Machine Learning
Unsupervised Learning and Dimensionality Reduction

Jorge Salvador Aguilar Moreno

***Abstract*—This report presents the implementation and exploration of two clustering algorithms—EM and K-Means—and three linear dimensionality reduction algorithms: RP, PCA and ICA. These algorithms are applied to two different datasets. The exploration process involves (1) applying each clustering and dimensionality reduction technique individually on the datasets, (2) executing clustering on dimensionality-reduced versions of the datasets, and (3) training neural network learners on transformed input spaces derived from the application of clustering and dimensionality reduction algorithms.**

***Index Terms*—unsupervised learning, dimensionality reduction, machine learning, expectation maximization, k-means, randomized projections, principal component analysis, independent component analysis.**

I. Introduction

Unsupervised learning is a branch of machine learning where the algorithm identifies patterns or structures within unlabeled data, revealing insights without pre-specified labels or categories. Unsupervised learning techniques can be broadly categorized into *clustering*, *association*, and *dimensionality reduction* methods. In this report, we focus on two clustering algorithms—Expectation Maximization (EM) and K-Means—and three dimensionality reduction techniques—Randomized Projections (RP), Principal Component Analysis (PCA), and Independent Component Analysis (ICA).

Clustering algorithms group data points based on similarity, enabling the identification of inherent patterns within the dataset. This study explores the following clustering techniques:

- **Expectation Maximization (EM):** EM is a probabilistic clustering algorithm that iteratively refines a model of the data distribution. It assumes data points are generated by a mixture of Gaussian distributions and adjusts the parameters to maximize the likelihood of the observed data. Unlike K-Means, EM can capture more complex and overlapping cluster shapes due to its probabilistic approach, making it suitable for diverse data distributions [1].
- **K-Means Clustering:** This algorithm partitions the dataset into a predefined number of clusters by iteratively minimizing the variance within each cluster. Each cluster is represented by its centroid, and data points are assigned to the cluster with the nearest centroid. K-Means is computationally efficient but assumes spherical clusters of similar size, which can limit its effectiveness with complex datasets [2].

Dimensionality reduction algorithms simplify high-dimensional data by projecting it into a lower-dimensional space, often retaining essential features while eliminating redundancy and noise. The three methods explored in this study are as follows:

- **Randomized Projections (RP):** RP reduces dimensionality by projecting data onto a randomly generated lower-dimensional subspace, effectively preserving distances between data points with minimal computational cost. RP is particularly efficient with very high-dimensional datasets and provides comparable results to more computationally intensive techniques [3].
- **Principal Component Analysis (PCA):** PCA projects data onto the directions (principal components) that capture the maximum variance. The algorithm identifies orthogonal axes and ranks them according to variance, retaining the components that capture the majority of data variability. PCA is widely used for its interpretability and ability to reduce dimensionality while retaining meaningful patterns in the data [2].
- **Independent Component Analysis (ICA):** ICA seeks to identify statistically independent components within the data, making it effective for separating mixed signals and identifying hidden factors. Unlike PCA, which captures variance, ICA focuses on independence, making it useful in applications where latent factors are not necessarily aligned with maximum variance [4].

The use of clustering or dimensionality reduction techniques as a preprocessing step for neural networks can enhance model performance significantly. Clustering algorithms, by organizing data into distinct groups, can reveal latent structures that help neural networks identify underlying patterns in complex datasets. Dimensionality reduction techniques improve the efficiency of neural networks by reducing the number of input features, thereby accelerating training time and reducing the risk of overfitting [3].

II. Datasets Overview

The following datasets served as the basis for implementing the unsupervised learning algorithms in this project:

1) **Concrete compressive strength dataset** (UCI Machine Learning Repository). This dataset contains 1,030 instances and 8 features: cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age.

2) **Diabetes dataset** (Kaggle). This dataset contains 768 instances and 8 features: pregnancies, glucose level, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age.

These datasets are interesting for unsupervised learning as they feature distinct data structures—binary in the diabetes dataset and continuous in the concrete dataset—allowing for exploration of clustering algorithms in both settings. Both datasets involve complex feature interactions, where the diabetes and material strength features contribute variably to groupings, providing insight into the algorithms' ability to capture latent structures. Additionally, sourced from real-world applications, these datasets contain inherent noise and variability, presenting a realistic setting for testing the robustness of clustering and dimensionality reduction methods.

III. Clustering Algorithms on Original Datasets

The first exploration involves applying the EM and K-Means clustering algorithms on the original diabetes and concrete datasets. For each dataset, the optimal number of clusters is determined by evaluating several clustering metrics. The results from this analysis is subsequently used in Section VII: "Neural Networks on Clustered Datasets." Additionally, a runtime comparison between the EM and K-Means algorithms is conducted at the end of this section, providing insights into the computational efficiency of each clustering approach across datasets and cluster counts.

A. Hypothesis

It is hypothesized that for all algorithms, the optimal number of components for the diabetes dataset will be 2, given its binary classification nature. In contrast, for the concrete dataset, a larger number of components is anticipated, as this dataset originally involves continuous labels, better suited to regression tasks in supervised learning. Regarding runtime performance, K-Means is expected to be faster and less computationally demanding than EM, especially as the number of components increases. This is because K-Means has the simpler objective of minimizing the sum of squared distances to cluster centers; whereas EM involves probability and covariance matrix calculations, which is more computationally intensive.

B. Experimental Methodology

To determine the optimal number of components for each clustering algorithm and dataset, a consistent methodology is applied. Each dataset is split into training and testing sets, but only the training data is used since this is an unsupervised learning task. The features are scaled, and different random states are used to capture model variability. The model is trained across a range of component values from 2 to 100. For EM, metrics such as BIC, AIC, silhouette score, and log-likelihood are evaluated, while for K-Means, metrics include inertia, silhouette score, VRC, and Davies-Bouldin index. These metrics are selected for the following reasons:

- **Bayesian Information Criterion (BIC)**: Metric balances goodness of fit with model complexity. Lower values indicate better models, as it penalizes models with more components.
- **Akaike Information Criterion (AIC)**: AIC penalizes the number of parameters to prevent overfitting. Lower AIC values suggest a better model fit.
- **Silhouette Score**: It assesses how similar an object is to its own cluster compared to other clusters, providing an evaluation of cluster separation. A higher silhouette score indicates more distinct and well-separated clusters.
- **Log-Likelihood**: It measures the likelihood of the data given a particular clustering model. For EM, higher log-likelihood values indicate a better fit of the model to the data.
- **Inertia**: Inertia measures the sum of squared distances between data points and their respective cluster centers. Lower inertia values signify tighter clusters, meaning points are closer to the centroid, which is favorable for K-Means clustering.
- **Variance Ratio Criterion (VRC)**: VRC assesses the ratio of between-cluster variance to within-cluster variance. Higher VRC values suggest better clustering, with more defined separation between clusters.
- **Davies-Bouldin Index**: It evaluates cluster quality by measuring the average similarity between clusters. Lower values indicate better separation, with less overlap between clusters.

A similar experiment is conducted to determine the runtime of each clustering algorithm and dataset as a function of the number of components.

C. Analysis and Results

Figure 1 and Figure 2 display the BIC, AIC, Silhouette score, and log-likelihood metrics obtained from applying EM to the diabetes and concrete strength datasets, respectively.

For the diabetes dataset, the optimal number of components is considered to be 2. This choice is supported by the highest silhouette score of 0.1617, which drops significantly after 3 or more components. While BIC, a metric that penalizes model complexity, indicates a minimum score of 8,166 at $n_{\text{components}} = 4$, and the AIC and log-likelihood exhibit slightly better results for this component. It was preferred to have well-separated clusters, as supported by the silhouette score.

For the concrete strength dataset, a clear optimal number of components was not readily apparent across all metrics. BIC suggested an optimal point at $n_{\text{components}} = 74$, with a minimum value of -8761. Other metrics, however, indicated that increased model complexity (i.e., a higher number of components) yielded improved results. Nevertheless, the rate of improvement became marginal as the slope of these curves flattened with more components. To balance model performance with the risk of overfitting, we selected $n_{\text{components}} = 74$ as an appropriate choice.

Figure 3 and Figure 4 display the Inertia, Silhouette score, VRC, and Davies-Bouldin index metrics obtained from ap-

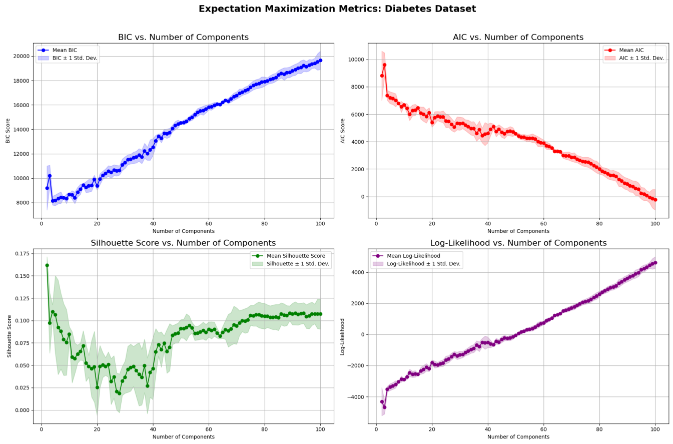

Fig. 1. EM on Diabetes Dataset: BIC (upper left), AIC (upper right), Silhouette Score (lower left) and Log-likelihood (lower right)

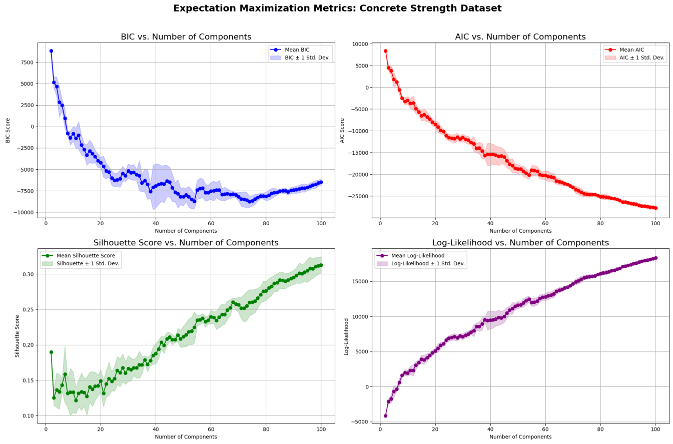

Fig. 2. EM on Concrete Dataset: BIC (upper left), AIC (upper right), Silhouette Score (lower left) and Log-likelihood (lower right)

plying the K-Means algorithm to the diabetes and concrete strength datasets, respectively.

For the diabetes dataset, the optimal number of components was determined to be 3, which contrasts with the initial hypothesis of 2 components due to the binary classification nature of the dataset. This choice is supported by the highest Silhouette score of 0.1864 and the VRC, which reached its maximum value of 122 at this point. While the elbow method using the inertia curve and the Davies-Bouldin index suggested a higher number of components, the most significant improvement was observed from 2 to 3 components, indicating that $n_{\text{components}} = 3$ is the best option.

For the concrete strength dataset, selecting an optimal number of components was less straightforward, as the metrics did not converge on a single value. The elbow method using inertia suggested an optimal point around 10 components, while both the Silhouette score and the Davies-Bouldin index favored the maximum value in the analysis, 100 components. The VRC suggested 2 or 4 as suitable options. Ultimately, 100 components were selected, as the Silhouette score is the preferred metric for achieving well-separated clusters, aligning with other experimental results.

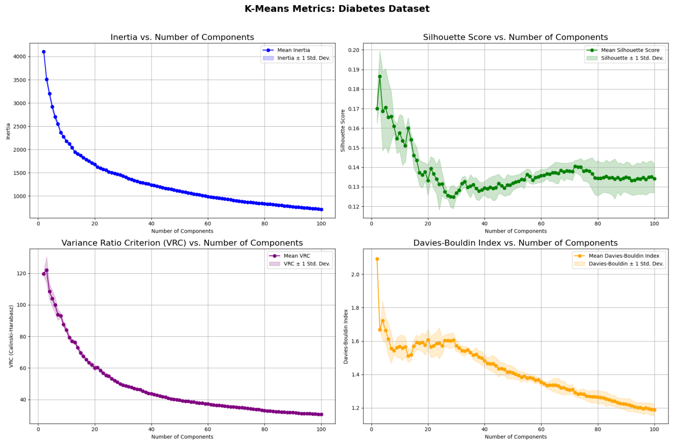

Fig. 3. K-Means on Diabetes Dataset: Inertia (upper left), Silhouette Score (upper right), VRC (lower left) and Davies-Bouldin Index (lower right)

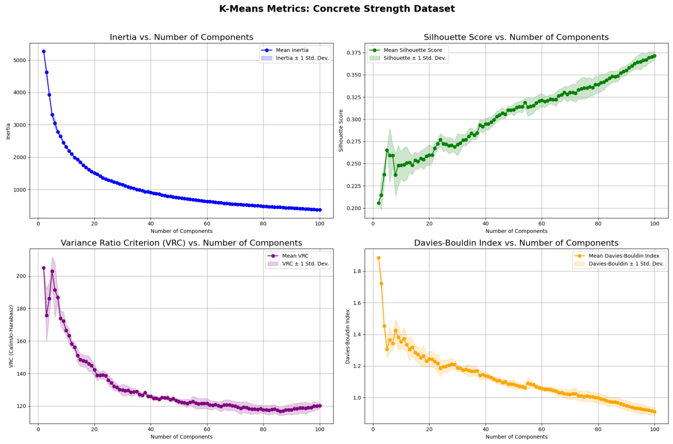

Fig. 4. K-Means on Concrete Dataset: Inertia (upper left), Silhouette Score (upper right), VRC (lower left) and Davies-Boulding Index (lower right)

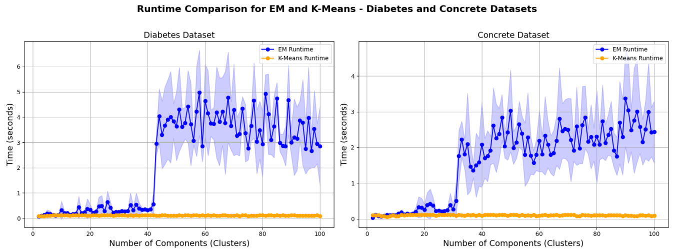

Fig. 5. Runtime Comparison for EM and K-Means Clustering Algorithms: Diabetes dataset (left) and Concrete dataset (right)

In Figure 5, a runtime comparison between the Expectation Maximization (EM) and K-Means clustering algorithms is displayed for both the diabetes and concrete strength datasets. For both datasets, K-Means exhibits significantly lower runtimes across all component counts, maintaining a relatively flat trend as the number of clusters increases. This efficiency is expected, as K-Means is computationally simpler, focusing only on minimizing the distance to cluster centers. In contrast, EM is more computationally intensive due to its iterative probability and covariance matrix calculations. As the number of components increases, EM's runtime shows a notable rise, with greater variability in time, particularly for component counts above 40. This trend underscores EM's sensitivity to model complexity, making K-Means the preferred choice when computational efficiency is a priority.

The clusters obtained from the diabetes and concrete strength datasets ”make sense” and reveal insights about the underlying structure of each dataset, though the alignment with natural groupings varies by dataset and algorithm. For the diabetes dataset, the optimal cluster count for EM was 2 components, likely reflecting the dataset's binary classification nature, as supported by the high silhouette score for this configuration. However, K-Means indicated an optimal 3 components, suggesting that there may be subtle subgroupings within the diabetic and non-diabetic classes that this algorithm detected. In contrast, for the concrete strength dataset, the optimal number of clusters was less clear. EM suggested 74 components based on BIC, while K-Means metrics pointed to a higher number of components, with silhouette scores peaking at 100, indicating that the data may not naturally partition into a small number of well-separated clusters due to the continuous nature of concrete strength. The choice of metrics also impacted these results; for example, silhouette score emphasized cluster separation, whereas BIC penalized complexity, favoring fewer clusters. To improve clustering results, modifications such as feature engineering to capture latent structure or hybrid models combining clustering with dimensionality reduction could be considered. Finally, the nature of these datasets influenced performance, with the diabetes dataset aligning more closely with clustering expectations than the continuous-label concrete dataset, which does not lend itself as naturally to discrete cluster formation.

IV. Dimensionality Reduction Algorithms on Original Datasets

This exploration applies the RP, PCA and ICA linear dimensionality reduction algorithms on the original diabetes and concrete datasets. For each dataset, the optimal number of clusters is identified through evaluation of various metrics. These dimensionality-reduced representations are used in Section VI: ”Neural Networks on Dimensionality-Reduced Datasets.” A runtime comparison of the RP, PCA and ICA algorithms is also included to highlight their computational efficiency across datasets and component counts.

A. Hypothesis

In this exploration, it is hypothesized that PCA will perform best in terms of preserving variance, as it is specifically designed to maximize explained variance in lower-dimensional representations by identifying principal components. This should result in more effective dimensionality reduction for both the diabetes and concrete datasets, as PCA's variance-maximizing approach aligns well with capturing key data patterns. ICA is expected to excel in creating components that are statistically independent, making it potentially advantageous for datasets where independent feature extraction is critical, though it may be sensitive to initialization and require more computation than PCA. RP, being a randomized method, is anticipated to be the fastest due to its lower computational complexity and lack of iterative calculations. However, RP may introduce more approximation error, especially for lower component counts, since it does not prioritize variance or independence directly. Therefore, we expect PCA to yield the most interpretable results, ICA to produce components with unique structure, and RP to outperform both in terms of runtime efficiency.

B. Experimental Methodology

To determine the optimal number of components for each dimensionality reduction algorithm and dataset, a consistent methodology is applied. Each dataset is split into training and testing sets, with only the training data used, as this is an unsupervised learning task. The features are scaled, and the model is trained across a range of component values from 2 to 5. Different metrics are evaluated for each algorithm: reconstruction error for RP; reconstruction error and explained variance ratio for PCA; and reconstruction error and kurtosis for ICA. These metrics are selected based on their relevance to each algorithm's objectives:

- **Reconstruction Error**: This metric measures the mean squared error between the original and reconstructed data, providing insight into how well the reduced components capture the original data's structure. Lower reconstruction error values indicate better approximations, making this metric relevant for all three algorithms.
- **Explained Variance Ratio (PCA-specific)**: This metric evaluates the amount of variance retained by the selected principal components in PCA. A higher explained variance ratio suggests that the principal components capture more of the dataset's original variability, indicating a more effective dimensionality reduction.
- **Kurtosis (ICA-specific)**: Kurtosis assesses the ”tailedness” or extremity of the distribution of independent components in ICA. Higher kurtosis values are preferred, as they suggest more non-Gaussian, independent sources have been extracted, which is the primary goal of ICA.

A similar experiment is conducted to evaluate the runtime of each dimensionality reduction algorithm on both datasets as a function of the number of components. This analysis provides insight into the computational efficiency of each method.

C. Analysis and Results

In Figure 6, RP shows a comparison of reconstruction errors across different numbers of components for both datasets. For both the diabetes and concrete datasets, the optimal number of components is identified as 3, yielding the lowest reconstruction errors. This suggests that 3 components capture sufficient variance in both datasets, making it a balanced choice between complexity and reconstruction accuracy.

Re-running the Random Projection (RP) algorithm with different random states resulted in minimal variation in reconstruction error across runs. This low variability suggests that RP reliably preserves data structure, making it consistent across runs and suitable for applications where stability is essential.

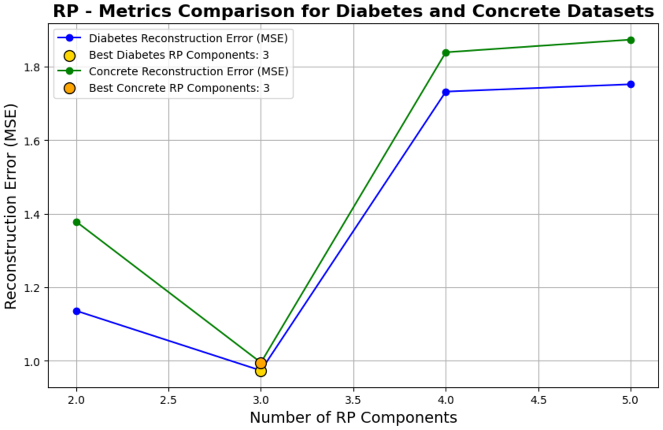

Fig. 6. Metrics Comparison in RP-Reduced Datasets: Reconstruction Error

In Figure 7, PCA is evaluated for both reconstruction error and explained variance ratio as a function of number of components. The Figure indicates that higher component numbers enhance data retention quality for both datasets, with 5 components being optimal for maximizing explained variance while keeping reconstruction error low.

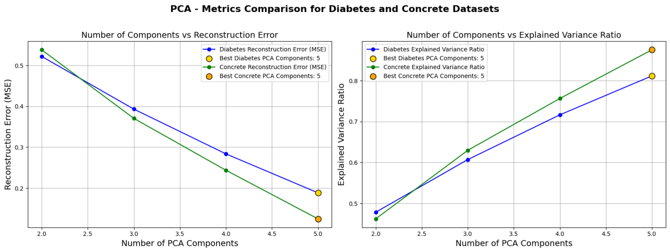

Fig. 7. Metrics Comparison in PCA-Reduced Datasets: Reconstruction Error (left) and Explained Variance Ratio (right)

In Figure 8, ICA is analyzed through reconstruction error and kurtosis. For both datasets, 4 components are considered optimal for achieving independent component structure, while 5 components minimize reconstruction error. Since a simpler model is preferred, 4 components is selected for analysis in other sections of this report.

Figure 9 illustrates the runtime comparison for RP, PCA, and ICA dimensionality reduction algorithms across different numbers of components. In both datasets, RP and PCA show relatively stable and low runtimes across component counts. ICA, however, exhibits a significant increase in runtime as the number of components reaches 5, especially in the concrete dataset. This indicates that while ICA provides independent components, it comes at a high computational cost, particularly with more components, making RP and PCA preferable choices for efficiency in larger datasets or real-time applications.

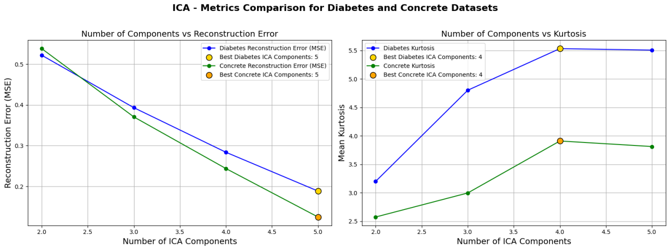

Fig. 8. Metrics Comparison in ICA-Reduced Datasets: Reconstruction Error (left) and Mean Kurtosis (right)

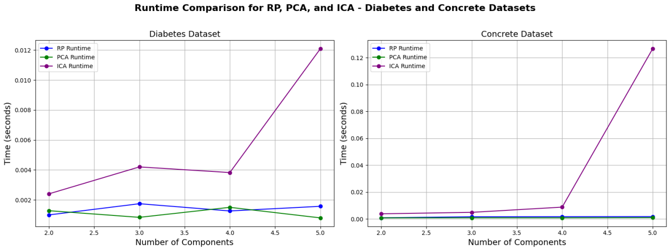

Fig. 9. Runtime Comparison for RP, PCA and ICA Dimensionality Reduction Algorithms: Diabetes dataset (left) and Concrete dataset (right)

The dimensionality reduction techniques reveal different perspectives on data structure. Using RP, the data is transformed into a lower-dimensional space while preserving pairwise distances to some extent, as indicated by the low reconstruction errors of approximately 1.0 with 3 components for both datasets. For PCA, the distribution of eigenvalues suggests that most variance is captured within the first few components. Figures illustrate that with 5 components, both datasets retain a substantial amount of explained variance (over 80%), and the reconstruction error is minimized, demonstrating PCA's effectiveness in capturing dominant linear patterns in the data. This eigenvalue distribution implies that the datasets have significant intrinsic dimensionality, but a few principal components sufficiently summarize their structure. In ICA, the kurtosis metric highlights the non-Gaussianity of independent components. The increasing kurtosis with additional components indicates that ICA successfully extracts statistically independent signals, particularly with 4 components. The independent components, characterized by higher kurtosis, suggest potentially meaningful latent variables. However, this independence comes at a computational cost, as seen in the runtime analysis, especially with more components in the concrete dataset. Given the nearly identical reconstruction error values for PCA and ICA, it suggests that both algorithms converge to a similar projection space, a convergence more likely to occur when using a small number of components. This similarity implies that both methods capture the most essential structure of the data effectively in lower-dimensional spaces.

The diabetes dataset likely has full rank equal to the number of original features, given typical settings; whereas the concrete dataset is also expected to have full rank, though principal components suggest that fewer dimensions capture the essential structure.

Noise affects dimensionality reduction algorithms in varying ways. RP is generally robust to noise due to its reliance on random linear combinations, which can mitigate random fluctuations. PCA is more sensitive, as it seeks directions of maximum variance, which noise can distort, potentially leading it to capture noisy features. ICA is the most affected, as it relies on independence assumptions that noise disrupts, making it harder to find meaningful components. Overall, RP is the least affected, PCA is moderately impacted, and ICA is the most sensitive to noise.

The diabetes and concrete datasets exhibit notable collinearity both qualitatively and quantitatively. Conceptually related features, such as blood pressure and BMI in the diabetes dataset or cement and fly ash in the concrete dataset, suggest overlapping information. Quantitatively, high correlations among certain feature pairs confirm this linear dependence. For example, glucose and insulin in diabetes or materials like cement and slag in concrete show moderate correlations, indicating redundancy. Calculating the variance inflation factor (VIF) further quantifies this, with higher values highlighting multi-collinearity where features may nearly be linear combinations of each other.

V. Clustering on Dimensionality-Reduced Datasets

In this exploration, the clustering algorithms are re-applied on the set of dimensionality reduction datasets. This will result in 12 combinations of results of datasets, dimensionality reduction, and clustering methods.

A. Hypothesis

The hypothesis for this exploration is that dimensionality reduction methods like PCA and ICA, which preserve key data characteristics, will improve clustering performance, with PCA expected to yield the best clustering input by maximizing variance. ICA may further enhance clustering in cases where independent features are critical. RP, while efficient, may perform less well due to its lack of variance or independence preservation. K-Means is expected to perform best on datasets with clear, spherical clusters, while EM may excel with more complex structures, though at a higher computational cost. Consequently, PCA combined with K-Means may yield the most effective results across datasets.

B. Experimental Methodology

This methodology applies dimensionality reduction first and then clustering to each dataset to identify optimal model configurations. After splitting each dataset into training and testing sets, only the training set is used for unsupervised learning. First, the data is standardized, and dimensionality reduction is applied with component values ranging from 2 to 5. For each transformed dataset, clustering is performed using EM or K-Means with component values from 2 to 10. The silhouette score is calculated for each clustering configuration to assess clustering quality, with the highest score identifying the best configuration for each specific dimensionality reduction setting.

Since silhouette scores are not directly comparable across different dimensionality reduction settings, a second metric is used to select the best-performing model overall: BIC is applied for EM and VRC for K-Means. This secondary metric evaluates each configuration to identify the optimal combination of dimensionality and clustering components. Finally, the best model is assessed with additional metrics, including AIC, silhouette score, and log-likelihood, for a comprehensive evaluation of the selected configuration.

Figure 10 provides a visual representation of the procedure for one specific combination: applying K-Means clustering on the RP-transformed diabetes dataset. This combination is part of the twelve configurations explored in these experiments. In this process, the silhouette score is first calculated for each combination of dimensionality reduction and clustering components. The top-performing models are then evaluated and compared using an additional metric to identify the optimal configuration.

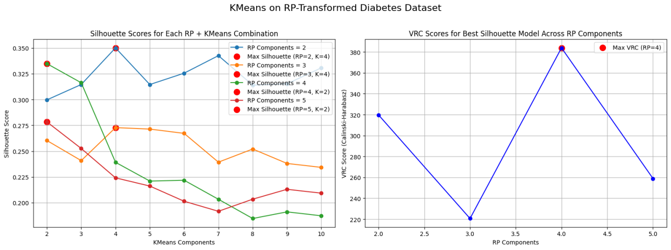

Fig. 10. K-Means clustering on RP-transformed diabetes dataset

C. Analysis and Results

The following tables present summary metrics for the optimal clustering algorithms applied to each dimensionality-reduced dataset, showcasing their performance across different configurations.

Based on Table I, the ICA-transformed data achieves the best overall results on the diabetes dataset when clustered with EM. ICA shows the lowest BIC (3378.63) and AIC (3330.01) scores, indicating a better balance between model complexity and fit quality compared to RP and PCA. Additionally, ICA yields the highest log-likelihood score (-1654.01), which suggests a closer fit of the model to the data. Although the silhouette score for RP (0.3661) is slightly higher than ICA (0.3623), the difference is minimal and does not outweigh the superior BIC, AIC, and log-likelihood values seen in the ICA model. Therefore, the model using ICA with 2 components and an EM configuration of 2 components is selected as the best-performing configuration.

TABLE I
EM ALGORITHM ON DIMENSION-REDUCED DIABETES DATASET

Metric	RP	PCA	ICA
Dim. Red. Components	2	2	2
EM Components	3	10	2
BIC Score	4117.95	4374.67	3378.63
AIC Score	4042.81	4113.89	3330.01
Silhouette Score	0.3661	0.3545	0.3623
Log-likelihood	-2004.41	-1997.95	-1654.01

Based on Table II. The PCA-transformed data with 2 components and 3 K-Means clusters shows the best overall performance of the K-Means algorithm on the dimension-reduced diabetes dataset. It has the highest silhouette score (0.38), indicating better-defined clusters, and the lowest Davies-Bouldin Index (0.8658), suggesting strong separation between clusters. Additionally, PCA achieves the highest Variance Ratio Criterion (449.13), meaning it retains the most variance in the reduced space. Inertia is lowest for ICA (766.12), but given the superior silhouette score, Davies-Bouldin Index, and variance retention with PCA, the PCA-based model is the best-performing configuration.

TABLE II
K-MEANS ALGORITHM ON DIMENSION-REDUCED DIABETES DATASET

Metric	RP	PCA	ICA
Dim. Red. Components	4	2	2
K-Means Components	2	3	2
Inertia	2486.40	955.81	766.12
Silhouette Score	0.33	0.38	0.38
Davies-Bouldin Index	1.0967	0.8658	1.0668
Variance Ratio Criterion	383.88	449.13	368.96

Table III presents the performance metrics of the EM algorithm on the dimension-reduced concrete dataset. Among these, the ICA-transformed data with 2 components and 4 EM clusters emerges as the best model. It has the lowest BIC (4494.19) and AIC (4385.77) scores, indicating a more parsimonious fit. Additionally, ICA yields the highest log-likelihood (-2169.88), suggesting a better model fit to the data. Although the silhouette score for ICA (0.4258) is slightly lower than RP's (0.4545), the other metrics strongly support ICA as the most effective configuration for the EM algorithm on this dataset.

TABLE III
EM ALGORITHM ON DIMENSION-REDUCED CONCRETE DATASET

Metric	RP	PCA	ICA
Dim. Red. Components	2	2	2
EM Components	2	5	4
BIC Score	6136.93	5494.65	4494.19
AIC Score	6085.07	5357.93	4385.77
Silhouette Score	0.4545	0.4281	0.4258
Log-likelihood	-3031.53	-2649.97	-2169.88

Table IV summarizes the performance of the K-Means algorithm on the dimension-reduced concrete dataset. Among these models, the RP transformation with 2 components and 2 K-Means clusters appears to be the best choice. This configuration yields the highest silhouette score (0.46), indicating better-defined and more separated clusters. Additionally, it has the highest Variance Ratio Criterion (958.71), suggesting stronger between-cluster separation relative to within-cluster cohesion. Although RP has a slightly higher Davies-Bouldin Index (0.8084) compared to ICA (0.7949), the superior silhouette score and Variance Ratio Criterion make RP the most favorable option for K-Means clustering on this dataset.

TABLE IV
K-MEANS ALGORITHM ON DIMENSION-REDUCED CONCRETE DATASET

Metric	RP	PCA	ICA
Dim. Red. Components	2	2	2
K-Means Components	2	4	3
Inertia	2270.80	873.57	620.35
Silhouette Score	0.46	0.43	0.43
Davies-Bouldin Index	0.8084	0.81	0.7949
Variance Ratio Criterion	958.71	690.63	680.21

When clustering experiments were reproduced on the datasets projected onto new spaces created by ICA, PCA, and RP, the clusters obtained were different from those on the original datasets. This variation is expected, as each dimensionality reduction technique produces distinct transformations that capture different aspects of the data. For example, PCA emphasizes variance by projecting data onto directions of maximum variance, often resulting in clusters aligned with major data trends. ICA, on the other hand, focuses on statistical independence, potentially revealing unique, non-Gaussian patterns in the data that PCA might miss. RP provides randomized transformations that may emphasize diverse structures based on each random projection's characteristics. In the resulting clusters, some configurations were more insightful based on metrics such as silhouette score, BIC, and AIC, indicating which dimensionality reduction-clustering combination offered better-defined and distinct clusters.

VI. NEURAL NETWORKS ON DIMENSIONALITY-REDUCED DATASETS

In this exploration, the diabetes dataset is used to train neural network models on input spaces transformed by various dimensionality reduction algorithms: RP, PCA and ICA.

A. Hypothesis

The hypothesis for this experiment is that applying dimensionality reduction techniques to the diabetes dataset will enhance the performance of neural network models by reducing noise and focusing on the most informative features. Among these methods, PCA is expected to yield the best results due to its emphasis on maximizing variance and preserving global structure, which should help the neural network capture the dataset's key patterns. ICA may also improve performance, especially if independent components align with meaningful patterns in the data. Conversely, RP is likely to show more variability in performance, as it randomly projects the data, potentially retaining less structure than the other methods.

B. Experimental Methodology

In this experiment, the diabetes dataset is used to train neural network models on input spaces transformed through various dimensionality reduction algorithms. The dataset is split into training and testing sets, and only training data is used for unsupervised learning. Each dataset undergoes standardization before applying a chosen dimensionality reduction algorithm, with specific component values for each (3 components for RP, 5 for PCA, and 4 for ICA). This components led to the best results per Section IV of this report.

For each transformation, an MLP neural network is trained with a configuration tailored to the transformation method. Hyperparameters were determined based on grid search. The training process is iterative, running for 1,000 epochs, with metrics collected at each iteration. Log-loss is calculated for the baseline and dimensionality-reduced models, providing a measure of model performance across iterations, and accuracy scores are used to compare each model's performance on both training and validation sets.

C. Analysis and Results

The plot in Figure 11 displays the iterative learning curves of log-loss for neural network models trained on various dimensionality-reduced versions of the diabetes dataset using the baseline (untransformed) data, RP, PCA and ICA transformations. The baseline model (blue) achieves the lowest log-loss for both training and validation, indicating strong generalization. The RP-transformed model (green) shows relatively stable but higher log-loss values, indicating less effective convergence compared to the baseline. PCA (orange) and ICA (purple) models start with higher initial log-loss values but show gradual improvement, with ICA achieving a more stable trajectory over iterations. However, both PCA and ICA exhibit slightly higher validation loss compared to the baseline, suggesting a modest overfitting tendency. Overall, the baseline model provides the best performance with validation log-loss of 0.48 at convergence, followed by PCA (log-loss of 0.52) and ICA (log-loss of 0.56), with RP lagging in convergence and generalization.

The plot in Figure 12 shows the iterative learning curves for accuracy of neural network models trained on the diabetes dataset. The baseline model (blue) achieves the highest accuracy, stabilizing at around 0.75-0.8 for both training and validation, indicating strong performance and generalization. The PCA-transformed model (orange) also performs well, achieving slightly lower but comparable accuracy, around 0.75 for both training and validation, demonstrating effective representation of the data. ICA (purple) shows a steady improvement, reaching an accuracy plateau slightly below PCA, suggesting reasonable but lower data retention. In contrast, the RP-transformed model (green) struggles with both training and validation accuracy, stabilizing at a significantly lower level (0.65), indicating that the RP transformation might not preserve critical data patterns for the neural network. Overall, the baseline and PCA-transformed models perform best, with RP showing the least effectiveness in this setup.

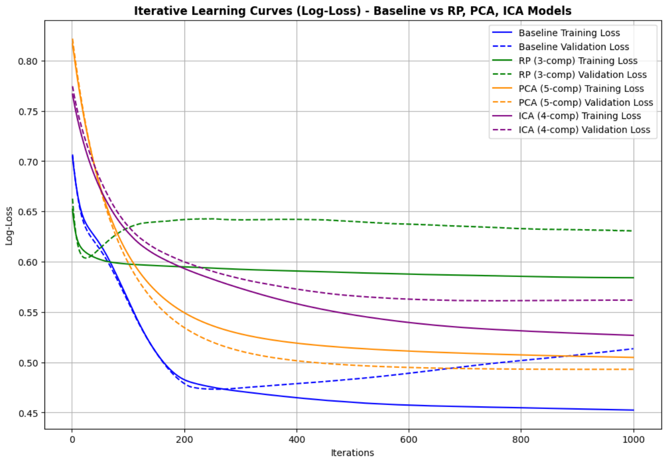

Fig. 11. Iterative Learning Curves (Log-Loss): Baseline vs. Dimensionality Reduction (RP, PCA and ICA) Models

At convergence, the training times for each model vary significantly. The baseline model is the slowest, taking 0.766 seconds, while RP is the fastest, completing in just 0.3643 seconds. PCA requires 0.500 seconds, and ICA takes slightly longer at 0.5477 seconds. This demonstrates that RP offers a substantial speed advantage, while the baseline model has the highest computational cost among the tested methods.

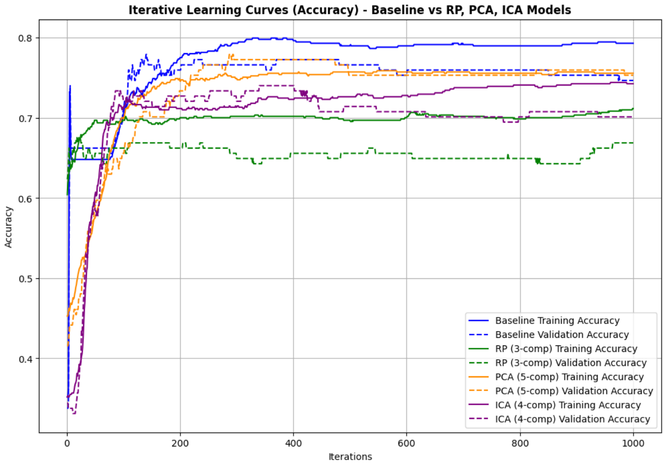

Fig. 12. Iterative Learning Curves (Accuracy): Baseline vs. Dimensionality Reduction (RP, PCA and ICA) Models

VII. Neural Networks on Clustered Datasets

In this exploration, the diabetes dataset is used to train neural network models on input spaces transformed by the clustering algorithms EM and K-Means.

A. Hypothesis

The hypothesis for this experiment is that training neural network models on the diabetes dataset with input spaces transformed by the clustering algorithms EM and K-Means will likely result in reduced performance compared to models trained on the original or dimensionally reduced data. This is because of the information loss that this clustering might lead. However, we expect EM to outperform K-Means due to its probabilistic approach, which provides richer, soft clustering representations that may better capture underlying patterns in the data. While overall results may be lower, EM's nuanced data representation is anticipated to yield comparatively better performance than K-Means.

B. *Experimental Methodology*

This experiment assesses the performance of neural network models trained on the diabetes dataset using various input transformations: baseline, EM, and K-Means clustering representations. First, the dataset is preprocessed, with the target labels encoded and data split into training and testing sets. Standard scaling is applied to the features to ensure consistent input distributions for all models. The baseline model is trained directly on the scaled features, while the EM and K-Means models use transformed input spaces from the clustering algorithms. For the EM model, the Gaussian Mixture Model with two components generates a probability distribution, and for K-Means, the three-cluster model produces transformed distances from the centroids. This components led to the best results per Section III of this report.

Each neural network is an MLP classifier with configuration tuned to the input transformation: logistic activation for the baseline model, and ReLU activation with layered configurations for EM and K-Means. Hyperparameters were determined based on grid search. The training process iteratively fits each neural network for 1,000 epochs, storing log-loss and accuracy scores for both training and validation sets. This iterative approach allows for detailed comparison of model convergence and overfitting trends.

C. *Analysis and Results*

Figure 13 compares the iterative learning curves in terms of log-loss for the baseline model, EM (2 components), and K-Means (3 components) models. The baseline model achieves the lowest log-loss values, demonstrating a smooth decline and stable convergence for both training and validation. In contrast, the EM model converges at a higher log-loss level than the baseline, with validation loss stabilizing at 0.65, indicating a less optimal fit. The K-Means model, though starting with a rapid initial decrease, exhibits high variability and could fail to converge as smoothly, with its log-loss fluctuating around 0.61 throughout the iterations. This instability in the K-Means model suggests that it struggles to achieve a stable clustering structure in this transformed space. Overall, the baseline model performs best, followed by the K-Means model if it converges, while EM has a slightly worse performance but a better defined convergence.

Figure 14 presents the iterative learning curves comparing accuracy. The baseline model demonstrates the highest accuracy, achieving stable training and validation accuracy around 0.75-0.8 after the initial training phase. In contrast, the EM model converges at a lower accuracy level, stabilizing slightly below 0.6. The K-Means model exhibits significant instability, with accuracy fluctuating widely across iterations and never reaching the levels of the baseline but above the EM models. This variability suggests that K-Means clustering struggles to create consistent clusters that enhance the neural network's performance in this context. Similar to the previous Figure, the baseline model performs best in terms of accuracy, followed by the K-Means model if converges properly, while EM performs the worst.

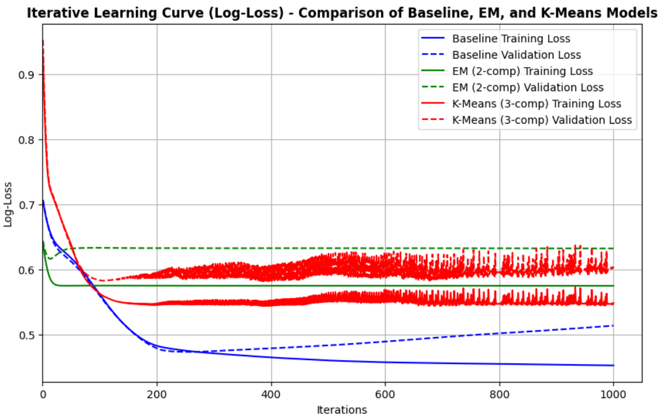

Fig. 13. Iterative Learning Curves (Log-Loss): Baseline vs. Clustering (EM and K-Means) Models

In terms of training time, the baseline model requires 0.7660 seconds, while EM is significantly faster at 0.1070 seconds, and K-Means falls in between with a runtime of 0.4925 seconds. EM is clearly the quickest method by a substantial margin, with K-Means moderately faster than the baseline model but still slower than EM.

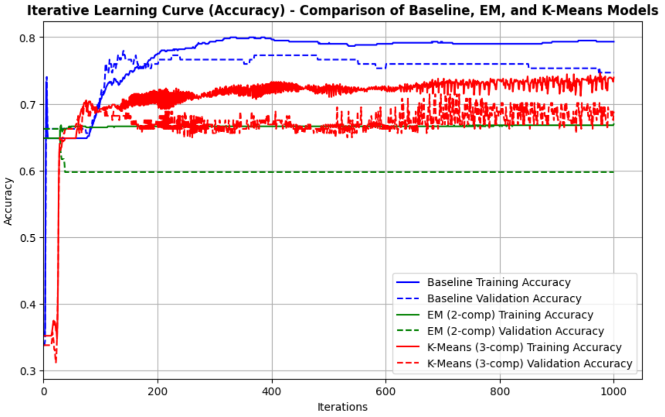

Fig. 14. Iterative Learning Curves (Accuracy): Baseline vs. Clustering (EM and K-Means) Models

VIII. Conclusion

In conclusion, this exploration assessed various unsupervised learning techniques—EM, K-Means, RP, PCA, and ICA—alongside their integration with neural networks on the diabetes and concrete datasets. Our initial hypotheses predicted that K-Means would be faster and more computationally efficient, which was confirmed by the results. K-Means consistently showed lower runtime than EM, particularly as component counts increased, affirming our expectation of K-Means' simplicity. However, EM outperformed in capturing complex, probabilistic cluster structures, especially on dimensionally reduced datasets.

In the dimensionality reduction segment, we hypothesized that PCA would best retain variance, making it ideal for clustering, while RP would offer speed and ICA would emphasize independent features. The results supported these expectations: PCA provided the best balance between variance retention and model performance, RP was the fastest, and ICA effectively extracted independent components but was computationally intensive. When clustering was applied to dimensionally reduced data, ICA-transformed datasets combined with EM achieved the best balance between model complexity and clustering performance, particularly for the diabetes dataset.

Finally, in the neural network experiments, we hypothesized that dimensionality reduction could enhance neural network performance by focusing on key features, with PCA expected to yield the best results. This was largely confirmed, as PCA-transformed inputs led to high accuracy and stable log-loss, although the baseline model generally retained the best overall performance without transformation. Clustering transformations (EM and K-Means) were anticipated to degrade performance, and results indicated that, while EM improved training speed, both clustering techniques fell short in accuracy and log-loss compared to the baseline.

Overall, our findings affirm that while dimensionality reduction can improve computational efficiency and clustering algorithms capture distinct data structures, using the original features generally yields the best neural network performance.

REFERENCES

[1] Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley & Sons.

[2] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

[3] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[4] Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. Pearson.