Analysis of Clustering and Dimensionality Reduction Techniques For Classification Problems

mshukla40@gatech.edu

Abstract

In this study, we explore the application of unsupervised learning and dimensionality reduction techniques on two datasets: Diabetes Prediction and MAGIC Gamma Telescope. We implement K-Means and Expectation Maximization clustering algorithms, along with Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Randomized Projections (RP) for dimensionality reduction. Our aim is to investigate how these algorithms interact with the data and affect clustering performance. By reapplying clustering on reduced datasets, we analyze improvements in data structure and classification accuracy. Additionally, we integrate clustering results as features to enhance neural network classifiers. The study demonstrates the impact of dimensionality reduction on clustering effectiveness and provides insights into optimizing machine learningA for complex datasets.

1 Introduction

In this study, we explore the impact of clustering and dimensionality reduction techniques on the Diabetes Prediction and MAGIC Gamma Telescope datasets. We begin by implementing K-Means and Expectation Maximization (EM) clustering algorithms on both datasets to identify inherent groupings. Next, we apply Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Random Projections (RP) to reduce the dimensionality of the datasets, aiming to uncover underlying structures and simplify the feature space. We then reapply K-Means and EM clustering on the reduced datasets to examine how the reduced feature space affects clustering performance. For the GAMMA dataset, we compare the performance of the reduced datasets against the baseline results of a neural network (NN) to evaluate the impact of dimensionality reduction. Additionally, we refit the NN (MLPClassifier) on the new input space derived from clustering outputs, comparing the results with the NN baseline. This comprehensive analysis demonstrates the benefits and challenges of using clustering and dimensionality reduction, providing insights into optimizing data preprocessing and enhancing model performance for complex datasets.

2 Datasets

2.1 Diabetes Prediction Dataset

The Diabetes Prediction dataset contains 21 attributes and is divided into 160,000 training samples and 68,000 test samples. It includes a mix of 14 binary, 4 categorical, and 3 continuous features.

Since k-means relies on distance measures, binary features might not be well-suited unless properly preprocessed.

Most of the features in this dataset have weak correlation. RP's ability to handle non-correlated features can be advantageous for GMM [1]. GMM assumes underlying Gaussian distributions for each cluster, and non-correlated features can contribute to forming these diversely shaped clusters [2]. RP might preserve the information needed for GMM to identify these clusters effectively.

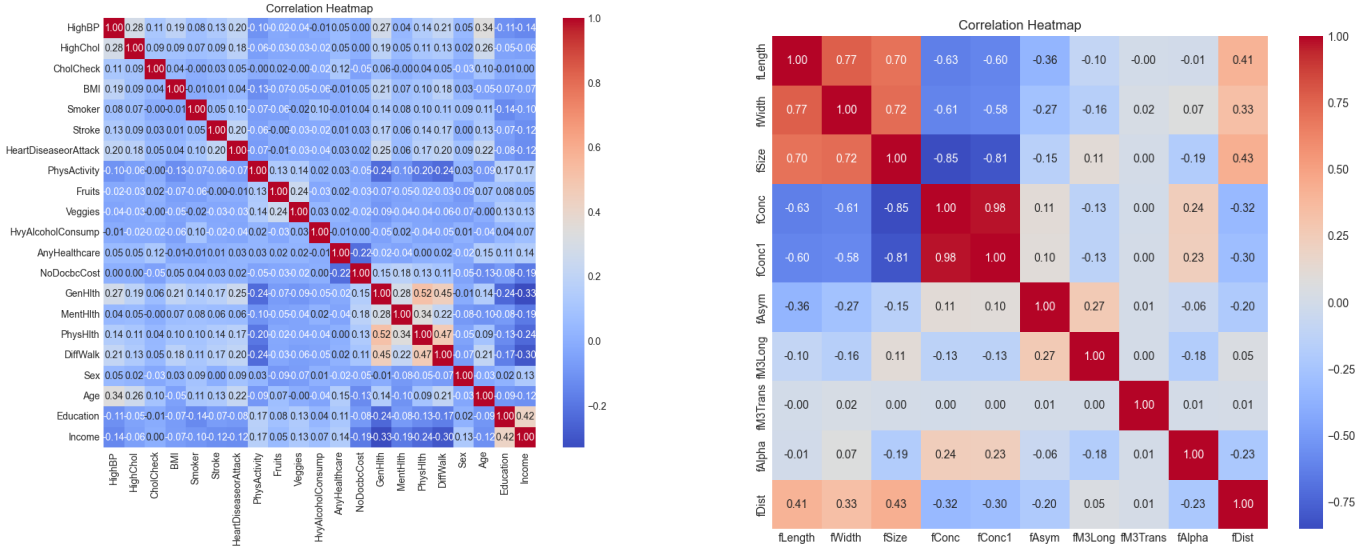
This dataset has Hopkins statistic 1 indicates a strong clustering tendency, likely due to the binary and ordinal nature of many features.

Adding clustering labels as features to the neural network input space is hypothesized to enhance the network's ability to distinguish between classes, potentially leading to improved performance compared to the baseline model

2.1.1 MAGIC Gamma Telescope Dataset

The dataset contains 10 features representing high-energy gamma particles and hadrons detected by the MAGIC telescope. The dataset is used for binary classification, distinguishing between gamma particles (signal) and hadrons.

The presence of strong positive correlations (values close to 1.0) between certain features suggests they might represent underlying groups or factors. For example, "fConc" and "fConc1" have a very high correlation (0.976),



(a) Feature Correlation of Dataset 1

(b) Feature Correlation of Dataset 2

Figure 1: Feature Correlation of Datasets

indicating they likely capture similar information. Dimensionality reduction techniques like PCA can help remove this redundancy, potentially improving the efficiency of clustering algorithms.

The presence of both positive and negative correlations can impact the performance of clustering algorithms like K-means, which assumes spherical clusters. GMM, which can handle non-spherical clusters, might be a better choice in this case.

3 K Means and EM Clustering

3.1 Dataset 1

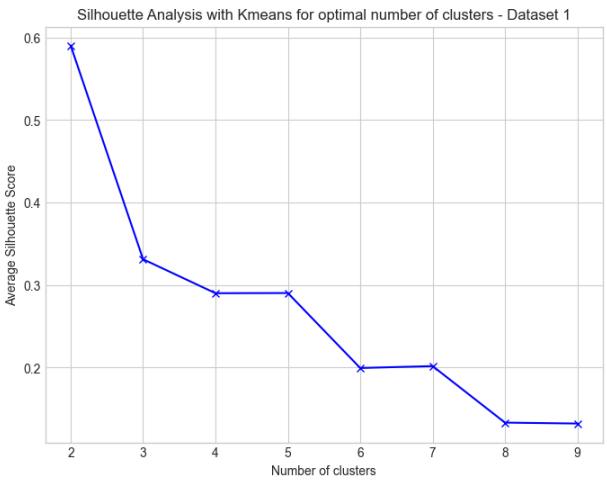
K-Means clustering was applied to the Diabetes Prediction dataset to explore its inherent clustering structure. The number of clusters (k) was varied from 2 to 9, and the performance of each clustering configuration was evaluated using multiple metrics, including the silhouette score, sum of squared distances, variance ratio criterion, and Davies-Bouldin Index. The highest silhouette score (0.59) was achieved with 2 clusters, indicating that this configuration provides the most cohesive clustering solution. As the number of clusters increases, the sum of squared distances generally decreases, which is expected since more clusters should reduce the distance within each cluster. The highest VRC was observed with 2 clusters, aligning with the silhouette score findings. A lower Davies-Bouldin Index indicates better clustering performance. The lowest value (0.785) was achieved with 2 clusters, further supporting the selection of 2 clusters as the optimal number for this dataset.

Expectation Maximization (EM) clustering, also known as Gaussian Mixture Model (GMM) clustering, was applied to the Diabetes Prediction dataset. The number of clusters (k) was varied from 2 to 10, and the performance of each clustering configuration was evaluated using multiple metrics, including silhouette score, variance ratio criterion (VRC), Davies-Bouldin Index, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and log-likelihood score. Different covariance types (full, tied, diagonal, spherical) were also explored to identify the best-fitting model for the data.

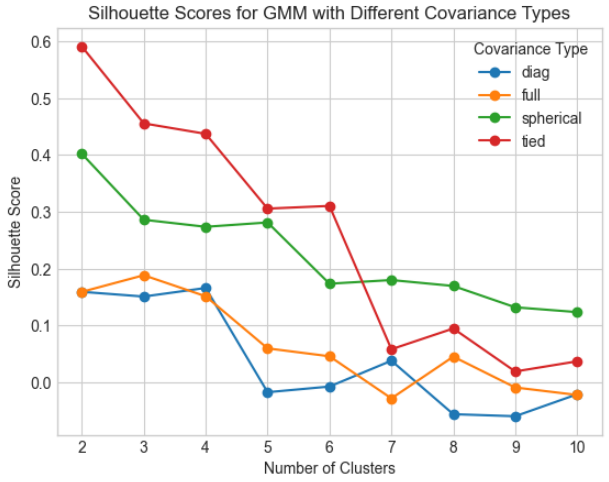
The tuning results indicates that the optimal clustering configuration is achieved with 2 clusters and a tied covariance type. This configuration provides the highest silhouette score, highest VRC, lowest Davies-Bouldin Index, and favorable AIC and BIC values. The results suggest that the data naturally forms two well-separated clusters.

3.2 Dataset 2

On Dataset 2, the highest silhouette score (0.397) was achieved with 2 clusters. This indicates that the data points within each cluster are moderately matched and distinct from points in other clusters. The sum of squared distances decreases as the number of clusters increases, indicating tighter clusters. However, the rate of decrease diminishes, suggesting diminishing returns with more clusters. [3] The VRC evaluates the ratio of the variance between clusters to

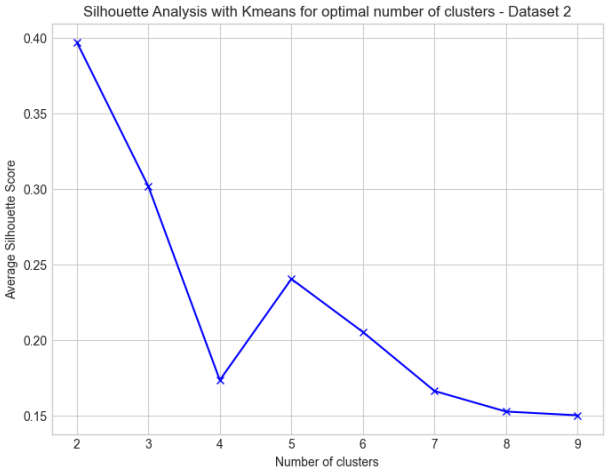


(a) Silhouette Score of Dataset 1 with K Means

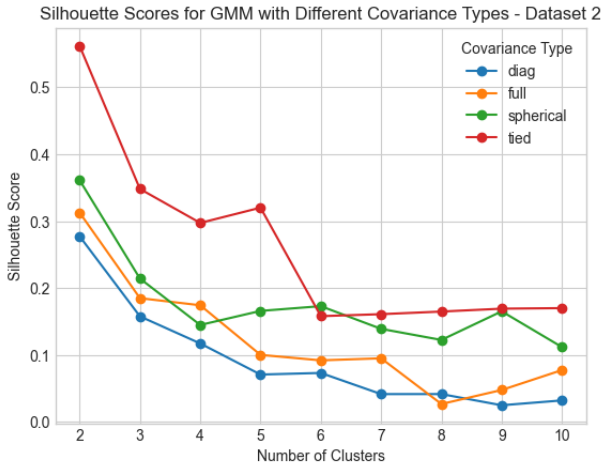


(b) Silhouette Score of Dataset 1 with Gaussian Mixture

Figure 2: Silhouette Score against number of clusters



(a) Silhouette Score of Dataset 2 with K Means



(b) Silhouette Score of Dataset 2 with Gaussian Mixture

Figure 3: Silhouette Score against number of clusters

the variance within clusters. A higher VRC indicates better clustering performance. The highest VRC was observed with 2 clusters (4591.169), which aligns with the silhouette score findings.

However, 3 clusters provide the lowest Davies-Bouldin Index (1.425), indicating that this configuration may also be considered for achieving better cluster separation

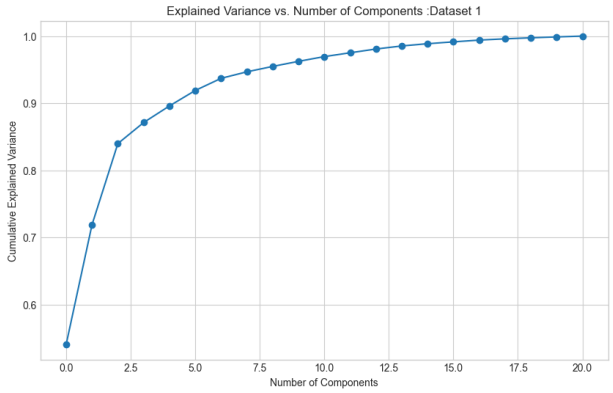
The analysis of Gaussian Mixture hyperparameters for the MAGIC Gamma Telescope dataset indicates that the optimal clustering configuration is achieved with 2 clusters and a tied covariance type. This configuration provides the highest silhouette score, the highest VRC, the lowest Davies-Bouldin Index, and favorable AIC and BIC values.

The tied covariance type consistently performs better across multiple metrics, indicating that assuming equal covariance across clusters helps in better modeling the data. The full covariance type, despite being more flexible, generally results in worse performance metrics due to over-fitting or capturing too much noise.

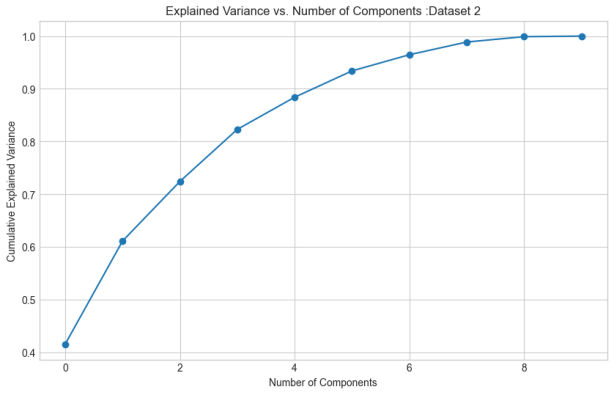
4 Dimensionality Reduction on both Datasets

4.1 Principal Component Analysis

For dataset 1, PCA effectively reduces the dimensionality from 21 to 5 while retaining 90% of the dataset's variance. The first few components capture the majority of the variance, indicating that the dataset has a significant amount of redundant or less informative features [4].

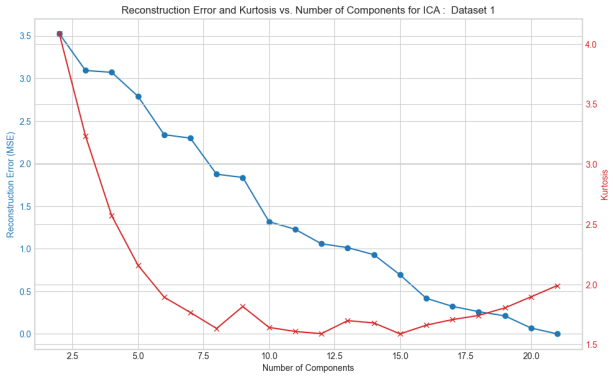


(a) PCA Explained Variance - Dataset 1

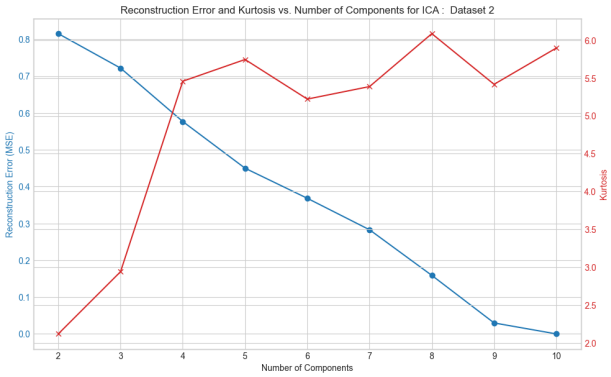


(b) PCA Explained Variance - Dataset 2

Figure 4: Explained Variance for varying Principal Components



(a) RP Explained Variance - Dataset 1



(b) RP: Kurtosis and Reconstruction Error - Dataset 2

Figure 5: Kurtosis and Reconstruction Error for varying Randomly Projected Components

For dataset 2, PCA reduces the dimensionality from the initial set of features to 5 components while retaining 90% of the dataset's variance.

4.2 Randomized Projection

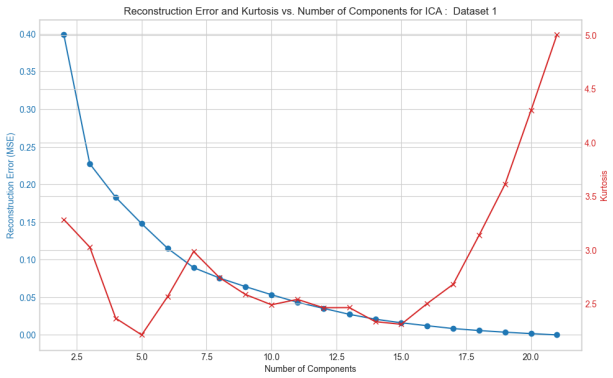
For dataset 1, RP shows that retaining all 21 components minimizes reconstruction error, highlighting that the original feature set is essential for accurate data representation. However, based on kurtosis, RP suggests that only 2 components capture the most significant non-Gaussian structures, indicating that significant dimensionality reduction is possible while preserving essential data characteristics.

RP on dataset 2 suggests that 10 components are needed to minimize reconstruction error, highlighting the need for a higher-dimensional space to accurately capture the dataset's structure. [1] Kurtosis-based analysis indicates that significant non-Gaussian structures can be captured with as few as 4 components, though 8 components would have provided a better balance.

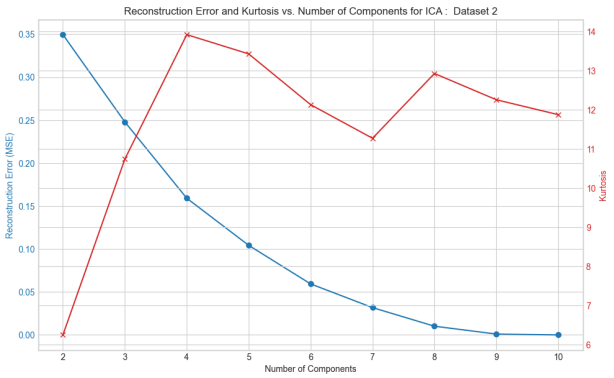
4.3 Independent Component Analysis

ICA on dataset 1 shows that retaining all 21 components minimizes reconstruction error and maximizes kurtosis to capture the independent structure in the data, but we'll choose 2 to reduce complexity, though this will affect the model performance. Unlike PCA, which aims to maximize variance, ICA focuses on statistical independence, suggesting that the dataset contains a complex structure that requires the full feature set to preserve.

ICA on dataset 2 indicates that 10 components are needed to minimize reconstruction error, similar to RP. The optimal number of components based on kurtosis is 4, suggesting that fewer components capture the most significant non-Gaussian structures in the dataset.



(a) ICA Explained Variance - Dataset 1



(b) ICA: Kurtosis and Reconstruction Error - Dataset 2

Figure 6: Kurtosis and Reconstruction Error for varying Independent Components

4.4 Clustering on Reduced Datasets

Clustering algorithms, K-Means and Expectation Maximization (EM), were applied to the Diabetes Prediction dataset after dimensionality reduction using PCA, ICA, and RP. The clustering performance was evaluated using several metrics: homogeneity score, completeness score, V-measure score, adjusted Rand score, adjusted mutual information score, silhouette score, accuracy score, and sum of squared distances.

For Dataset 1:

- GMM achieves significantly higher Accuracy scores in almost all cases compared to K-means. This suggests that the data might contain some non-spherical clusters, which GMM can handle better.
- **Interestingly,** the superior performance of RP over PCA and ICA in this specific diabetes dataset suggests that preserving pairwise distances is more crucial than capturing principal components or independent components to identify the separation between diabetic and non-diabetic patients. Our data has complex, nonlinear relationships with known outliers which RP can much better.
- The counter-intuitive result of ICA performing poorly on this dataset with low correlated features suggests that though features show a degree of independence, it doesn't guarantee true statistical independence. ICA seems to struggle since the underlying components exhibit non-linear relationships. [5]
- While the data might have some non-spherical clusters that favor GMM, the separation between diabetic and non-diabetic patients might also be well-represented by distances between data points, which K-means with RP could leverage effectively in this dataset.

For Dataset 2:

- In almost all cases, GMM achieves significantly higher Silhouette Score, Homogeneity, Completeness, V-Measure, Adjusted Rand Score, and Accuracy scores compared to K-means. GMM assigns data points to clusters probabilistically and thus captures multimodal distributions and outliers more effectively than K-means. Though it is slower than K-means due to the iterative Expectation-Maximization algorithm.
- ICA with K-means shows the best performance among K-means methods, with the highest silhouette, homogeneity, completeness, V-measure, adjusted rand, and accuracy scores. ICA with GMM shows high homogeneity, completeness, and V-measure scores, suggesting effective clustering. However, the accuracy score is significantly lower, indicating that ICA might be identifing features that are independent but not necessarily the most discriminative for the binary classification task of distinguishing between gamma particles and hadrons. This is expected since we chose a lower component count for reduction even though the kurtosis and reconstruction analysis showed we'd need all 10 components for ICA for a good feature transformation.
- While RP with K-means performs similarly to PCA with K-means, with slightly lower scores in some metrics. This suggests that RP might not have preserved the information needed for good separation unlike PCA which identifies the directions of maximum variance, which are often the most informative for clustering . Whereas, GMM with RP shows a moderate improvement over both GMM with PCA and ICA as it can model more complex cluster shapes (e.g., elliptical clusters) and uses a probabilistic approach. GMM can potentially benefit from the spreading effect of RP, which might reveal underlying structures not captured by PCA or ICA.

Clustering Algorithm	K-means			Gaussian Mixture		
Dim. Red.	PCA	ICA	RP	PCA	ICA	RP
Time (sec)	0.13	0.05	0.05	2.7	6.4	2.6
Silhouette Score	0.64	0.50	0.74	0.64	0.63	0.74
Homogenity Score	0.003	0.038	0.005	0.003	0.003	0.004
Completeness Score	0.004	0.028	0.006	0.003	0.003	0.005
V-Measure Score	0.003	0.032	0.006	0.003	0.003	0.005
Adjusted Rand Score	0.038	0.103	0.049	0.034	0.034	0.043
Accuracy Score	0.782	0.27	0.786	0.782	0.21	0.794

Table 1: Performance of Clustering on Dataset 1 After Dimensionality Reduction

Clustering Algorithm	K-means			Gaussian Mixture		
Dim. Red.	PCA	ICA	RP	PCA	ICA	RP
Time (sec)	0.05	0.05	0.05	0.81	0.40	0.83
Silhouette Score	0.39	0.44	0.38	0.62	0.62	0.56
Homogenity Score	0.009	0.023	0.008	0.087	0.089	0.083
Completeness Score	0.009	0.04	0.009	0.26	0.26	0.23
V-Measure Score	0.009	0.035	0.008	0.13	0.13	0.122
Adjusted Rand Score	0.03	0.08	0.034	0.106	0.11	0.11
Accuracy Score	0.398	0.673	0.381	0.708	0.291	0.7086

Table 2: Performance of Clustering on Dataset 2 After Dimensionality Reduction

5 Analysis of Neural Network Performance on MAGIC GAMMA Dataset with Reduced Features vs Additional Cluster Labels

We run baseline Neural Network (NN) on the MAGIC GAMMA dataset compared to using PCA, ICA, and GRP for dimensionality reduction before feeding the data into the NN. Retuning the NN parameters for each scenario helps in optimizing performance. However, since the number of components for DR is data-dependent and not algorithm-dependent, it is kept constant.

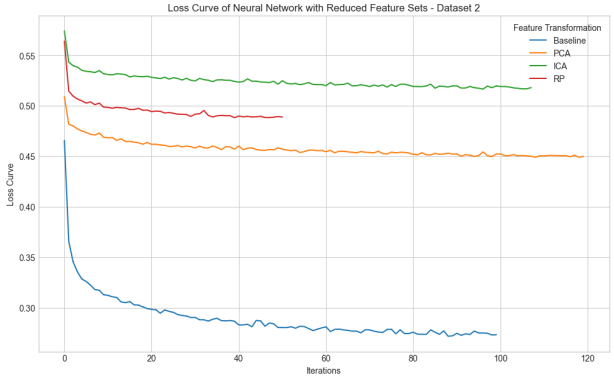
- Since PCA outputs both positive and negative values, the `tanh` activation function, which outputs values between -1 and 1, helps in normalizing and centering the data around zero, leading to faster and more stable convergence
- The `relu` activation function is particularly suited for the sparse and non-Gaussian nature of the ICA components. It helps in effectively learning from the independent features by mitigating the vanishing gradient problem.
- The baseline NN has higher accuracy and F1-Score compared to all DR techniques (PCA, ICA, RP). This suggests that the original feature space is more informative for the NN than the reduced feature spaces.
- PCA performs better than ICA and RP but worse than the baseline. This suggests that while PCA captures the most variance, some important information for classification is still lost during dimensionality reduction.
- ICA shows the lowest performance among the DR techniques, with the lowest accuracy (0.74) and F1-Score (0.70). This suggests that the number of independent components extracted by ICA is not fully able to bifurcate the data into two clear clusters.
- GRP achieves a moderate improvement over both PCA and ICA. While GRP might not preserve the exact correlations between features, it might have retained more of the information needed for the NN to classify gamma rays.

Table 3: Neural Network Performance Metrics for different Feature Engineering Techniques

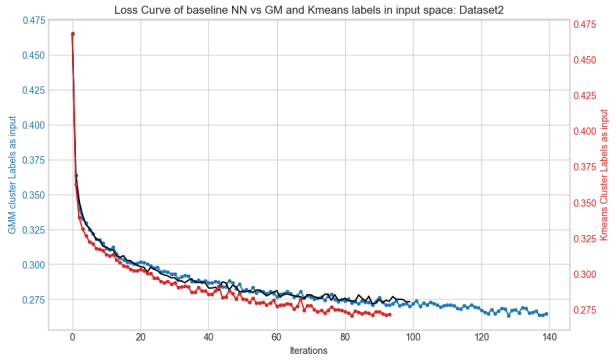
Metric	Baseline	PCA	ICA	RP	Kmeans Clusters	GM Clusters
Accuracy	0.86	0.78	0.74	0.77	0.87	0.88
F1-Score	0.86	0.77	0.70	0.75	0.87	0.88
MSE	0.13	0.21	0.25	0.22	0.12	0.12

Algo	Activation	Hidden Layer Size	Learning Rate
PCA	tanh	(10, 10)	0.001
RP	relu	(50,)	0.001
ICA	relu	(100,)	0.01

Table 4: Best Hyperparameters selected post-tuning



(a) Loss Curve for each Dimensionality reduction technique



(b) Loss Curve for each Cluster Label added to input space

Figure 7: Comparing Baseline NN Performance against Feature Transformations vs Cluster Label Additions

- Explore feature selection techniques that can identify a smaller subset of the original features that are most informative for the classification task

It shows a significant improvement in the performance of the Neural Network (NN) when cluster labels from K-means and Gaussian Mixture Model (GMM) are added to the original input space compared to using dimensionality reduction techniques (PCA, ICA, GRP). Here's a breakdown of the key observations:

- Both K-means and GMM cluster labels lead to an increase in accuracy, F1-score, and a decrease in MSE compared to the baseline NN. This suggests that the cluster labels provide valuable information for the NN to classify gamma rays.
- GMM achieves slightly better performance than K-means in all metrics. This could be because GMM can capture varying cluster shapes unlike K-means. Since our data has imbalance, GMM can classify these points better.
- Adding cluster labels can be seen as a form of feature engineering. These labels provide a new feature that directly reflects the cluster a data point belongs to, which can be easier for the NN to learn from compared to the original, potentially more complex features.

6 Results

1. Diabetes Dataset - The performance of K-means with RP on the diabetes dataset is an interesting finding. While GMM might be generally better suited for non-spherical clusters, RP seems to have preserved the information K-means needs to effectively separate the data points in this specific case.
2. MAGIC Gamma Dataset - The presence of strong correlations between features (as seen in the correlation matrix) might have affected the effectiveness of PCA. PCA prioritizes capturing variance, and these correlations might have led it to focus on redundant information not necessarily relevant for clustering. The data likely contains non-spherical clusters, as evident by the superior performance of GMM compared to K-means. Dimensionality reduction techniques might not have always captured the structure needed for optimal K-means performance.
3. While dimensionality reduction simplifies the feature space for our dataset 2, it can lead to a significant performance drop for neural networks. The baseline model using the original features remains the best performing, highlighting the importance of comprehensive feature sets for accurate classification in the MAGIC Gamma Telescope dataset. We could potentially improve ICA performance by increasing the number of components we

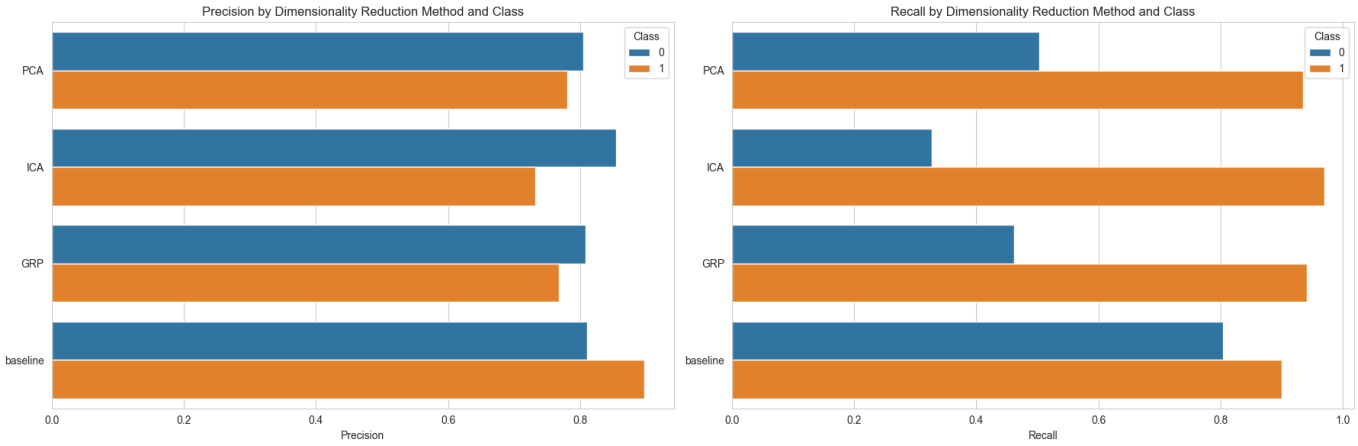


Figure 8: Classification Report on Reduced Dataset

reduce and minimize reconstruction error. We should also perform multiple runs with different random seeds for RP to evaluate the stability and robustness of the results.

4. PCA may reduce the dimensionality effectively but does not always preserve the class-specific information, leading to poorer recall for class 0. ICA focuses on statistical independence, which may not align well with the class boundaries in the data leading to very low recall (0.327), indicating that many gamma particles are not being identified correctly.

5. The inclusion of cluster labels from both K-means and GMM significantly enhances the NN's performance, surpassing the baseline. This suggests that clustering captures meaningful patterns and structures in the data that the NN can leverage for better predictions. [6]

Conclusion

With minimal feature engineering, K means performs much better than GM with higher silhouette score and lower wall clock time. But with reduced features, GM overperforms as its probabilistic nature is able to capture the nuances in the data labels more accurately. Overall, these results suggest that incorporating cluster labels obtained from unsupervised learning techniques can be a valuable approach to improve the performance of Neural Networks for classification tasks. If the relationships between features and classes are non-linear, linear DR methods (PCA) might not be effective. ICA, while non-linear, might not align well with class boundaries. Using SMOTE should help balance the classes and improve class 0 classification. Considering the relatively low performance on standard metrics, exploring other clustering algorithms like DBSCAN or hierarchical clustering might also be beneficial.

References

[1] Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 245-250).

[2] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[3] Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. International Journal of Advanced Research in Computer Science and Management Studies, 1(6), 90-95.

[4] Jolliffe, I. T. (2002). Principal Component Analysis. Springer Series in Statistics.

[5] Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. Neural Networks, 13(4-5), 411-430.

[6] Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), 478-487.