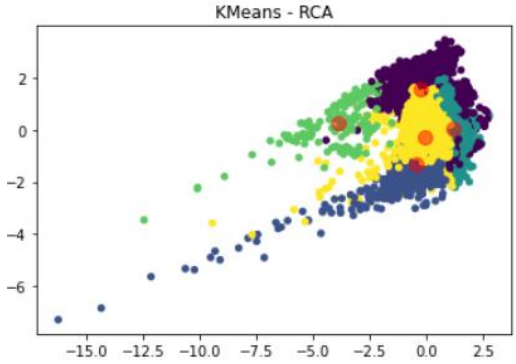


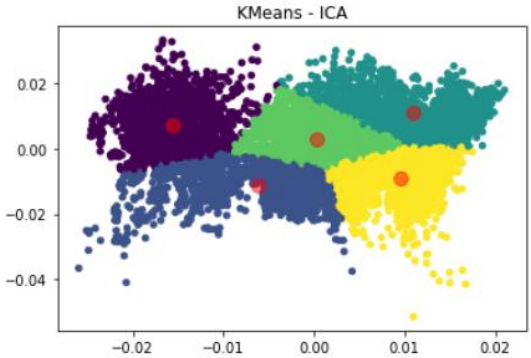
- **Randomized Projections**

Try various dimensionality of the target projection space; when components are equal to 12, the Silhouette Score is the best. The cluster number is 5.

- **ICA**

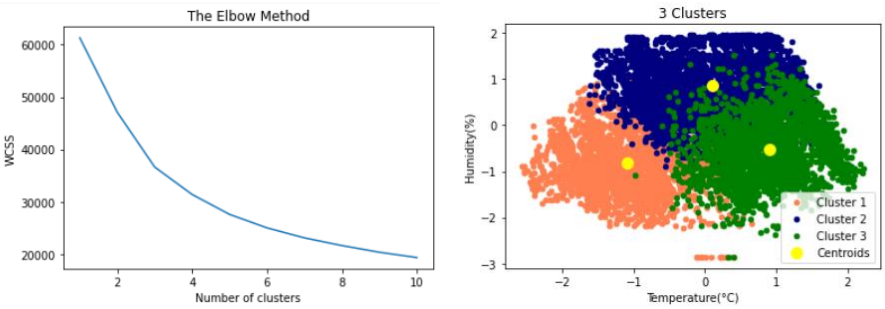
Applied various components; when components are equal to 2, the Silhouette Score is the best. The cluster number is 5.





- **Feature Selection using Random Forest**

Random forests consist of 4 –12 hundred decision trees, each of them built over a random extraction of the observations from the dataset and a random extraction of the features. Each tree is also a sequence of yes-no questions based on a single or combination of features. At each node (this is at each question), the three divides the dataset into two buckets, each of them hosting observations that are more similar among themselves and different from the ones in the other bucket. Therefore, each feature's importance is derived from how "pure" each of the buckets is. After performing feature selection on the original dataset, following features were obtained: *'Hour', 'Temperature(°C)', 'Humidity(%)', 'Wind speed (m/s)', 'Visibility (10m)', 'Dew point temperature(°C)', 'Solar Radiation (MJ/m2)'*. Use the Elbow method to determine the optimal number of clusters, which is 3.



1.2.2 Summary of K-Means

To measure the performance of the clustering algorithm, Sklearn provides a metric called Silhouette Score. This measure checks how much the clusters are compact and well separated. The more the score is closer to one, the better the clustering is.

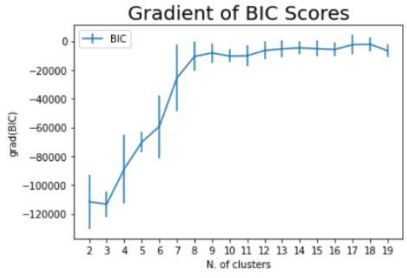
Table 1 shows the K-Mean clustering algorithm by each feature selection and dimension reduction method with Silhouette Score and Features or Components numbers.

(Table 1)

	K-means	PCA	R.P.	ICA	R.F.
Silhouette Score	0.3218	0.3267	0.2086	0.3591	0.2466
Features/ Components	15	10	12	2	7
Clusters	5	5	5	5	3

In this experiment on the K-means algorithm, the best feature selection and dimension reduction method is Independent component analysis. It has the best Silhouette Score is 0.3591 under the condition of the K=5 clusters.

1.3 Expectation Maximization



Expectation-Maximization algorithm comes under Gaussian models and is a way of thinking and modeling rather than an algorithm. In this algorithm, clusters are modeled as Gaussian distributions and not by their means. There is some correspondence between all data points and all clusters rather than correspondence between each data point to its cluster, as in the case of k-means clustering. We use the gradient of

Bayesian Information Criterion (BIC) to determine the number of clusters. This criterion gives us an estimation of how good the GMM in predicting the data we have. We can see that from the starting of cluster nine, there is no steep increase in gradient, so that we will pick 8 clusters.

1.3.1 Feature Selection & Dimension reduction

- **PCA**

The first ten components contribute to more than 90% variance.

	0	1	2	3	4	5	6	7	8	9
0	-3.014294	0.247630	-0.679963	-0.684526	0.737702	-1.322253	0.687272	1.022983	-0.068142	0.754092
1	-2.859904	0.734432	-0.573277	-1.003121	0.600002	-1.675753	0.725728	0.852378	0.270377	-0.320285
2	-2.895760	0.654897	-0.595077	-0.938961	0.628313	-1.531951	0.741040	0.757123	0.235132	-0.200406
3	-2.877791	0.676871	-0.589806	-0.946782	0.626391	-1.480283	0.758020	0.646811	0.270656	-0.304426
4	-3.118120	0.032958	-0.732680	-0.610665	0.769502	-0.993353	0.763200	0.642298	-0.020233	0.635820

- **Randomized Projections**

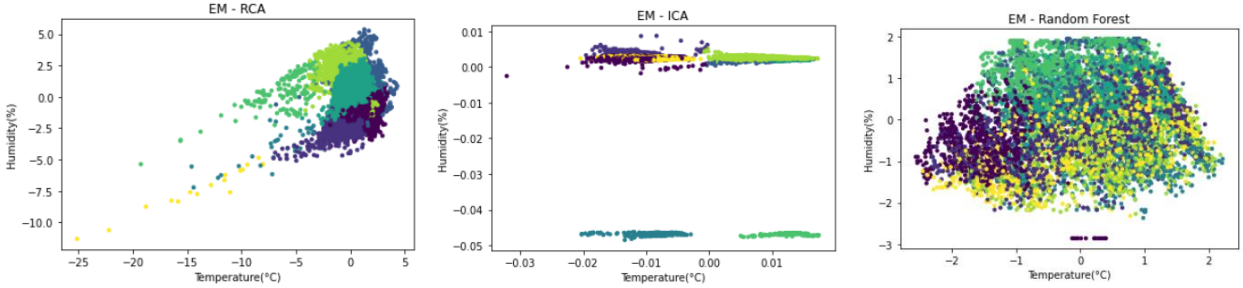
Try various dimensionality of the target projection space. When components are equal to 5, the Silhouette Score is the best. The cluster number is 8.

- **ICA**

Applied various components; when components are equal to 4, the Silhouette Score is the best. The cluster number is 8.

- **Feature Selection using Random Forest**

After performing feature selection on the original dataset, following features were obtained: 'Hour', 'Temperature(°C)', 'Humidity(%)', 'Wind speed (m/s)', 'Visibility (10m)', 'Dew point temperature(°C)', 'Solar Radiation (MJ/m2)'.



1.3.2 Summary of Expectation Maximization

(Table 2)

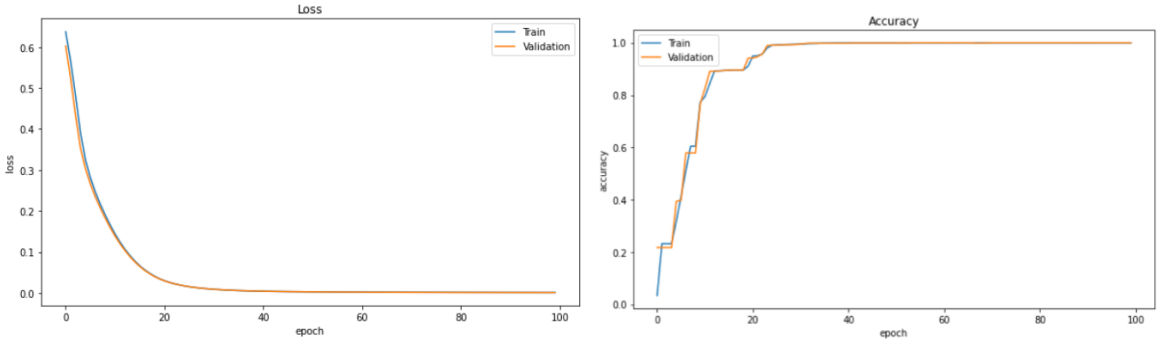
	EM	PCA	R.P.	ICA	R.F.
Silhouette Score	0.2503	0.2681	0.3114	0.3516	0.1409
Features/ Components	15	10	5	4	7
Clusters	8	8	8	8	8

Table 2 shows the Expectation-Maximization clustering algorithm by each feature selection and dimension reduction method. In this experiment, the best feature selection and dimension reduction method is Independent component analysis. It has the best Silhouette Score is 0.3516 under the condition of the K=8 clusters.

1.3.3 Neural Network on Cluster Output

Use the clustering results from Expectation-Maximization 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. The Accuracy is 0.9989.

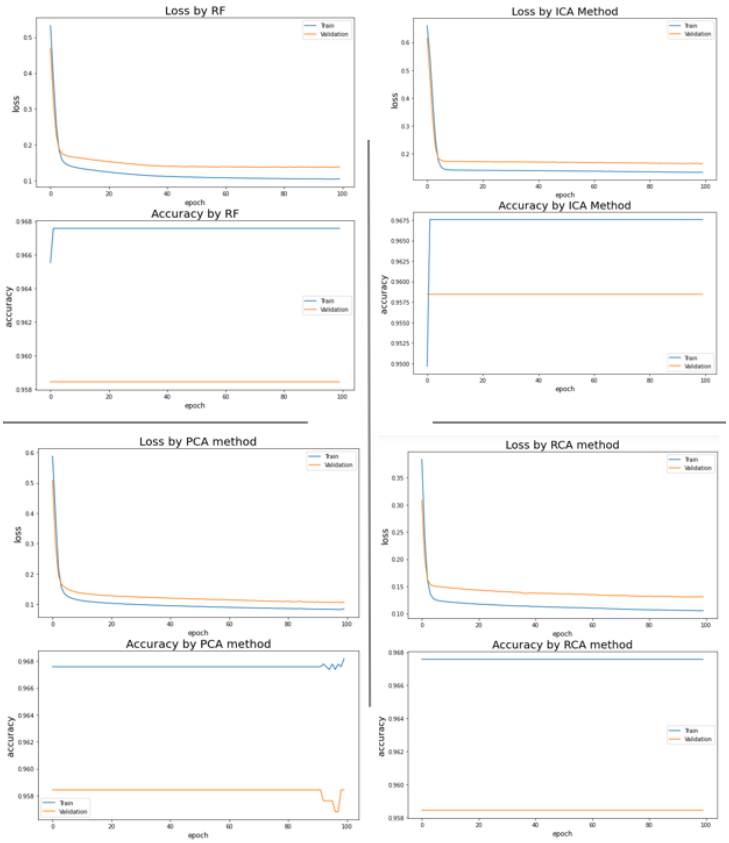
```
[[150   0   0   0   0   0   0   1]
 [  0 171   0   0   2   0   0   0]
 [  0   0 133   0   0   0   0   0]
 [  0   0   0 566   0   0   0   0]
 [  0   0   0   0 471   0   0   0]
 [  0   0   0   0   0 127   0   0]
 [  0   0   0   0   0   0 497   0]
 [  0   0   0   0   0   0   0 510]]
```



1.4 Artificial Neural Network

(Table 3)

	ANN	PCA	R.P.	ICA	R.F.
Accuracy	96.27%	96.84%	96.76%	96.76%	96.76%
Features/ Components	15	10	5	4	7
Elapsed Time	-	11.2929	10.7754	10.5324	10.3877



With two hidden layers from the previous assignment, the best-performed model has been picked for this experiment. the Accuracy is 96.27%

From Accuracy and loss, it is evident that feature and dimension reduction can improve performance. Compared to the four techniques, the Accuracy scores are very close, but by plotting the loss and Accuracy, the processes are not the same.

The PCA method has the best accuracy score, which is 96.88%. And PCA Method reduced five features, and the run time is 10.30 seconds.

2. IBM HR-Employee-Attrition & Performance

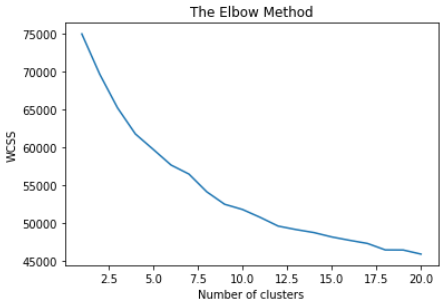
2.1 Dataset Explore

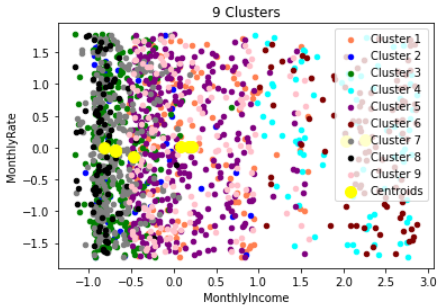
By reviewing the dataset, three variables are not appropriate to use to cluster or classify the dataset. And on this step, it is the first feature selection by using the understanding of the dataset.

1. The "PerformanceRating" column data were number rating records and were output data. The Performance Rating in a specific number has two category values: with lower rating=3 or higher rating=4.
2. Ignored "EmployeeCount" and "EmployeeNumber" columns since both columns have no help for cluster. "EmployeeCount" and "EmployeeNumber" are for index or count purposes.
3. Ignored the " PercentSalaryHike'" column since the " PercentSalaryHike'" information is positively correlated with the output, "Performance Rating." No further information is needed if considering the " PercentSalaryHike'" as an input attribute. After modification, I encoded the dataset by OneHotEncoder Method. The following is the shape of the dataset: [X (1470, 53), Y (1470)]

2.2 K-Means

The Elbow method was used to determine the optimal number of clusters. As from the figure below, the optimum value of k is 9. We can visualize the distribution with these two variables.

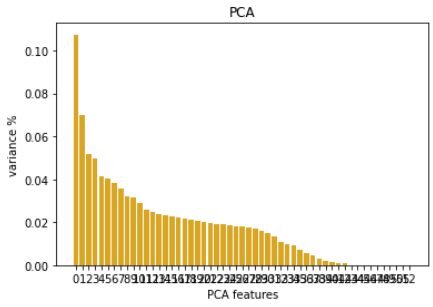


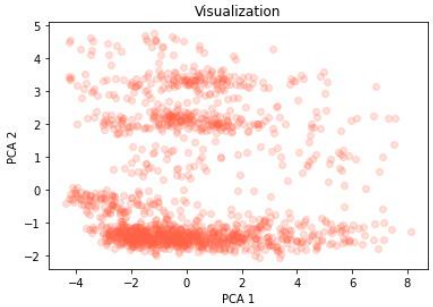


2.2.1 Feature Selection & Dimension reduction

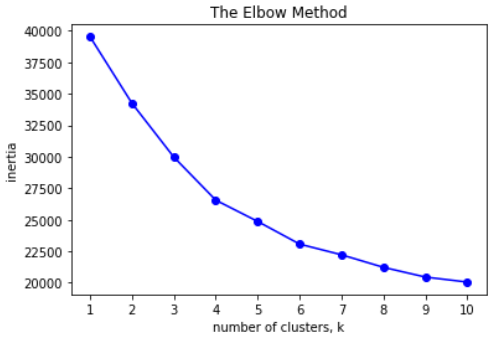
- **PCA**

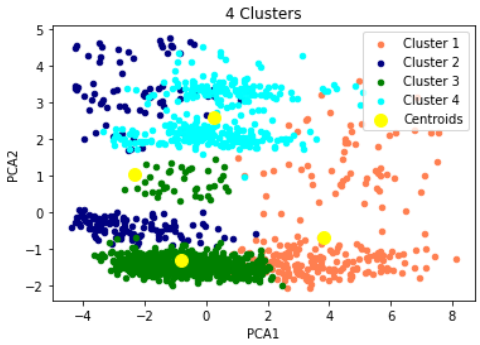
The PCA was applied to the modified dataset of IBM HR-Employee-Attrition & Performance, and the variance ratio was plotted. By selected the first 30 PCA to test the experiment. Now with these first two components, we can visualize the distribution.





The Elbow method was used to determine the optimal number of clusters. As from the figure below, the optimum value of k is 4.



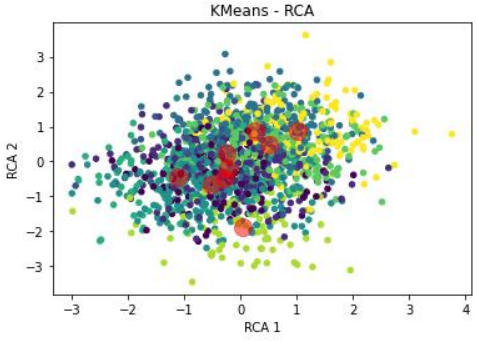


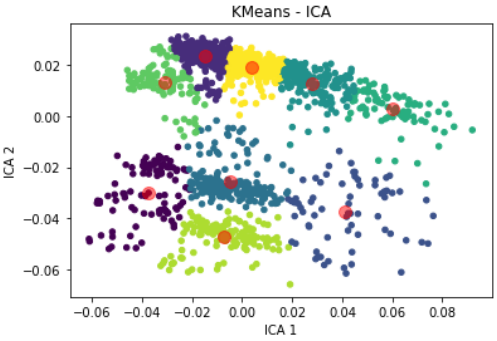
- **Randomized Projections**

Try various dimensionality of the target projection space. When components are equal to 42, the Silhouette Score is the best. The cluster number is 9.

- **ICA**

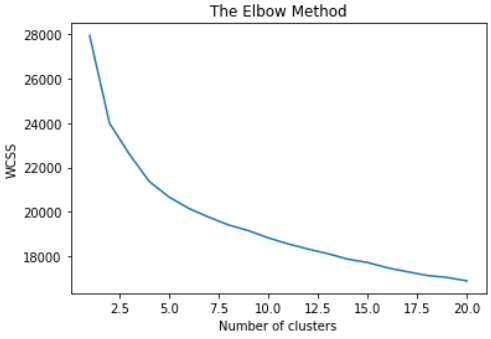
Applied various components; when components are equal to 2, the Silhouette Score is the best. The cluster number is 9.

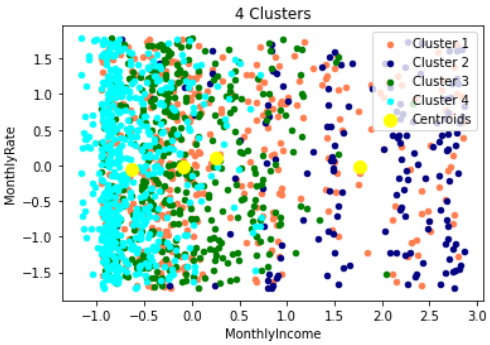




- **Feature Selection using Random Forest**

After performing feature selection on the original dataset, following features were obtained: *'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager', 'JobSatisfaction', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'RelationshipSatisfaction', 'HourlyRate', 'JobInvolvement', 'EnvironmentSatisfaction', 'DistanceFromHome', 'Education', 'DailyRate', 'Age'*. Use the Elbow method to determine the optimal number of clusters which is 4.





2.2.2 Summary of K-means

Silhouette Score is the measurement to check how much the clusters are compact and well separated. The more the score is closer to one, the better the clustering is.

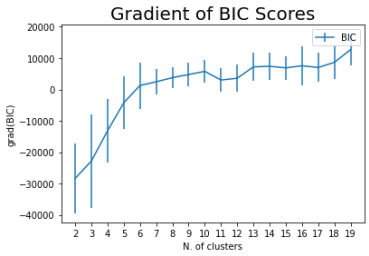
Table 4 shows the K-Mean clustering algorithm by each feature selection and dimension reduction method with Silhouette Score and Features or Components numbers.

(Table 4)

	K-means	PCA	R.P.	ICA	R.F.
Silhouette Score	0.1193	0.0850	0.0937	0.4266	0.0872
Features/ Components	53	30	42	2	19
Clusters	9	9	9	9	4

In this experiment on the K-means algorithm, the best feature selection and dimension reduction method is Independent component analysis. It has the best Silhouette Score is 0.4266 under the condition of the K=9 clusters.

2.3 Expectation Maximization



We use the gradient of the Bayesian Information Criterion (BIC) to determine the number of clusters. This criterion gives us an estimation of how good the GMM in predicting the data we have. We can see that from the starting of cluster six, there is no steep increase in gradient, so that we will pick 5 clusters.

2.3.1

- **PCA**

The PCA was applied to the modified dataset of IBM HR-Employee-Attrition & Performance, and the variance ratio was plotted. By selected the first 30 PCA to test the experiment.

- **Randomized Projections**

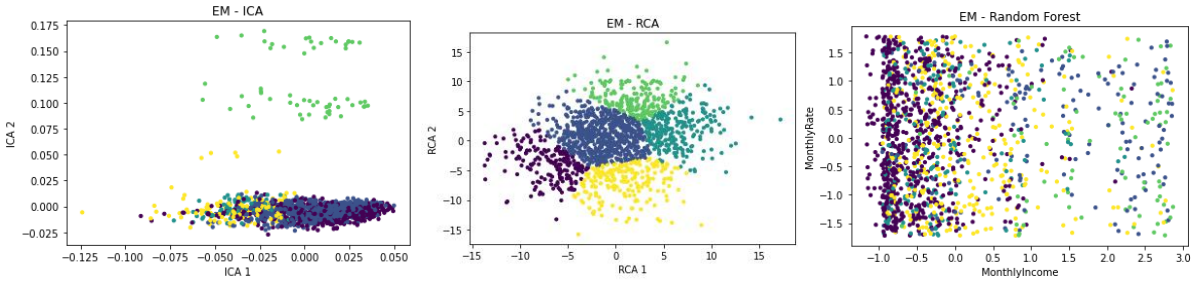
Try various dimensionality of the target projection space. When components are equal to 2, the Silhouette Score is the best. The cluster number is 5.

- **ICA**

Applied various components; when components are equal to 4, the Silhouette Score is the best. The cluster number is 5.

- **Feature Selection using Random Forest**

Following features were obtained: 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager', 'JobSatisfaction', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'RelationshipSatisfaction', 'HourlyRate', 'JobInvolvement', 'EnvironmentSatisfaction', DistanceFromHome', 'Education', 'DailyRate', 'Age'.



2.3.2 Summary of Expectation Maximization

(Table 5)

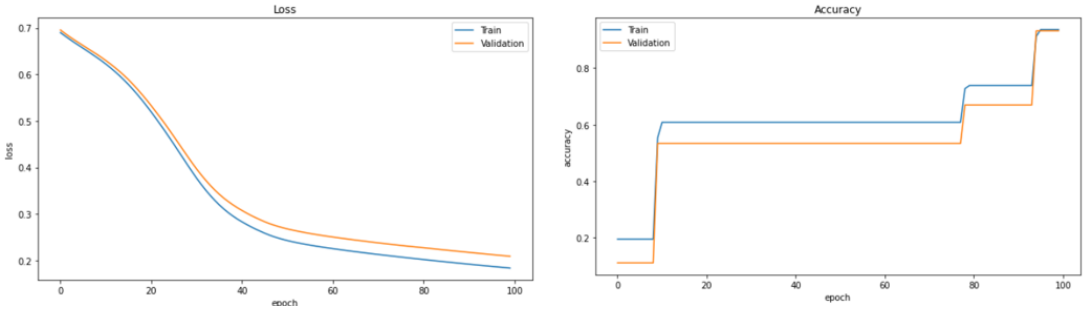
	EM	PCA	R.P.	ICA	R.F.
Silhouette Score	0.0806	0.0971	0.3018	0.3681	0.0760
Features/ Components	53	30	2	4	19
Clusters	5	5	5	5	5

Table 5 shows the Expectation-Maximization clustering algorithm by each feature selection and dimension reduction method. In this experiment, the best feature selection and dimension reduction method is Independent Component Analysis. It has the best Silhouette Score is 0.3681 under the condition of the K=5 clusters.

2.3.3 Neural Network on Cluster Output

Use the clustering results from Expectation-Maximization 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. The Accuracy is 0.8549.

```
[[  0   0   0   0  17]
 [  0  92   0   0   0]
 [  0   0 109   0   0]
 [  0   0   0  47   0]
 [  0   0   0   0 176]]
```

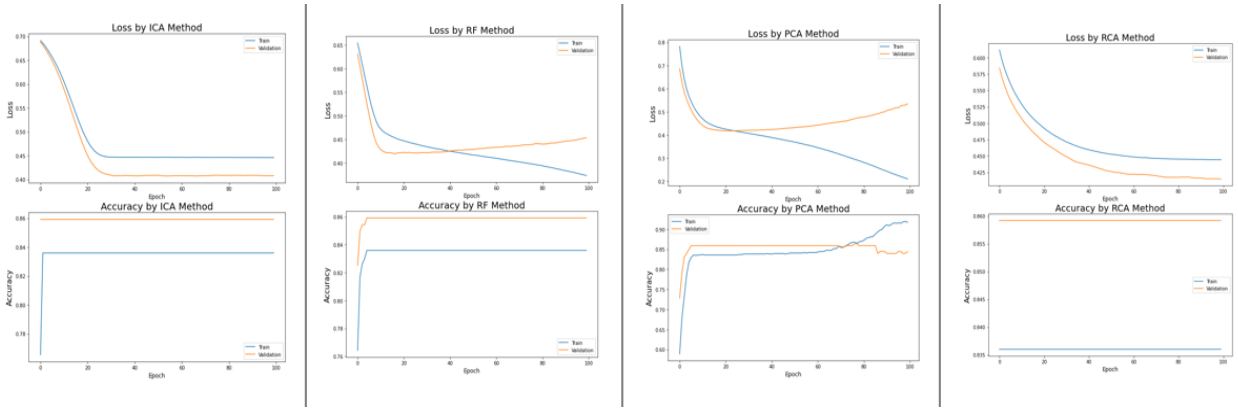


2.4 Artificial Neural Network

(Table 6)

	ANN	PCA	R.P.	ICA	R.F.
Accuracy	82.77%	84.81%	85.94%	85.94%	85.94%
Features/ Components	53	30	2	4	7
Elapsed Time	-	5.043	4.080	4.586	4.501

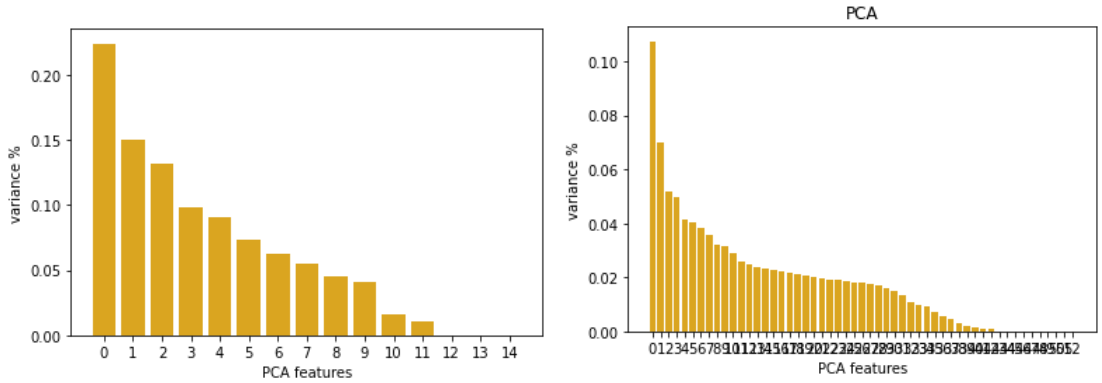
With two hidden layers from the previous assignment, the best-performed model has been picked for this experiment. The Accuracy is 82.77%. From Accuracy and loss, it is evident that feature and dimension reduction can improve performance. Compared to the four techniques, the Accuracy scores are very close or the same, but by plotting the loss and Accuracy, the processes are not the same. The ICA, Randomized Projections, and Random forest method has the best accuracy score is 85.94%. And can save computing by approximately 1 second.



3. Conclusion

Since there is no definite answer for finding the right number of clusters as it depends on (a) Distribution shape, (b) scale in the data set (c) Domain knowledge. Thus, Finding the number of clusters is a very subjective problem. For example, I try various component numbers in order to find the best Silhouette Score on Seoul Bike-Sharing Demand Dataset. But I try the math way, running all the possible component numbers in each method and choosing the best results on the IBM Dataset.

In my experiment, the ICA method has better results in every situation. Because PCA can't well reconstruct each variable's variation, and many variables are categorical variables, and the numbers in the same variable are rarely continuous.



Moreover, In the IBM dataset, the information is unique to each person. ICA's goal is to separate the individual sources by maximum statistical independence, so in this case, ICA might be a perfect method to transform the features. In conclusion, Dimensionality Reduction plays a significant role in machine learning, especially when working with hundreds of features.

```
Index(['PerformanceRating', 'Age', 'Attrition', 'BusinessTravel', 'DailyRate',
       'Department', 'DistanceFromHome', 'Education', 'EducationField',
       'EnvironmentSatisfaction', 'Gender', 'HourlyRate', 'JobInvolvement',
       'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus',
       'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'Over18',
       'OverTime', 'RelationshipSatisfaction', 'StandardHours',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole',
       'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')
```