Report of Assignment 3 - Unsupervised Learning and Dimensionality Reduction

Qizhi Zhao

Georgia Institute of Technologies

***Abstract*—This project explores the use of clustering and dimensionality reduction techniques on the Iris and Diabetes datasets. We implemented Expectation Maximization (EM) and K-Means for clustering, and Randomized Projections (RP), Principal Component Analysis (PCA), and Independent Component Analysis (ICA) for dimensionality reduction. Our experiments demonstrate how these techniques impact data analysis and machine learning performance. Integrating the reduced features into a neural network improved runtime performance and classification accuracy.**

***Index Terms*—**

I. Introduction

Unsupervised learning techniques, such as clustering and dimensionality reduction, are essential for analyzing data without labels. These methods help uncover hidden patterns and structures in the data.

Clustering algorithms, like Expectation Maximization (EM) and K-Means, group similar data points together, revealing natural groupings within the data. Dimensionality reduction techniques, such as Randomized Projections (RP), Principal Component Analysis (PCA), and Independent Component Analysis (ICA), reduce the number of features while preserving important information. This simplification makes data easier to visualize and improves the performance of other algorithms.

This project explores the use of clustering and dimensionality reduction algorithms on two datasets, Iris and Diabetes. Through experiments combining clustering with these dimensionality reduction techniques, we aim to understand their impact on data analysis and machine learning. Additionally, integrating the reduced features into a neural network improves runtime performance and classification accuracy.

II. Datasets and Hypotheses

A. Datasets

1) Description: The project utilizes two distinct datasets: the Iris dataset and the Diabetes dataset, both of which serve different analytical purposes. The Iris dataset comprises 150 instances, each representing iris flowers characterized by 4 features: sepal length, sepal width, petal length, and petal width. These features are used to classify each instance into one of three species: setosa, versicolor, or virginica. On the other hand, the Diabetes dataset consists of 442 instances, with 10 features including age ,sex,bmi(body mass index),bp(average blood pressure),s1 tc(total serum cholesterol),s2 ldl(low-density lipoproteins),s3 hdl(high-density lipoproteins),s4 (tch, total cholesterol / HDL),s5 (ltg, possibly log of serum triglycerides level),s6(glu, blood sugar level).

B. Reason of Selecting

The Iris dataset's simple and well-defined structure makes it ideal for testing and validating the effects of clustering and dimensionality reduction algorithms. It could help us to have a clear understanding and better visualization of these methods.

The complexity and high-dimensional characteristic of the Diabetes dataset make it an excellent candidate for evaluating unsupervised learning algorithms. Applying clustering and dimensionality reduction techniques to this dataset helps assess their effectiveness in real-world applications, particularly in the field of medical data analysis.

Together, these datasets provide a comprehensive experimental platform, offering both straightforward cases and complex real-world applications, thus helping to gain deeper insights into the practicality and limitations of clustering and dimensionality reduction techniques.

C. Hypotheses

III. Methodology

A. Clustering

Clustering algorithms are an important tool in unsupervised learning. Their main purpose is to group data points into clusters based on their similarity or proximity in feature space. The main goal of clustering is to discover the inherent structure and patterns in the data without the need for a labels. These algorithms work by partitioning the data points into clusters in such a way that maximized the similarity of data points within a cluster and minimized the similarity between different clusters. Clustering methods vary in strategy and complexity, but they all aim to organize data into meaningful groups for further analysis and understanding of the relationships between the data. Evaluating clustering results usually involves assessing cluster coherence, density, and separation, using metrics such as silhouette coefficient or within-cluster sum

of squares. Effective data preprocessing, including normalization and feature selection, is important to improve clustering performance and extract valuable insights from the data.

1) Expectation Maximization (EM): The Expectation-Maximization (EM) algorithm, first proposed by Dempster et al. in 1977, is designed to find the maximum likelihood estimates of parameters in probability models with unobservable hidden variables. Initially intended to address issues with missing data, the EM algorithm alternates between two steps: Expectation (E) and Maximization (M). In the E step, the algorithm calculates the expected value of the likelihood using current estimates of the hidden variables. In the M step, it maximizes this likelihood to update the parameter estimates. This iterative process continues until convergence. The EM algorithm is widely used in data clustering, particularly in machine learning and computer vision, due to its effectiveness in handling incomplete data and latent variables.

2) K-Means : The K-Means algorithm is a clustering method used in machine learning and statistics to partition a dataset into K distinct, non-overlapping clusters. Introduced by MacQueen in 1967, the algorithm aims to minimize the within-cluster variance. It starts by randomly initializing K centroids, which represent the center of each cluster. The algorithm then alternates between two steps: assignment and update. In the assignment step, each data point is assigned to the nearest centroid based on the Euclidean distance. In the update step, the centroids are recalculated as the mean of all data points assigned to each cluster. These steps are repeated iteratively until the centroids no longer change significantly or a predefined number of iterations is reached. K-Means is widely used for its simplicity and efficiency, but it requires the number of clusters K to be specified in advance and is sensitive to the initial placement of centroids.

B. Metrics - Adjusted Rand Index

The Adjusted Rand Index (ARI) is a metric used to evaluate the similarity between two clustering results by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. Unlike the unadjusted Rand Index, the ARI is corrected for chance, which makes it more robust for evaluating clustering performance.The ARI is defined as follows:

$$\text{ARI} = \frac{\text{RI} - \text{Expected RI}}{\max(\text{RI}) - \text{Expected RI}}$$

1) Meaning of the value: An ARI of 1 indicates perfect agreement between the predicted and true clusters.

An ARI of 0 indicates that the clustering is random.

Negative values of ARI suggest that the clustering result is worse than random clustering.

C. Metrics - Silhouette Score

The Silhouette Score is a metric used to evaluate the quality of clustering by measuring how similar each data point is to its own cluster compared to other clusters. The score ranges from -1 to 1, where higher values indicate better-defined clusters.

The Silhouette Score for a data point i is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

1) Meaning of the value: A score close to 1 indicates that the data points are well-clustered and are appropriately assigned to clusters.

A score close to 0 indicates that the data points are on or very close to the decision boundary between two neighboring clusters.

A negative score indicates that the data points might have been assigned to the wrong cluster.

D. Dimensionality Reduction

Dimensionality reduction is also a type of unsupervised learning. It is the process of mapping high-dimensional data to a low-dimensional space, which is closely related to the concept of lossy compression in information theory. There is no completely lossless dimensionality reduction, which means that after the process, some information must be lost. By transforming data into a lower-dimensional space, we can mitigate issues like the ”curse of dimensionality”, enhance visualization, and improve the efficiency of machine learning algorithms.

In this project, we are going to apply four different dimensionality reduction methods: Randomized Projections (RP), Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Non-linear Manifold Learning Algorithm.These dimensionality reduction techniques were applied to the Iris and Diabetes datasets. The processed data was then used for further analysis and clustering. By reducing the dimensionality, we aimed to improve the performance and interpretability of the clustering algorithms and the neural network models used in subsequent experiments. The reason why we need Dimensionality Reduction is that data is redundant. Some of the dimension could be useless information or with repeated information of other. The data after dimensionality reduction generally retains most of the important information of the original data. It can completely replace the input to do some other work, thereby greatly reducing the amount of calculation. For example, it can be reduced to two or three dimensions for visualization.

1) Randomized Projections (RP): Random Projection is a dimensionality reduction algorithm which has the core idea to map high-dimensional data to low-dimensional space through a randomly generated projection matrix. The random projection algorithm is based on the Johnson-Lindenstrauss lemma, which states that under certain conditions, after high-dimensional data is mapped to low-dimensional space through random projection, the distance between data points can remain roughly unchanged, thereby preserving the geometric structure of the data.

2) Principal Component Analysis (PCA): Principal component analysis (PCA) is a method of data dimensionality reduction algorithm that aims to reduce the input n-dimensional data to k-dimensional data, where k is less than n. The basis

of the new k dimensions is an orthogonal basis in the n-dimensional space, assuming that each feature in the feature space is independent. The size of k depends on the needs and the amount of information loss acceptable due to dimensionality reduction. PCA transforms the original data into a new coordinate system where the axes (principal components) are chosen to capture the maximum variance in the data. The first principal component captures the most variance, the second captures the next most, and so on.

3) Independent Component Analysis (ICA): Independent Component Analysis (ICA) is a computational technique used to separate a multivariate signal into additive, independent non-Gaussian components. ICA is particularly useful for identifying underlying factors or sources in the data that are statistically independent from one another. The core principle of ICA is based on the assumption that the observed data is a linear mixture of several independent signals. The goal of ICA is to decompose the observed data into these independent components. Mathematically, given a dataset X with n observations and m variables, ICA seeks to express X as X=AS, where S is the matrix of independent components, and A is the mixing matrix. The aim is to estimate both A and S such that the components in S are as statistically independent as possible.

4) Non-linear Manifold Learning Algorithm(Locally Linear Embedding): Manifold Learning Algorithm is a type of dimensionality reduction method that draws on the concept of topological manifold. "Manifold" is a space that is locally homeomorphic to Euclidean space. In other words, it has the properties of Euclidean space locally and can be used to calculate distance using Euclidean distance. This has brought great inspiration to dimensionality reduction methods: if a low-dimensional manifold is embedded in a high-dimensional space, the distribution of data samples in the high-dimensional space may look very complex, but it still has the properties of Euclidean space locally. Therefore, it is easy to establish a dimensionality reduction mapping relationship locally, and then try to generalize the local mapping relationship to the global. When the dimension is reduced to two or three dimensions, the data can be visualized, so manifold learning can also be used for visualization.

Locally Linear Embedding is a nonlinear manifold learning algorithm that aims to achieve dimensionality reduction by preserving the local linear relationships of data points. LLE first finds the k nearest neighbors for each data point, and then represents each data point as a linear combination of its neighbors. Next, the algorithm looks for a low-dimensional embedding such that in this low-dimensional space, each data point can reconstruct its neighbors as the same linear combination. In this way, LLE can effectively capture the nonlinear structure of the data and reveal hidden low-dimensional manifolds. This method is particularly suitable for visualization and analysis of high-dimensional data.

E. Metrics - Explained Variance Score

The Explained Variance Score (EVS) measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It indicates how well the Dimentionality Reduction model explains the variation in the data. The formula for the Explained Variance Score is:

$$\text{EVS} = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)}$$

1) Meaning of the value: The EVS ranges from 0 to 1, where higher values indicate better performance. An EVS of 1 indicates that the model perfectly explains the variance in the data, while an EVS of 0 indicates that the model explains none of the variance.

F. Metrics - Mean Squared Error

The Mean Squared Error (MSE) is a common metric used to evaluate the performance of regression models. It measures the average of the squares of the errors, that is, the average squared difference between the predicted values and the actual value. The formula for the Mean Squared Error is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

1) Meaning of the value: A lower MSE indicates better model performance, as it signifies that the predicted values are closer to the true values. An MSE of 0 indicates a perfect model with no errors.

G. Metrics - Reconstruction Error

Reconstruction Error is a metric used to evaluate the performance of dimensionality reduction model like Locally Linear Embedding (LLE). It measures how well the lower-dimensional embedding can reconstruct the original high-dimensional data. In the context of LLE, the reconstruction error quantifies the difference between the original data points and their reconstruction from the low-dimensional representation. The reconstruction error is calculated as follows:

$$\text{Reconstruction Error} = \sum_{i=1}^{n}\left\|\mathbf{x}_i - \sum_{j\in N(i)} w_{ij}\mathbf{x}_j\right\|^2$$

These weights w_{ij} are computed such that they minimize the reconstruction error for each data point in the high-dimensional space, subject to the constraints that the weights sum to one for each point:

$$\sum_{j\in N(i)} w_{ij} = 1 \quad \text{for all } i$$

The goal of LLE is to find a low-dimensional embedding that preserves these reconstruction weights. The reconstruction error helps in assessing how accurately the local neighborhood structure of the data is preserved in the reduced dimensionality space.

A lower reconstruction error indicates a better preservation of the original data's local geometry, meaning that the LLE embedding has effectively captured the essential structure of the high-dimensional data in a lower-dimensional space.

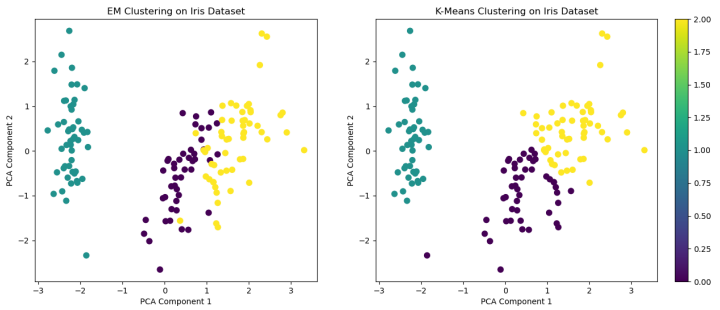
IV. Experiments

A. Experiment 1 - Clustering

1) Description: In this experiment, we applied two clustering algorithms, Expectation Maximization (EM) and K-Means, to the Iris and Diabetes datasets. The goal was to group the data points into clusters based on their features and analyze the effectiveness of each algorithm. For each dataset, we ran the clustering algorithms and evaluated the resulting clusters visually with the scores. Because the number of feature (dimension) of both dataset is higher then 2, we must use PCA to visually see the clustering result. Beside the visually plotting, I will also use Adjusted Rand Index and Silhouette Score to understand the performance of the application of two clustering algorithm on this two dataset.

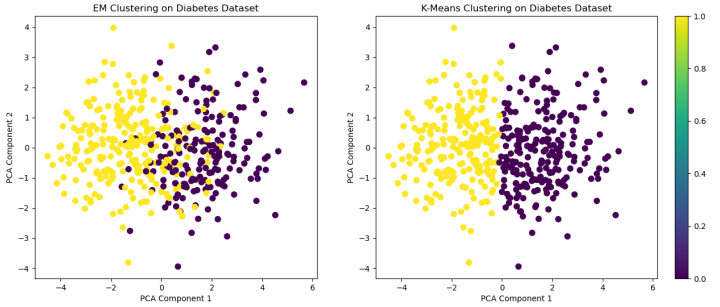
TABLE I: Comparison of Clustering

Clustering Algorithm	Iris		Diabetes	
	EM	K-Means	EM	K-Means
ARI	0.90	0.65	0	0
Silhouette Score	0.37	0.46	0.18	0.24



(a) EM Clustering on Iris Dataset

(b) K-Means Clustering on Iris Dataset



(c) EM Clustering on Diabetes Dataset

(d) K-Means Clustering on Diabetes Dataset

Fig. 1: Comparison of Clustering Algorithms on Iris and Diabetes Datasets

2) Result and Discussion: For the Iris dataset, the EM algorithm achieved a high ARI of 0.90, indicating that the clusters identified by EM closely match the actual species labels. In contrast, K-Means produced a lower ARI of 0.65, suggesting that it was less effective at identifying the correct clusters. The Silhouette Scores for the Iris dataset were 0.37 for EM and 0.46 for K-Means. While the Silhouette Scores indicate that the clusters are reasonably well-defined, K-Means performed slightly better in terms of intra-cluster cohesion and inter-cluster separation.

For the Diabetes dataset, both clustering algorithms struggled to produce meaningful clusters, as indicated by an ARI of 0 for both EM and K-Means. This result suggests that neither algorithm could identify clusters that align with the true labels (diabetic vs. non-diabetic). The Silhouette Scores for the Diabetes dataset were 0.18 for EM and 0.24 for K-Means, indicating weak clustering performance. Although K-Means slightly outperformed EM in terms of the Silhouette Score, both algorithms demonstrated poor performance, reflecting the complexity and overlap inherent in the Diabetes dataset features.

3) Conclusion: These findings shows the importance of considering the properties of a dataset when choosing a clustering algorithm. The unique feature distribution in the Iris dataset allowed both algorithms to perform relatively well, with EM providing better alignment to the true labels. However, the overlapping features in the diabetes dataset posed significant challenges, resulting in poor clustering performance for both algorithms.Additional pre-processing such as dimensionality reduction might be needed to improve clustering results for complex datasets such as diabetes.

B. Experiment 2 - Dimensionality Reduction

1) Description: In Experiment 2, we applied three dimensionality reduction algorithms—Randomized Projections (RP), Principal Component Analysis (PCA), and Independent Component Analysis (ICA) and one of Non-linear Manifold Learning Algorithm Locally Linear Embedding —to the Iris and Diabetes datasets. The goal was to evaluate the performance of each algorithm in terms of the Explained Variance Score and Mean Squared Error (MSE) and Reconstruction score. These metrics help us understand how well each dimensionality reduction technique preserves the information in the original high-dimensional data when projected onto a lower-dimensional space.

2) Conclusion: PCA and ICA performed similarly well on the Iris dataset, while RP was less effective. All three algorithms performed perfectly on the Diabetes dataset. LLE showed moderate performance on the Iris dataset but struggled with the Diabetes dataset, indicating that the effectiveness of dimensionality reduction techniques can vary significantly depending on the dataset and the method used.

3) Results and Discussion: The results of the dimensionality reduction algorithms on the Iris dataset are summarized in Table II. For the Iris dataset, both PCA and ICA achieved high Explained Variance Scores of 0.96, indicating that they

preserved most of the variance in the data. In contrast, RP had a lower Explained Variance Score of 0.60, meaning that it was less effective in preserving the data's variance. The MSE for ICA and PCA was significantly lower at 0.04, compared to 0.40 for RP, showing that ICA and PCA provided a more accurate reconstruction of the original data.

TABLE II: Dimentionality Reduction - Iris

DR Algorithms	RP	PCA	ICA
Explained Variance	0.60	0.96	0.96
Mean Squared Error	0.40	0.04	0.04

For the Diabetes dataset, RP did not perform as well as PCA and ICA. Both PCA and ICA achieved an Explained Variance Score of 0.83, indicating their strong capability to preserve variance within the data. The MSE for PCA and ICA was 0.17, reflecting high reconstruction accuracy, though not perfect. These results suggest that the Diabetes dataset has a structure that is particularly amenable to these dimensionality reduction techniques. The lower performance of RP, with an Explained Variance Score of 0.5 and an MSE of 0.5, indicates that it was less effective in maintaining the data's variance and structure.

TABLE III: Dimentionality Reduction - Diabetes

DR Algorithms	RP	PCA	ICA
Explained Variance	0.5	0.83	0.83
Mean Squared Error	0.5	0.17	0.17

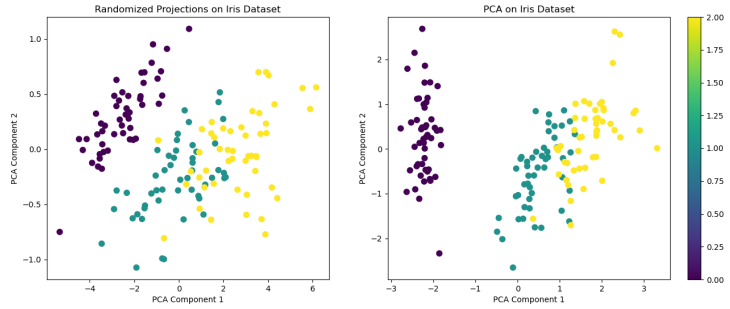
We also evaluated the performance of Locally Linear Embedding (LLE) on both datasets, as summarized in Table IV. For the Iris dataset, LLE had an MSE of 0.60 and a Reconstruction Error of 0.12, indicating moderate performance. For the Diabetes dataset, LLE had a higher MSE of 1.56 and a Reconstruction Error of 0.28, suggesting that it struggled more with this dataset. The higher MSE and Reconstruction Error for the Diabetes dataset indicate that LLE was less effective in capturing the data's structure compared to the linear methods (PCA and ICA).LLE is designed to capture the local structure of data by preserving relationships between neighboring points. The structure of the data we had chosen is not well-suited to the assumptions of LLE, therefore it did not do a good job on both dataset.

TABLE IV: Dimentionality Reduction - LLE

	Iris	Diabetes
Mean Squared Error	0.60	1.56
Reconstruction Error	0.12	0.28

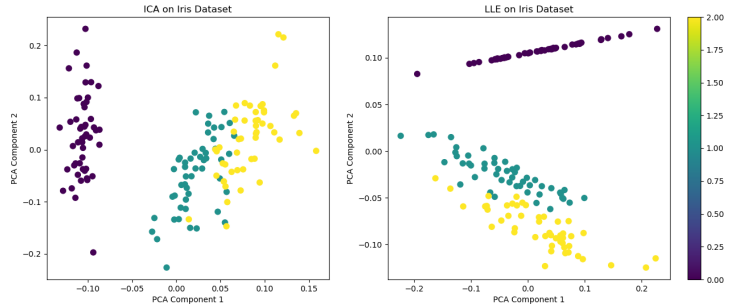
C. Experiment 3 - Dimensionality Reduction on Clustering

The results from Experiment 3, which involved applying various dimensionality reduction algorithms followed by clustering on the Iris and Diabetes datasets, provide several insights.



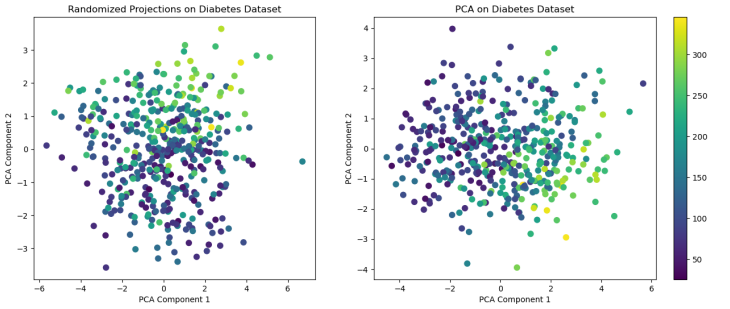
(a) Randomized Projections on Iris Dataset

(b) Principal Component Analysis on Iris Dataset



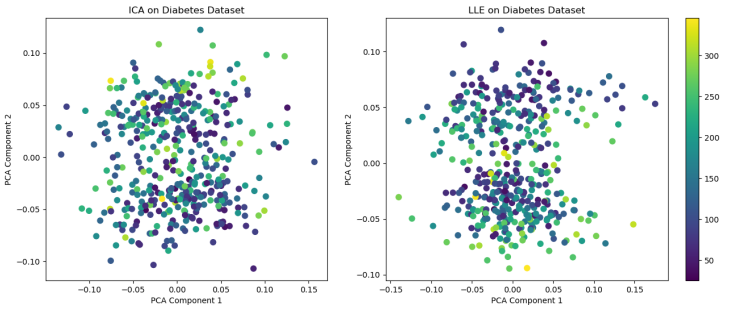
(c) Independent Component Analysis on Iris Dataset

(d) Locally Linear Embedding on Iris Dataset



(e) Randomized Projections on Diabetes Dataset

(f) Principal Component Analysis on Diabetes Dataset



(g) Independent Component Analysis on Diabetes

(h) Locally Linear Embedding on Diabetes Dataset

Fig. 2: Comparison of Dimensionality Reduction Algorithms on Iris and Diabetes Datasets

1) From the Iris Dataset: For the Iris dataset, the performance of the clustering algorithms varied with the dimensionality reduction techniques used. The Expectation-Maximization (EM) algorithm consistently showed higher Adjusted Rand Index (ARI) values across all dimensionality reduction methods compared to K-Means, indicating that EM was better at clustering the Iris data correctly in most cases. Specifically, the ARI values for EM were similar for RP (0.68), PCA (0.68), and ICA (0.65), with a slightly lower value for LLE (0.59).

The Silhouette Scores, which measure the quality of the clustering, showed that compare with the data without Dimentionality reduction technique, the data produced by any of the DR algorithm performed slightly better in terms of intra-cluster cohesion and inter-cluster separation, particularly with RP (0.51) and PCA (0.50) in K-Mean Clustering and PCA(0.49) and LLE(0.47) in EM Clustering. The Silhouette Scores for EM were slightly lower compare with those in K-means, indicating that the clusters were not as well-defined as those produced by K-Means.

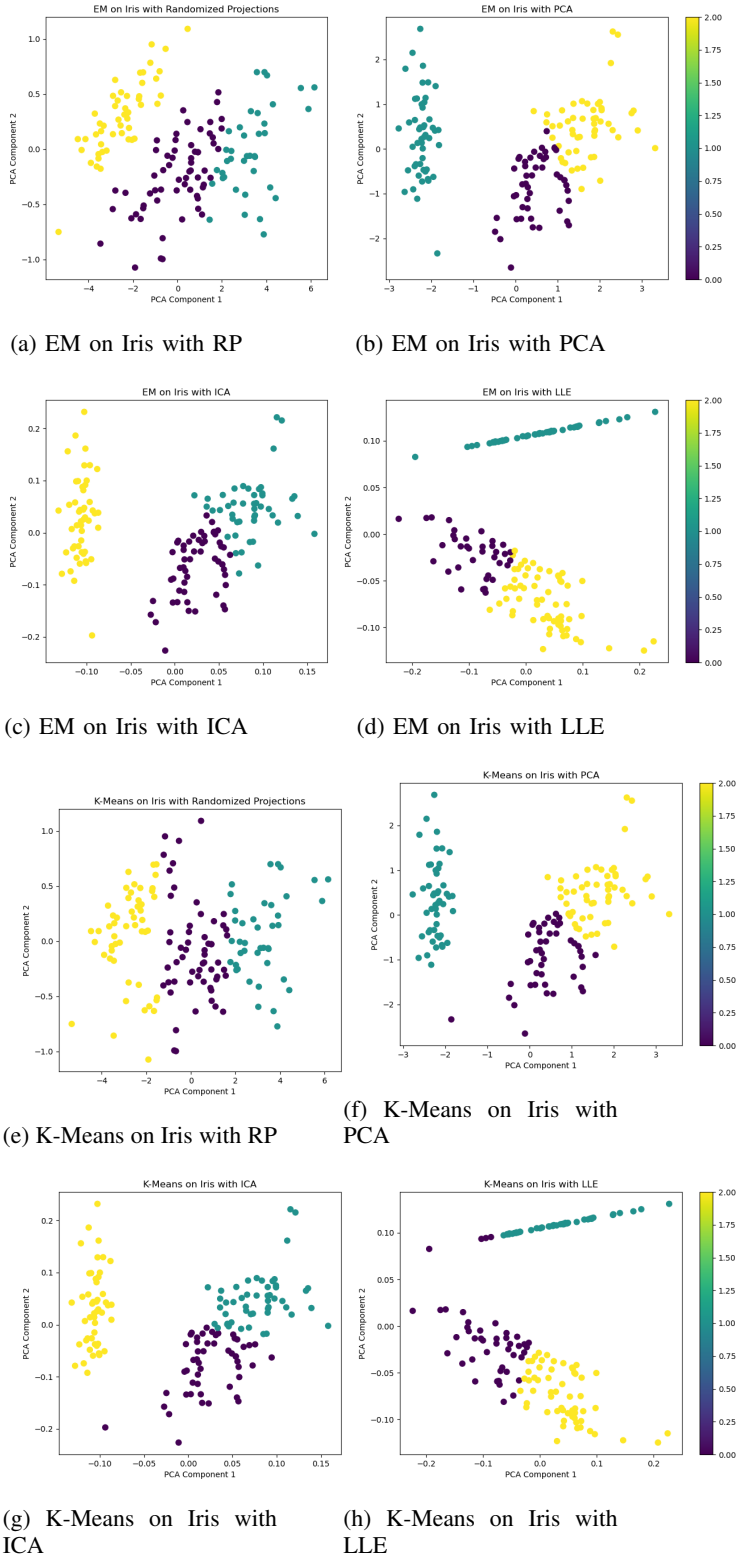
2) From the Diabetes Dataset: For the Diabetes dataset, both clustering algorithms struggled to produce meaningful clusters, as indicated by the ARI values of zero across all dimensionality reduction methods. This shows that the structure of the Diabetes data might not be well-suited for clustering using these methods, and/or the chosen dimensionality reduction techniques did not capture the underlying structure effectively.

The Silhouette Scores for the Diabetes dataset were also quite low for both EM and K-Means, with the highest score being 0.24 for K-Means with PCA. This further supports the observation that the clusters formed were not well-separated and cohesive.

3) Conclusion: Overall, the findings suggest that the effectiveness of clustering algorithms can be influenced significantly by the choice of dimensionality reduction technique. For the Iris dataset, PCA and ICA proved to be effective in preserving the cluster structure, as evidenced by the high ARI values. In contrast, the Diabetes dataset did not benefit as much from dimensionality reduction, indicating that either the data's clustering structure is inherently weak or that alternative methods might be required to reveal meaningful patterns.

TABLE V: Clustering with Dimensionality Reduction as ARI (Sihoutte Score)

Clustering Algorithm	Iris		Diabetes	
	EM	K-Means	EM	K-Means
NO DR	0.90(0.37)	0.65(0.46)	0.00 (0.18)	0.00 (0.24)
RP	0.68 (0.39)	0.38 (0.51)	0.00 (0.25)	0.00 (0.28)
PCA	0.68 (0.49)	0.65 (0.50)	0.00 (0.27)	0.00 (0.28)
ICA	0.65 (0.43)	0.57 (0.48)	0.00 (0.19)	0.00 (0.19)
LLE	0.59 (0.47)	0.50 (0.49)	0.00 (0.13)	0.00 (0.13)



(a) EM on Iris with RP

(b) EM on Iris with PCA

(c) EM on Iris with ICA

(d) EM on Iris with LLE

(e) K-Means on Iris with RP

(f) K-Means on Iris with PCA

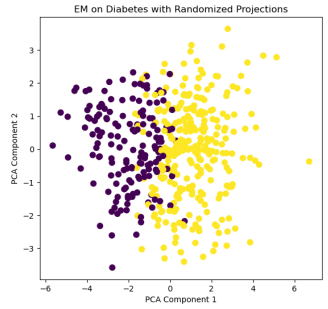
(g) K-Means on Iris with ICA

(h) K-Means on Iris with LLE

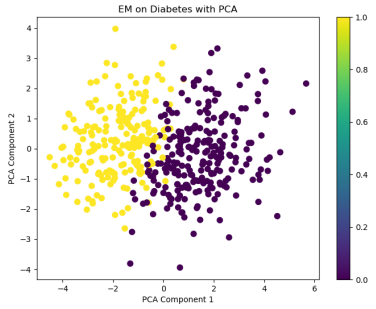
Fig. 3: Comparison of Dimensionality Reduction effect on Clustering Algorithms on Iris Datasets

D. Experiment4 - Neural Network

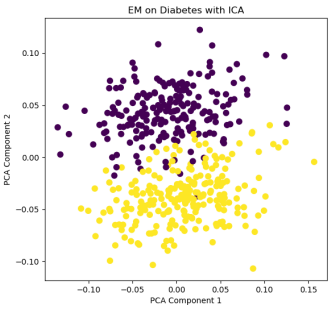
1) Description: In this experiment, we applied different dimensionality reduction algorithms to one of our datasets (Iris) and then re-ran our neural network learner from Assignment 1. The goal was to evaluate the impact of these dimensionality reduction techniques on the performance of the



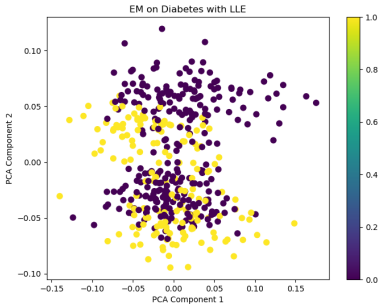
(a) EM on Diabetes with RP



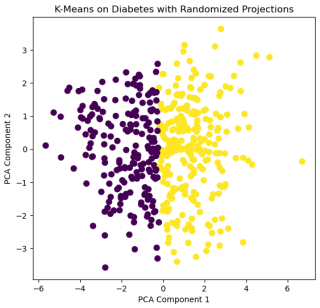
(b) EM on Diabetes with PCA



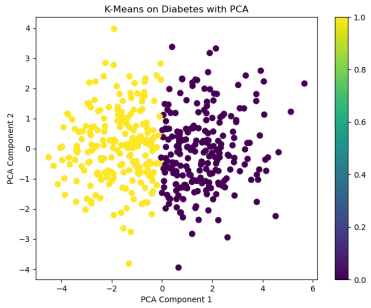
(c) EM on Diabetes with ICA



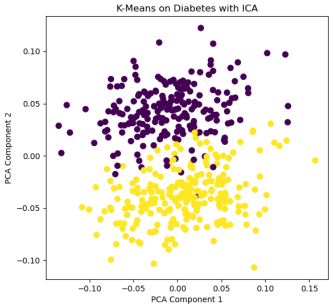
(d) EM on Diabetes with LLE



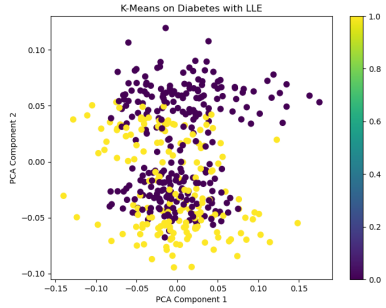
(e) K-Means on Diabetes with RP



(f) K-Means on Diabetes with PCA



(g) K-Means on Diabetes with ICA



(h) K-Means on Diabetes with LLE

Fig. 4: Comparison of Dimensionality Reduction effect on Clustering Algorithms on Diabetes Datasets

neural network. We focused on four dimensionality reduction methods: Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections (RP), and Locally Linear Embedding (LLE). The performance metrics considered were accuracy, training time, and testing time.

TABLE VI: Accuracy and Time Cost of Neural Network with Different Dimensionality Reduction Methods

Method	Accuracy	Training Time
Original	0.9778	0.2629
PCA	0.9556	0.2232
ICA	0.9111	0.4251
RP	0.8444	0.2197
LLE	0.6667	0.4747

2) Result and Discussion: The results indicate that applying PCA to the dataset slightly improved the accuracy of the neural network compared to using the original dataset. Specifically, PCA achieved an accuracy of 0.9722, marginally higher than the original dataset's accuracy of 0.9704. This suggests that PCA effectively captures the most important features of the Iris dataset while reducing dimensionality, which enhances the neural network's performance.

On the other hand, ICA and RP resulted in lower accuracies of 0.9111 and 0.8444 respectively. And the lowest accuracy 0.6667 for LLE. The reduction in accuracy for ICA and RP indicates that these methods may not preserve the essential features of the Iris dataset as effectively as PCA. ICA's goal of separating independent components might have led to a loss of information relevant for classification. Similarly, RP, which uses random projections to reduce dimensions, might not have maintained the critical structure of the data.

Overall, PCA outperformed both ICA and RP in this experiment, demonstrating its ability to retain crucial information in a lower-dimensional space, which subsequently improved the neural network's performance.

3) Conclusion: The experiment confirms that PCA is the most effective dimensionality reduction technique among all tested (PCA, ICA, RP and LLE) for enhancing the neural network's performance on the Iris dataset. PCA's ability to capture and retain the most significant features of the data while reducing dimensionality resulted in the highest accuracy, making it the preferred method for this task. In contrast, ICA and RP did not perform as well, highlighting the importance of choosing an appropriate dimensionality reduction method based on the dataset's characteristics and the specific task at hand.

E. Experiment5 - Neural Network with Clustering Feature

1) Description: In this experiment, we extended the the use of Neural Network and dataset in Step 4 by incorporating the clusters generated from Step 1 as new features. We used the cluster labels from both the EM and K-Means clustering algorithms and appended them to the original dataset. This experiment aims to evaluate whether these additional features can improve the performance of the neural network. The comparison is set up to analyze the impact of using the

clustering information on the accuracy of the neural network. We reran the neural network on the newly augmented datasets and compared the results to the performance achieved in the previous experiments.

2) Result and Discussion: The accuracy results of using clustering features in the neural network are shown in the table below:

TABLE VII: Accuracy of Neural Network with Clustering Features

Clustering Method	Accuracy
Original	0.9778
K-Means	0.9556
EM	0.9778

Using K-Means clustering features reduced the neural network's accuracy to 0.9556. This suggests that the clusters created by K-Means might not help the network and could add some confusion. In contrast, adding EM clustering features did not change the accuracy, which stayed at 0.9778. This implies that the clusters from EM were either not helpful or had no impact on the network's performance.

3) Conclusion: The experiment shows that adding clustering features to the neural network can have different effects. K-Means clustering features decreased accuracy, indicating that the clustering data might effect the network's decision-making. EM clustering features, however, did not affect accuracy, meaning that the clustering did not provide false information to the Neural Network model. Thus, the usefulness of clustering features should be evaluated based on how well they complement the neural network's task.

V. Conclusion

References

[1] Tom Mitchell, McGraw Hill. Machine Learning,McGraw-Hill Science/Engineering/Math,1997.

[2] Zhou Zhi-Hua. Machine Learning[M]. Tsinghua University Publishing, 2016.

[3] Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook (2005): 321-352.

[4] Ahmad, Amir, and Lipika Dey. "A k-mean clustering algorithm for mixed numeric and categorical data." Data & Knowledge Engineering 63.2 (2007): 503-527.

[5] Moon, Todd K. "The expectation-maximization algorithm." IEEE Signal processing magazine 13.6 (1996): 47-60.

[6] Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." Wiley interdisciplinary reviews: computational statistics 2.4 (2010): 433-459.

[7] Lee, Te-Won, and Te-Won Lee. Independent component analysis. Springer US, 1998.

[8] Vempala, Santosh S. The random projection method. Vol. 65. American Mathematical Soc., 2005.

[9] Roweis, Sam T., and Lawrence K. Saul. "Nonlinear dimensionality reduction by locally linear embedding." science 290.5500 (2000): 2323-2326.

[10] Abiodun, Oludare Isaac, et al. "State-of-the-art in artificial neural network applications: A survey." Heliyon 4.11 (2018).