Unsupervised Learning & Dimensionality Reduction

Juejing Han, jhan446@gatech.edu

***Abstract*—**K-Means Clustering outperforms Expectation Maximization Clustering. Cluster-enhanced datasets exhibit improvement. Linear and non-linear dimensionality reduction (DR) algorithms perform well on strongly collinear data, while Isomap outperforms most linear algorithms on data with weak collinearity. DR data may be computationally expensive because it may require a more complex model.

1 EXPERIMENT DESIGN

1.1 Datasets & Hypotheses

Records of diabetes (8 features) and rice (7 features) with binary targets are used. Pre-processed datasets (including missing data elimination and scaling) are referred to as Data1 (1932 samples) and Data2 (3810 samples). Original data characteristics: Both datasets are imbalanced and full rank – each feature provides unique and independent information; **Data2 exhibits strong collinearity while Data1 has weak collinearity** (Fig. 1 and 2); **Data1 has more outliers and is more imbalanced**.

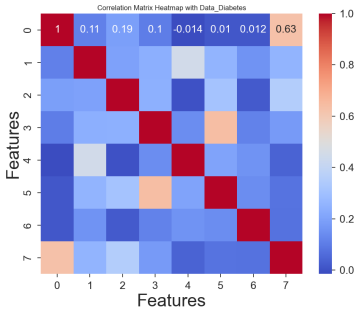

***Figure 1*—**Correlation Matrix for Data1

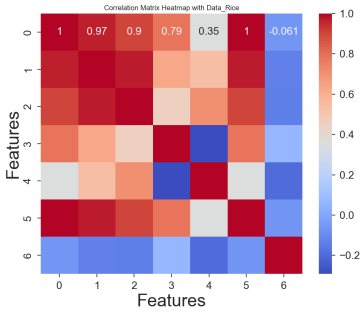

***Figure 2*—** Correlation Matrix for Data2

Hypothesis 1: Since both datasets perform well with K-nearest Neighbor, which is based on Euclidean distance, they should also perform well with Euclidean-distance-based K-means clustering.

Hypothesis 2: Because of Data2's strong collinearity, it should perform well with both linear and non-linear dimensionality reduction algorithms.

1.2 Clustering Algorithms

Expectation Maximization clustering (EM) is a probabilistic model-based clustering algorithm that finds the maximum likelihood of parameters. EM assumes that the data is generated from Gaussian Mixture Models and assigns data points into its probable cluster.

K-Means clustering (Kmeans) is a Euclidean-distance-based algorithm that assigns data into non-overlapping clusters with the nearest centroid. Its goal is to minimize within-cluster variances so that clusters are internally homogeneous. It assumes spherical clusters with equal variances.

1.3 Dimensionality Reduction Algorithms

Principal Component Analysis (PCA) is a linear method. It uses Singular Value Decomposition to transform data into a lower dimension where most variation in the data can be described[1].

Independent Component Analysis (ICA) is a linear method. It separates multivariate signals into constituent sources by assuming at most one component is Gaussian distributed, while the other components are statistically independent[2]. It aims to capture independence between components.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA
[2] https://en.wikipedia.org/wiki/Independent_component_analysis

Sparse Random Projection (SRP) is also a linear method. It projects the data into a lower dimension with a sparse random matrix and aims to approximate the preservation of pairwise distances or similarities between data points[3].

Isomap (IMP) is a manifold learning method used for nonlinear dimensionality reduction. With a combination of nearest-neighbor graph and Multidimensional Scaling, it focuses on preserving the intrinsic geometric structure and relationship when mapping data into a lower dimension[4].

2 RESULTS & ANALYSIS

2.1 Experiment 1 – Clustering

2.1.1 *Expectation Maximization Clustering (EM)*

Akaike and Bayesian Information Criterion (AIC, BIC) are estimators of prediction error used for probabilistic model selection[5]. BIC applies a larger penalty and tends to favor simpler models; therefore, BIC is chosen as the primary metric. The optimal cluster number (k) for Data1 is 7, as the BIC reaches its minimum (Fig. 3). In Fig. 4, AIC and BIC exhibit similar patterns and tend to converge beyond 12, which is selected as the value of k.

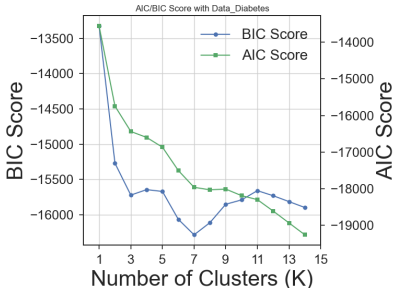

Figure 3—AIC/BIC Score in EM with Data1

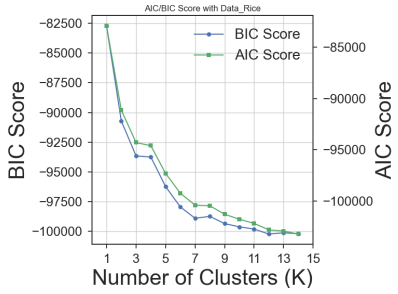

Figure 4—AIC/BIC Score in EM with Data2

2.1.2 *K-means Clustering (Kmeans)*

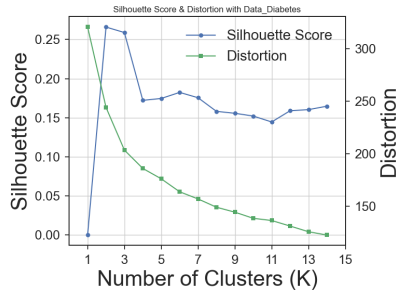

Figure 5—Silhouette Score & Distortion in Kmeans with Data1

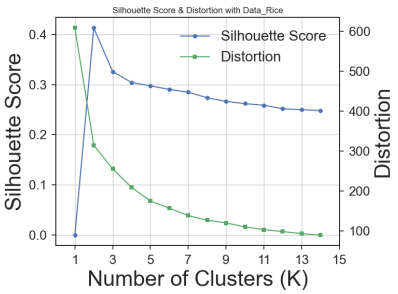

Figure 6—Silhouette Score & Distortion in Kmeans with Data2

Distortion is the average squared distance from cluster centroids to data points. Silhouette Score measures the similarity of data points in their clusters compared to them in neighboring clusters, and it considers the trade-off between the cohesion and separation of clusters. In Fig. 5 and 6, Distortion does not show a clear inflection point, but Silhouette Score peaks when k=2, which is the optimal choice for both datasets.

2.1.3 *Compare and Contrast*

Kmeans identifies 2 clusters, matching the true labels, whereas EM identifies more clusters, which may align with the data's imbalance nature. In the clustering heatmaps (Fig. 7 and 8), Kmeans divides data points into two well-defined clusters, while the interpretation of the meaningfulness of EM clusters is challenging. This discrepancy may result from metric differences, as EM also finds k=2 using the Silhouette

[3] https://en.wikipedia.org/wiki/Random_projection
[4] https://en.wikipedia.org/wiki/Isomap
[5] https://en.wikipedia.org/wiki/Bayesian_information_criterion

Score. Aside from the different metrics they employ, Kmeans assumes spherical clusters and enforces hard clustering, while EM permits soft clustering. Besides, the imbalanced data nature might also lead EM to identify more clusters while using the BIC metric.

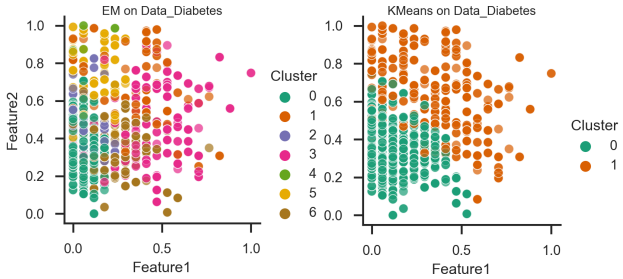

Figure 7—Heatmap on Data1 (Left: EM, Right: Kmeans)

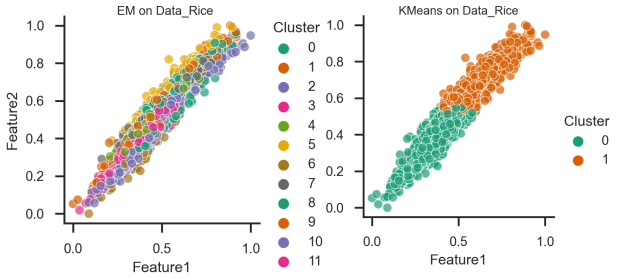

Figure 8—Heatmap on Data2 (Left: EM, Right: Kmeans)

Integrate the predicted labels into the original features and test with an untuned Neural Network model (weak learner). Kmeans (F1 score is 0.71 for Data1 and 0.93 for Data2) outperforms EM (0.70 and 0.92) Besides, Adjusted Rand Index (ARI) measures the similarity between true and predicted labels. Kmeans has a higher ARI (0.13 and 0.69 for Data1 and Data2, respectively) than EM (0.12 and 0.30). Therefore, **Hypothesis 1 holds true**. Data1 has a lower ARI because it has more outliers and is highly imbalanced, and Kmeans underperforms with Data1 because Kmeans is sensitive to outliers.

Change the metric to Silhouette Score, EM identifies k=2 for both datasets, resulting in better performance. Kmeans is a distance-based algorithm and is sensitive to outliers, since Data1 has significant outliers, outlier handling techniques on Data1 may lead to improvement for Kmeans.

2.2 Experiment 2 – Dimensionality Reduction

2.2.1 *Principal Component Analysis (PCA)*

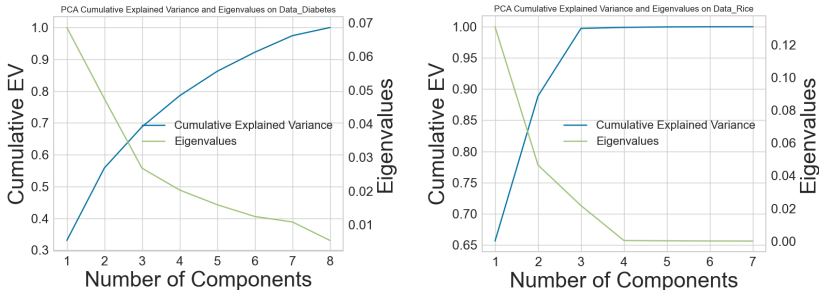

Figure 9—PCA CEV and Eigenvalue (Left: Data1, Right: Data2)

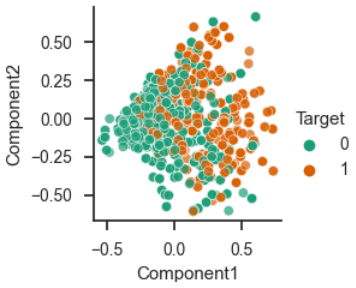

Figure 10—PCA-Reduced Data1

For Data1, the eigenvalue decreases as the number of components increases, while for Data2, the eigenvalue decreases and then plateaus when the number of components reaches 4 (Fig. 9). Explained variance (EV) represents the portion of variance explained by each component[6], and PCA captures all the variability in the data when Cumulative EV (CEV) reaches 1. For Data1, by setting CEV > 90% as the threshold, the optimal number of components (n) is 6 (Fig. 9 Left). Meanwhile, CEV reaches 1 when there are three components for Data2, resulting in 3 as the optimal choice (Fig. 9 Right).

To save space, the first two components of PCA-reduced Data1 are demonstrated in Fig. 10. Data1, which contains more outliers, provides an opportunity to assess the algorithm's performance with noisy data. In PCA, these two components are orthogonal. The scatter plots in the original Data1 pair plots exhibit similar patterns, and the PCA-reduced data shows a reduction in the number of outliers.

A weak learner is utilized to evaluate the performance of PCA, and the F1 scores are 0.67 for Data1 and 0.92 for Data2. PCA excels with Data2 because of its strong collinearity, aligning with PCA's linear nature.

[6] https://www.baeldung.com/cs/pca

However, PCA may not perform well with Data1, primarily due to its weak collinearity. The lower score for Data1 is partially due to the model bias identified in assignment 1. Besides, PCA is sensitive to noise, which can introduce additional variance and increase dimensionality. Given that Data1 has more outliers than Data2, these factors may also contribute to its underperformance.

2.2.2 *Independent Component Analysis (ICA)*

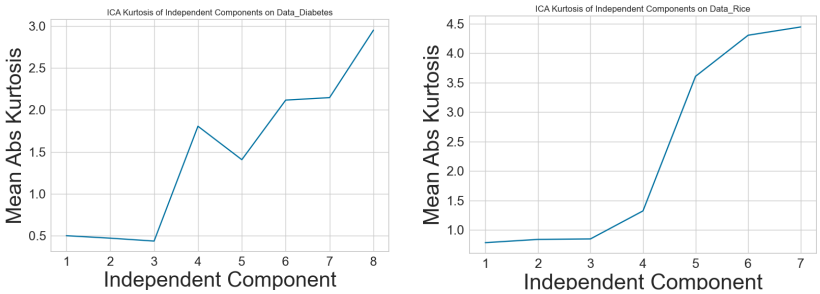

Figure 11—ICA CEV and Eigenvalue (Left: Data1, Right: Data2)

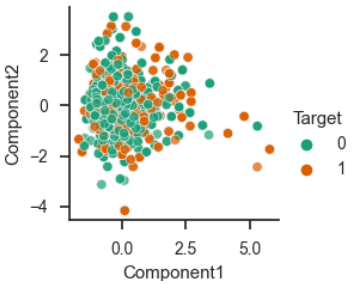

Figure 12—ICA-Reduced Data1

In Fig. 11, kurtosis gradually increases and reaches its maximum when the component number (n) is equal to the feature number (N). However, when n=N, dimensionality reduction is meaningless. Therefore, n=6 is chosen for Data1, corresponding to the local plateau, and n=6 for Data2, where kurtosis tends to level off.

In Fig. 12, ICA components are independent. Similar scatter plots can also be found in the original Data1 pair plots, and the ICA-reduced pattern exhibits a reduction in the number of outliers, a byproduct of dimensionality reduction.

The F1 scores are 0.77 for Data1 and 0.93 for Data2. As a linear algorithm, ICA also excels with Data2, which has strong collinearity. In the case of Data1, ICA has a higher score than PCA. Given Data1's weak collinearity, ICA's assumption of independence may better capture underlying patterns, especially those patterns that are not well represented by orthogonal components sought by PCA. However, the presence of noise in Data1 may cause ICA to underperform, resulting in components that contain signal and noise. ICA achieves higher scores than PCA with both datasets, indicating ICA captures meaningful components.

2.2.3 Sparse *Random Projection (SRP)*

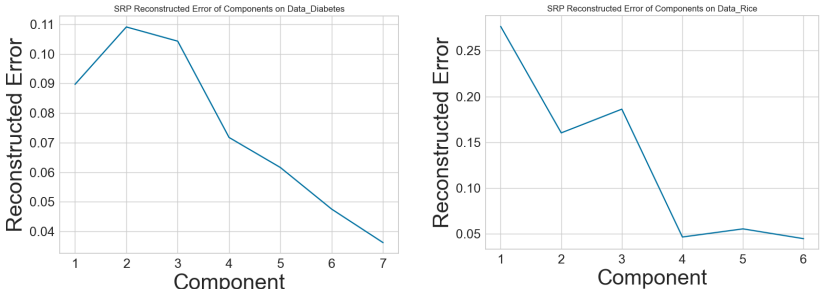

Figure 13—SRP CEV and Eigenvalue (Left: Data1, Right: Data2)

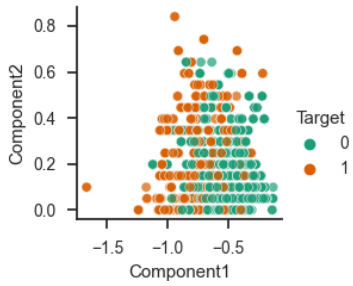

Figure 14—SRP-Reduced Data1

Reconstruction error (RE) quantifies the disparity between the original data and its reconstructed counterpart. For Data1, RE exhibits an overall decreasing pattern as the number of components increases, and n=4 is an elbow point, where RE slows its rate of decrease, making it an optimal choice (Fig. 13 Left). Regarding Data2, RE generally decreases as the number of components increases and tends to stabilize beyond 4 components, leading to the selection of n=4 for Data2 (Fig. 13 Right).

Similar scatter patterns can also be observed in the original Data1 pair plots, and the SRP-reduced data exhibits noise reduction (Fig. 14). However, the label distribution is notably different, and this divergence implies that SRP may underperform on Data1, which is further verified by the weak learner test.

The F1 scores are 0.60 for Data1 and 0.92 for Data2. As a linear algorithm, SRP also performs well with Data2, because Data2 is strongly collinear. SRP has a lower score with Data1, because Data1 has weak collinearity and more outliers. Noise from Data1 may disrupt the sparsity structure and the pairwise distances between data points, and hence cause SRP to underperform. Additionally, SRP behaves differently with different random seeds due to its inherent randomness, so implementing multiple runs and carefully choosing the random seed may improve performance.

2.2.4 Isomap (IMP)

Reconstruction error (RE) is also a relevant metric in Isomap (IMP). RE shows a linear decrease as the number of components increases (Fig. 15 Left). By setting RE < 0.4 as the threshold, n=6 is determined to be the optimal choice for Data1. In the case of Data2, RE generally decreases, and n=3 corresponds to the inflection point, beyond which RE levels off (Fig. 15 Right), making it the optimal choice.

IMP-reduced Data1 exhibits a different scatter pattern from the original Data1 pair plots. It is non-linear with high dispersion (Fig. 16). IMP projects data onto a lower-dimensional manifold, so the change in scatter patterns may be a result of a representation of the non-linear structure present in Data1. However, IMP-reduced Data1 exhibits more outliers, which has the potential to cause the algorithm to underperform.

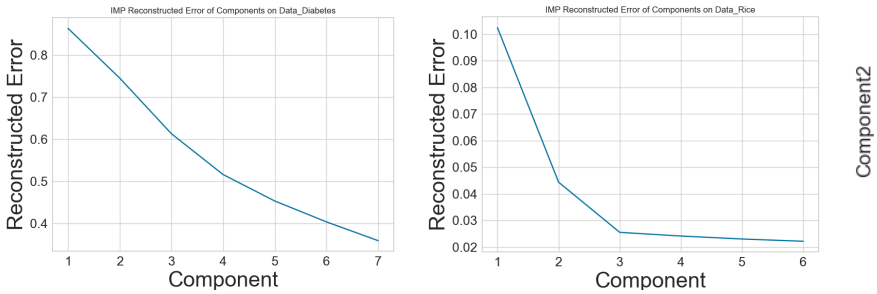

Figure 15—IMP CEV and Eigenvalue (Left: Data1, Right: Data2)

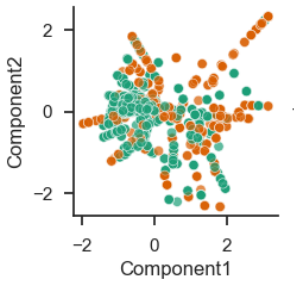

Figure 16—IMP-Reduced Data1

IMP scores are 0.70 for Data1 and 0.93 for Data2. As a non-linear algorithm, IMP performs well with Data2, which has a clear linear-nonlinear mixed structure. For Data1, IMP achieves a higher score than PCA and SRP. It is because of Data1's weak collinearity, which is more suitable for a non-linear algorithm. However, IMP has a lower score than ICA, it could be due to the noise in Data1 which leads to distortion in the computed manifold for IMP.

Both linear and non-linear algorithms perform well with Data2 because of Data2's strong collinearity. Therefore, **Hypothesis 2 holds true**. Besides, Data2 achieves similar scores with different algorithms (0.92-0.93), while Data1 exhibits varying scores (0.60-0.77). Other than model/algorithm bias, Data1 has more outliers and is more imbalanced than Data2, resulting in the tendency of underperformance.

2.3 Experiment 3 – Clustering with SRP and IMP

Apply EM and Kmeans to reduced datasets from section 2.2, using the same strategy from section 2.1 to determine the optimal cluster number (figures not displayed but generated by code).

2.3.1 SRP + Clustering

SRP is selected for demonstration because 1) SRP-reduced Data2 shows distinct structures from IMP, and 2) cluster-enhanced SRP datasets exhibit improvement. Kmeans evenly divides SRP-reduced data into 2 clusters (Fig. 17 Right and 18 Right), aligning with the original data clustering and the true labels. It is because 1) the reduced data successfully captures the underlying data structure, and 2) the metric (Silhouette Score) used by Kmeans. EM divides reduced datasets into 5 and 9 clusters, which is fewer than the original data clustering (7 and 12 clusters), implying feature reduction may lead to cluster reduction.

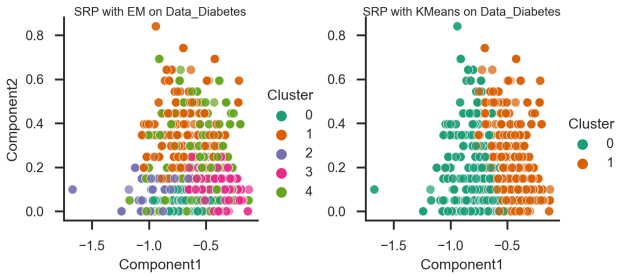

Figure 17—SRP Heatmap on Data1 (Left: EM, Right: Kmeans)

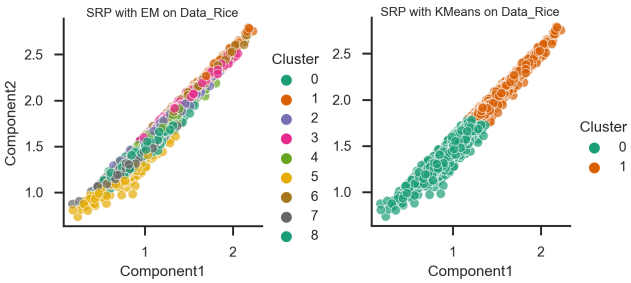

Figure 18—SRP Heatmap on Data2 (Left: EM, Right: Kmeans)

A weak NN is introduced for evaluation. For reduced Data1, F1 score of EM and Kmeans are 0.63 and 0.62, respectively. They show enhancement compared to the reduced dataset without clustering (0.59). Reduced Data1 is highly imbalanced with more outliers, and EM suggests more than 2 clusters, which may align with the data characteristics, resulting in better performance. For reduced Data2, F1 score of EM and Kmeans are 0.915 and 0.919, respectively, and Kmeans still exhibits superiority. This is because reduced Data2 has fewer outliers and the choice of k=2 in Kmeans effectively divides the data points.

2.3.2 IMP + Clustering

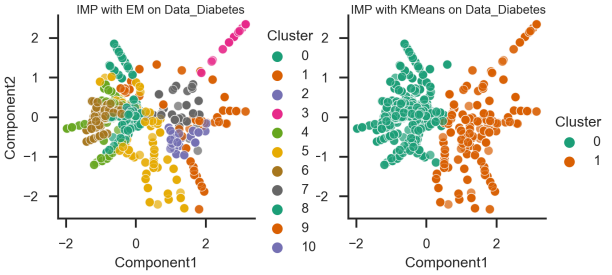

Figure 19—IMP Heatmap on Data1 (Left: EM, Right: Kmeans)

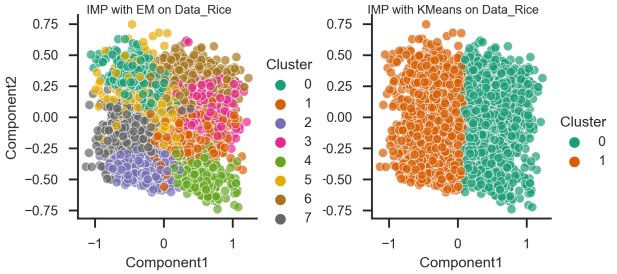

Figure 20—IMP Heatmap on Data2 (Left: EM, Right: Kmeans)

The scatter pattern of reduced Data1 has high dispersion, leading to more outliers and, subsequently, more EM clusters (Fig. 19 Left) and underperformance (F1 score is 0.65, while without clustering the score is 0.66). This could be attributed to outliers in Data1. As mentioned in section 2.2.4, IMP is sensitive to noise, which can distort its manifold. NN results show that Kmeans improves the performance of IMP-reduced datasets: 0.66/0.93 for reduced Data1/Data2 before clustering, and 0.68/0.94 after clustering.

Furthermore, Fig. 18 and 20 illustrate the potential distinction between linear and nonlinear algorithms while working with Data2, which has a clear linear-nonlinear mixed structure: SRP captures the linearity, while IMP captures the non-linearity.

2.4 Experiment 4 – NN Learner with ICA and IMP on Data2

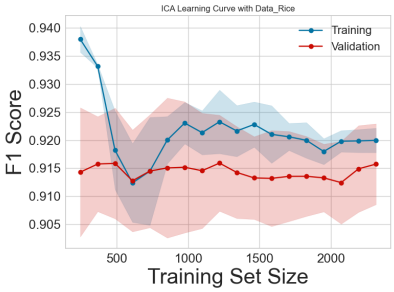

Figure 21—ICA Learning Curve with Data2

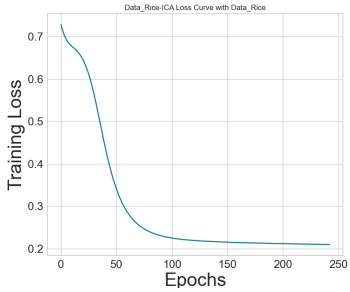

Figure 22—ICA Loss Curve with Data2

For ICA-reduced data, increasing sample size mitigates overfitting (Fig. 21). Both training and validation scores peak at 1200 samples and converge thereafter. Additional samples may lead to further improvement.

The model bias is acceptable, and the variance is small (within 0.1). Loss curve converges within 100 epochs, indicating effective model learning (Fig. 22).

IMP-reduced data exhibits a similar pattern. Overfitting is reduced with more samples; both scores peak at 1300 samples and converge afterward (Fig. 23). Adding more samples does not lead to enhancement. The model bias is acceptable, and the variance is smaller than ICA, indicating that IMP generalizes better than ICA. IMP loss curve converges within 3 epochs (Fig. 24), signaling less computational cost than ICA.

Figure 23—IMP Learning Curve with Data2

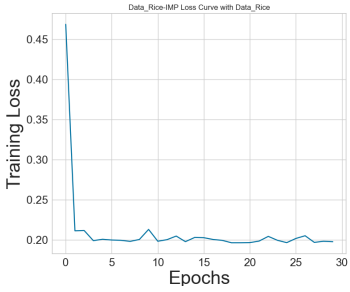

Figure 24—IMP Loss Curve with Data2

ICA-reduced data and IMP-reduced data achieve higher scores than the original data, demonstrating the superiority of dimensionality reduction techniques (Table 1). Grid search favors a smaller learning rate (0.001) with two hidden layers for ICA, leading to the longest training time and more epochs to converge (Fig. 22). IMP with one hidden layer and a larger learning rate (0.1) requires shorter training time than the original data, which needs two hidden layers with a learning rate of 0.01. IMP corresponds to a simpler model, likely due to its fewer features (n=3). ICA and IMP outperform the original data, indicating their successful extraction of informative components from Data2, which has a strong collinearity. This also verifies that **Hypothesis 2 is true**.

Table 1—Evaluation of ICA and IMP with Data2.

Clustering	Training time (s)	Testing time (s)	Accuracy	F1 Score
ICA	0.62233	0.00014	0.933	0.924
IMP	0.15071	0.00037	0.937	0.930
Original	0.24827	0.00031	0.930	0.919

2.5 Experiment 5 – NN Learner with EM and Kmeans

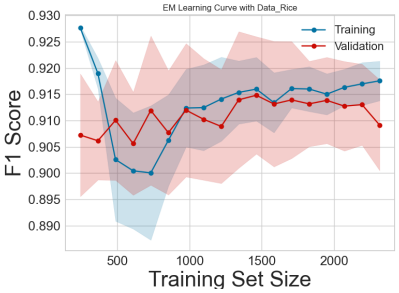

Figure 25—EM Learning Curve with Data2

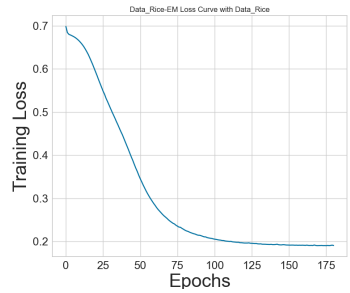

Figure 26—EM Loss Curve with Data2

Incorporate the predicted labels (from EM/Kmeans) into the original features and test the new dataset with the NN learner. For both cluster-enhanced datasets, adding more samples initially mitigates overfitting. However, the training score subsequently drops, while the validation score exceeds the training score (Fig. 25 and 27). This phenomenon is possibly attributed to data splitting, making the validation set more representative of the test data distribution. It disappears when a different data-splitting method is employed. Both scores tend to converge beyond 1000 samples. Additional samples do not lead to further

improvement. The model bias is acceptable, and the Kmeans variance is smaller than EM, indicating that Kmeans generalizes better.

Loss curves exhibit a continuous decrease followed by stabilization after 100 and 150 epochs for EM-enhanced and Kmeans-enhanced data, respectively (Fig.26 and 28), indicating effective model learning and faster convergence with EM.

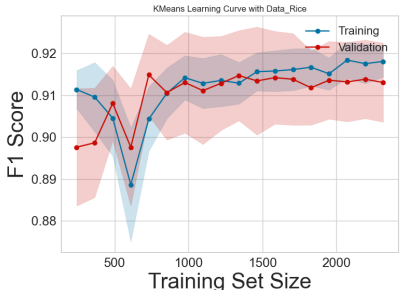

Figure 27—Kmeans Learning Curve with Data2

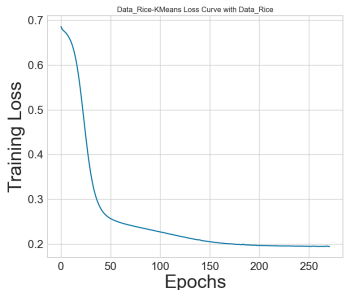

Figure 28—Kmeans Loss Curve with Data2

Table 2 illustrates that Kmeans-enhanced data achieves the highest scores at a lower computational cost compared to the original data, and EM-enhanced data reaches the second highest score with the shortest running time. This verifies that **Hypothesis 1 is true**. Kmeans identifies k=2, which matches the true labels. EM uses a soft clustering approach, allowing overlaps in clusters, which may be beneficial to imbalanced data. A simpler model for Kmeans-enhanced data and a weaker regularization (which allows a quicker convergence) for EM-enhanced data make cluster-enhanced datasets more computationally efficient.

Table 2—Evaluation of EM and Kmeans with Data2.

Clustering	Training time (s)	Testing time (s)	Accuracy	F1 Score
EM	0.60925	0.00037	0.937	0.929
Kmeans	0.79271	0.00020	0.941	0.933
Original	0.81275	0.00040	0.937	0.928

3 CONCLUSIONS

Kmeans outperforms EM with both datasets and performs better with Data2 than Data1 because Data1 has more outliers. For Data2, Kmeans achieves the highest score in the fine-tuned Neural Network test, and EM also exhibits improvement. It is because Kmeans successfully identifies k=2 for the binary datasets, and EM allows overlap clusters, which might be beneficial to imbalanced data. With the additional information from clustering, cluster-enhanced data requires a simpler model and less running time.

Data2 is strongly collinear, so it works well with linear and non-linear dimensionality reduction algorithms. Data1 exhibits weak collinearity, and it performs better with the non-linear algorithm, while ICA also works effectively with Data1. Fine-tuned Neural Network results further prove that both linear (ICA) and non-linear (IMP) algorithms show improvement on Data2. However, with reduced features, the data may demand more running time since it may require a more complex model to achieve a higher score.

Furthermore, it is important to consider the trade-off in dimensionality reduction, as it can lead to information loss. Striking the right balance between dimensionality reduction and information retention is essential for effectively implementing dimensionality reduction methods.

After dimensionality reduction (SRP and IMP), Kmeans still outperforms EM, because it identifies k=2, which matches the true labels. Reduced datasets generally correspond to reduced EM clusters. However, Data1 has more outliers and is more imbalanced, so IMP-reduced Data1 has more clusters, resulting in underperformance.