Evaluating Clustering Performance on Political and Social Data Through Dimensionality Reduction Techniques

Tyler Dickson
Georgia Institute of Technology
tdickson30@gatech.edu

***Abstract*—This paper examines the application of unsupervised clustering and Dimentionality Reduction (DR) techniques on political contribution and school shooting donation datasets, addressing the gap in understanding how DR impacts clustering performance in complex, and often categorical socio-political feature space. We applied Expectation Maximization (EM) and K-means clustering with three DR techniques—PCA, RP, and ICA. Through implementation over two datasets, PCA most effectively reduced feature space while still enhancing K-means clustering.**

We then implement DR techniques to a NN to maximize test accuracy on the school shooting dataset. substantial feature reductions occurred in all DR implementations, with PCA yielding better cluster separability than RP and ICA. Notably, ICA reduced dimensionality but led to poorly grouped clusters, highlighting a tradeoff between optimizing test accuracy and clustering quality. Our findings affirm PCA's effectiveness in preserving structure for clustering across datasets of varying complexity.

I. Introduction

Unsupervised clustering is essential for identifying patterns and categorizing data without labels. Iterative algorithms like EM and K-means uncover these patterns by estimating the probability of each data point belonging to a cluster. In this paper, we apply EM and K-means to two datasets from prior work in *Assignment 1* and perform multiple iterations of clustering and DR.

We first apply EM and K-means to **Dataset 1** as a baseline. Then, we use three DR techniques—RP, PCA, and ICA—to reduce the dataset's dimensionality and re-run clustering on the transformed data. We repeat this process for **Dataset 2**.

Additionally, we incorporate the three DR techniques into a previously implemented NN to observe performance changes. Finally, we add new features derived from EM and K-means clustering projections and re-evaluate them within the NN.

II. Problem Space

Dataset 1 examines political contributions, predicting whether donations align with party affiliation based on **9** distinct features: location (city, state, zip code), occupation, date, transaction amount, and party (Democrat, Republican, Other). Clustering this data aims to reveal socio-political trends within these features.

Dataset 2 explores the relationship between school shootings and subsequent donation patterns to pro-gun and anti-gun committees, predicting whether a 'donation spike' to pro or anti-gun committees transpired. I define a 'Donation Spike' as an instance where both the donation frequency and amount were **50%** higher **3** days after the shooting, when compared to the moving **3** day average from **2003 - 2024** (excluding the top **10%** of top donations as outliers). Key features include quantities killed, injured, casualties, donation amounts, spikes in donations, and various shooting characteristics. After one hot encoding, the resulting number of features is **68**. This analysis seeks to uncover how specific aspects of shootings might influence donation behaviors across political affiliations.

III. Dataset 1:Political Contributions Analysis

Applying EM and K-means algorithms should reveal distinct clusters in political contributions. We expect EM to outperform K-means because of its ability to model complex, non-spherical cluster shapes, which aligns with the socio-economic and categorical nature of the data. Among DR techniques, ICA is anticipated to perform best, as it builds on PCA reduced data, enhancing component independence. Overall, post-DR EM should yield the highest performance improvement, as it is sensitive to clearly separated clusters, which DR techniques are expected to produce.

A. *Categorical Data Considerations & Implementation*

for clustering in **Dataset 1**, I used target encoding to transform categorical variables—city, state, zip code, and occupation—by grouping them based on the average transaction amount (donation) for each category. This approach prioritizes the transaction amount, identified as the most influential predictor in previous work with XGBoost on this dataset. Target encoding with mean transaction amounts preserves the key information within each category while maintaining a continuous data format that suits clustering. **Dataset 1** is condensed into groupings by date and transaction amounts due to the high observation of instances (over **32 million**), which caused fluctuations in the data. To address the variance between large and small transaction amounts, I applied a log transformation to the transaction amount feature. This transformation reduced

skewness and stabilized variance, allowing clustering algorithms to form meaningful clusters without distortion from extreme values. Descriptive statistics are generated for each cluster to support analysis, but due to space limitations, they are not graphically represented.

For **Dataset 2**, I applied one-hot encoding to all boolean and non-binary variables to enable a performance comparison across datasets. **Dataset 2** contains **354** observations in relation to number of school shootings cataloged by the Wall Street Journal's comprehensive list of School Shooting Data from **1999** to present [1].

All features are standardized using scikit-learn's StandardScaler [3] to ensure equal contribution to distance calculations.

For EM algorithms, I selected the optimal number of clusters by calculating Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) scores across Gaussian Mixture Models (GMM) with 1–10 clusters. For K-means algorithms, I used Silhouette Scores (SS), cross-referenced with the Elbow method, to determine the ideal cluster count.

B. *Hypothesis: Dataset 1*

Applying EM and K-means algorithms should reveal distinct clusters in political contributions. We expect EM to outperform K-means because of its ability to model complex, non-spherical cluster shapes, which aligns with the socio-economic and categorical nature of the data. Among DR techniques, ICA is anticipated to perform best, as it builds on PCA reduced data, enhancing component independence. Overall, post-DR EM should yield the highest performance improvement, as it is sensitive to clearly separated clusters, which DR techniques are expected to produce.

C. *Expectation Maximization: Dataset 1*

The EM algorithm produced clusters visualized in Figure 1. Both AIC and BIC scores steadily decreased with more clusters, stabilizing at **10**, indicating an optimal balance between complexity and fit.

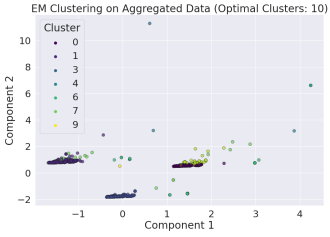

Fig. 1. EM Clustering: DS 1

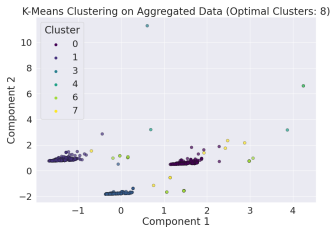

Fig. 2. K-means Clustering: DS 1

Clusters **0**, **1**, and **2** contain the highest observations. Specifically, **Cluster 0** represents a stronghold for Democrat donations with **386** contributions averaging **$2.3 million** each. **Cluster 1** includes **480** Republican-affiliated donations, with a mean donation size of **$4.7 million**.

These results align with the hypothesis that EM would effectively capture underlying patterns in political donations based on socio-economic data, showcasing the algorithm's ability to identify natural groupings in complex datasets.

D. *K-means Clustering: Dataset 1*

Applying the K-means algorithm, the optimal number of clusters was determined to be **8** based on the elbow method and silhouette score, with clusters visualized in Figure 2.

K-means identified key clusters, including a Democrat-leaning cluster with **463** contributions averaging around **$4.5 million** and a Republican-leaning cluster with **494** donations, averaging **$4.7 million**. Additionally, K-means detected a high-value cluster with an average transaction of **$238 million**, similar to the high-value cluster seen in EM.

K-means produced a slightly different cluster structure, identifying **8** clusters instead of EM's **10**, supporting the hypothesis that EM's probabilistic approach is sensitive to finer variations, while K-means yields more balanced groupings.

E. *Randomized Projections: Dataset 1*

RP is a DR technique that projects high-dimensional data into a lower-dimensional space using a randomly generated projection matrix. During projection, the reconstruction error (y-axis in Figure 3) is calculated using the pseudo-inverse of the projection matrix. This error represents the difference between the original and reconstructed data, indicating the amount of information lost in the DR process. [**?**]

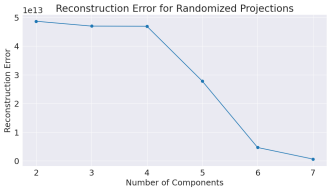

Fig. 3. RP: Reconstruction Error

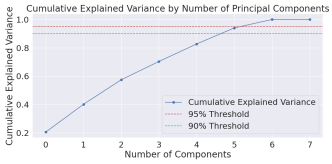

Fig. 4. PCA: Explained Variance

Consistent with the hypothesis, reconstruction error decreases gradually with more components, with a sharp drop around component **4**. By component **7**, nearly all information is retained, suggesting that a **7-dimensional space** effectively captures the original structure. Both RP and PCA identified **7** as an optimal number of components, though PCA more clearly visualizes variance maximization. By component **6**, **94.09%** of variance is captured. We proceed with **7** components for a comprehensive model.

F. *Principle Component Analysis: Dataset 1*

PCA reduces high-dimensional data by projecting it onto principal components that maximize variance. Each principal component corresponds to an eigenvector of the covariance matrix, with its eigenvalue indicating the variance it captures. Higher eigenvalues represent components that better preserve the data's structure. [5]

Our implementation used explained variance to decide how many components to retain. As shown in Figure 4, cumulative explained variance reaches nearly **100%** by Component **7**, suggesting that **7** components approximate the original data with minimal loss. [**?**] Notably, PCA accounts for **94.09%** of variance by component **6**. This indicates that **6** components may produce efficient clustering results as well.

Table I highlights that Component **1** explains **20.56%** of the variance, and Component **2** accounts for **19.58%**. Together, these two components capture a substantial portion of the data's variance, providing a concise representation.

TABLE I
EXPLAINED VARIANCE RATIO BY PRINCIPLE COMPONENT

Component	1	2	3	4	5	6	7	8
Variance (%)	20.56	19.58	17.32	12.92	12.35	11.36	5.91	0.0
Eigenvalue	1.646	1.567	1.386	1.035	0.989	0.909	0.473	0.0

G. *Independent Component Analysis: Dataset 1*

ICA transforms data by maximizing statistical independence among components, extracting unique, independent features. ICA identifies components with non-Gaussian distributions, capturing distinct structures within the data. We used kurtosis, a measure of non-Gaussianity, to assess component independence, aiming to maximize average absolute kurtosis. Higher kurtosis indicates components with more independent, non-Gaussian features. Figure 5 shows that average absolute kurtosis steadily increases with the number of components, peaking at **7** components.

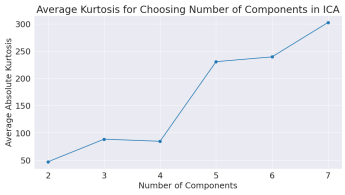

Fig. 5. ICA: Kurtosis

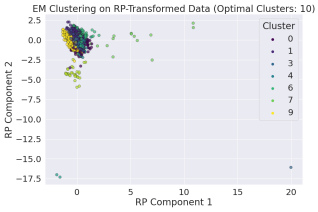

Fig. 6. EM Clustering: RP

H. *EM Clustering on RP Transformed Dataset 1*

RP was initially applied to Dataset 1, resulting in an optimal component number of **7** (III-D) as this minimized reconstruction error and retained essential structure while reducing computational complexity.

Using the EM algorithm on the RP-transformed data, AIC and BIC analyses identified **10** clusters (Figure 6).

Cluster **7** primarily contains Democrat-affiliated donations with a mean transaction amount of approximately **$3.1 million**. Cluster **1** is largely composed of Republican-affiliated donations, with a mean transaction of around **$1.3 million**. Cluster **6** stands out with a significantly higher transaction amount average of **$238 million**, suggesting a concentration of high-value contributions.

The RP transformation allowed EM to capture meaningful clusters despite the DR. This aligns with the hypothesis that lower dimensional representations can still capture complex patterns in donation behaviors.

I. *EM Clustering on PCA Transformed Dataset 1*

PCA reduced Dataset 1 to **6** components (sectionIII-L), capturing **100%** of the cumulative explained variance (Table I). Components **1 (20.56%)** and **2 (19.58%)**—the two with the highest variance—were selected to visualize clustering results.

We determined the optimal cluster count using BIC and AIC scores, with both scores favoring **10** clusters.

Cluster **0** shows a strong Democrat affiliation with **424** records and a moderate average transaction amount. Cluster **3** represents the largest Republican-affiliated donations, totaling **452** records with a relatively lower transaction average. Cluster **5** includes mixed affiliations with some of the highest transaction amounts.

The results support the hypothesis, showing distinct patterns similar to the RP-transformed data. However, PCA enhanced cluster differentiation, emphasizing specific affiliations and transaction levels more clearly.

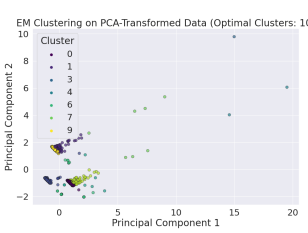

Fig. 7. EM Clustering: PCA

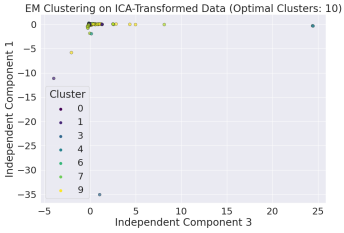

Fig. 8. EM Clustering: ICA

J. *EM Clustering on ICA Transformed Dataset 1*

After standardizing and reducing Dataset 1 with PCA to retain **95%** of the variance, ICA was applied, optimizing for kurtosis. The ideal dimensionality was determined to be **7** components, where maximum average kurtosis was achieved (figure 5)

Using the EM algorithm, BIC and AIC scores both indicated **10** optimal clusters, visualized with the two ICA components showing the highest kurtosis (Figure 8)

Notably, cluster **0** represents predominantly Democrat-affiliated donations with a total of 368 records. Whereas cluster **2** primarily contains Republican-affiliated donations with **385** records. Cluster **7** shows significant variance in donation size, with a very high transaction amount of around **$344 million**, potentially capturing unique outliers.

Contrary to expectations, the clusters showed limited separation in the ICA-transformed space, likely due to categorical data transformed into continuous values, which may yield independent features without strong nonlinear separability.

K. *K-means Clustering on RP Transformed Dataset 1*

K-means clustering was applied to the RP transformed dataset with **7** components, selected based on minimized reconstruction error (Section III-D, Figure 3). The optimal cluster count was determined as **2** by silhouette scores, differing from the **10** clusters identified in the EM-RP results.

Cluster **0** includes most data points, covering a range of affiliations (Republican, Democrat, and others) with an average transaction amount around **$2.76 million**. Cluster **1** contains only **5** records, representing high-value donations averaging **$247 million**, likely outliers.

The lower number of clusters in K-means suggests that K-means perceives the data structure as more similar in this RP space. The algorithm groups data into broader and more generalizable categories. This differs from the EM algorithm, which produced finer distinction with **10** clusters. The contrast could indicate that K-means prioritizes variance minimization within clusters, while EM prioritizes subtle differences.

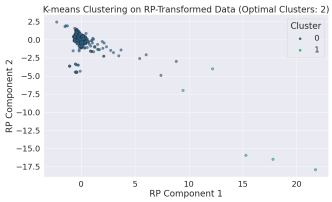

Fig. 9. K-means Clustering: RP

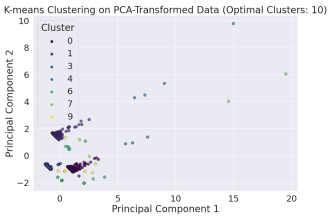

Fig. 10. K-means Clustering: PCA

L. *K-means Clustering on PCA Transformed Dataset 1*

K-means clustering was applied to the PCA-transformed dataset using **7** components, selected based on maximizing explained variance (Section III-E). The silhouette score indicated an optimal cluster count of **9**, with clusters plotted on the first two principal components (Figure 10).

Cluster **0** represents a large portion of Democrat-affiliated transactions and mean transaction amount around **3** million. Cluster **1** is dominated by Republican-affiliated donations, with a similar average transaction amount of **2.4** million. Cluster **3** captures high-value contributions from mixed affiliations. Finally, other clusters contain various affiliations with more moderate transaction values, indicating diverse groups with smaller contributions.

The affiliation-dominated clusters confirm that K-means on PCA-transformed data successfully identified distinct groupings based on political affiliation and transaction size. This aligns with the hypothesis, as PCA's dimension reduction retained key variance, allowing K-means to capture clusters with clear characteristics.

M. *K-means Clustering on ICA Transformed Dataset 1*

To apply K-means on the ICA-transformed dataset, we first reduced it to **7** components using PCA (Section 4). The optimal cluster count was determined by silhouette scores, resulting in **10** clusters. The two ICA components with the highest kurtosis were used for visualization (Figure 11), as these maximize non-Gaussianity and offer clearer feature separation.

Results of this implementation were inconsistent with our hypothesis. The reduced separation observed between clusters compared to other DR techniques may be due to plotting the two ICA components with the highest kurtosis. Unlike PCA, which maximizes variance and often enhances cluster separation, ICA maximizes statistical independence. As a result, high-kurtosis components may not reflect the primary directions of spread in the data, leading to tighter, less visually separated clusters.

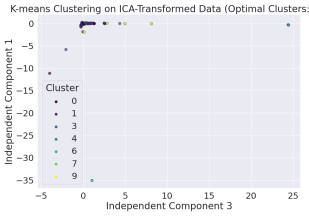

Fig. 11. K-means Clustering: ICA

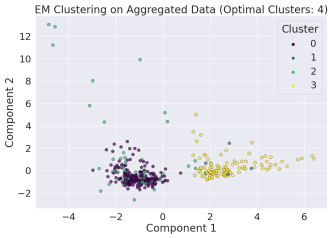

Fig. 12. EM Clustering: DS 2

IV. Dataset 2: School Shootings and Donation Patterns

A. *Hypothesis: Dataset 2*

Given the high number of features in Dataset 2, DR will be crucial for uncovering meaningful patterns. PCA followed by EM should yield the most interpretable clusters by capturing variance in key features like donation spikes and casualties. K-means may produce broader but less distinct groupings. ICA applied on PCA reduced data may highlight unique feature combinations, but cluster separation could be limited due to the high independence in feature representation.

B. *Expectation Maximization: Dataset 2*

Applying EM to Dataset **2** yielded differing scores: BIC indicated **3** clusters, while AIC suggested **9**. We chose **3** clusters for a clearer grouping, reflecting broader trends rather than finer distinctions. Figure 12 illustrates the resulting clusters, highlighting distinct donation patterns in response to school shooting characteristics. Cluster **0** shows frequent but smaller donation spikes with strong pro-gun affiliations and high casualty counts. Cluster **1** includes fewer incidents and lower casualties but has significantly higher average transaction amounts, likely driven by a few high-value donations. Cluster **2** features moderate casualties with frequent donation spikes, indicating sustained pro-gun funding.

These results partly support the hypothesis. While clustering reveals distinct donation behaviors, limited cluster separation suggests that high feature complexity may hinder accurate pattern prediction.

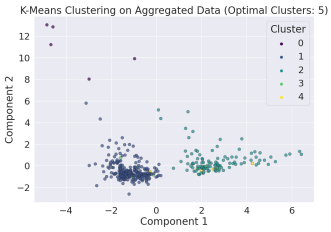

Fig. 13. K-means Clustering: DS 2

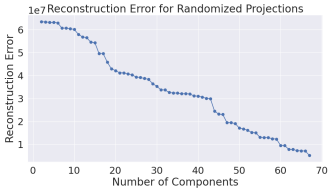

Fig. 14. RP: Reconstruction Error

C. *K-means: Dataset 2*

Analysis of K-means on Dataset 2 revealed an optimal cluster count of **5** determined through the silhouette score. The clustering revealed meaningful groupings in donation behaviors relative to the characteristics of school shootings (figure 13). Cluster **1** depicts incidents with high casualties and a large number of pro-gun affiliations but low donation spikes. This may indicate consistent but lower-value donations. Cluster **2** stands out with both high pro-gun affiliations and the highest donation spikes, likely reflecting targeted responses to specific incidents. Meanwhile, Clusters **0**, **3**, and **4** feature lower casualty incidents, fewer pro-gun affiliations, moderate average transaction amounts, and minimal donation spikes.

This clustering aligns with the hypothesis by suggesting distinct donation patterns tied to incident severity and party affiliation. Though the diversity within clusters indicates complex underlying factors affecting donation responses.

D. *Randomized Projections: Dataset 2*

The reconstruction error was minimized with **67** components (Figure 14). This only slightly reduces the dimensionality from the original of **68**.

This outcome is consistent with expectations, as RP tends to retain a dimensional space close to the original since the algorithm calculates reconstruction error using the pseudo-inverse of the projection matrix, which contains the same dimensions as the original space. The implication of working in a **67-dimensional space** is that while dimensionality is slightly reduced, we still maintain most of the dataset's complexity. This enables future clustering implementations to capture more detailed clustering patterns without significant information loss. The drawback is the reduction of only **1** dimensions will not significantly reduce computational efficiency.

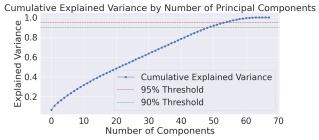

Fig. 15. PCA: Explained Variance

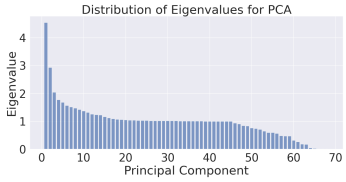

Fig. 16. PCA: Eigenvalue Distribution

E. *Principle Component Analysis: Dataset 2*

After running PCA on dataset 2, we observe from Figure 15 that **55** components are needed in order to account for **95%** of the variance within the model. The eigenvalue distribution (Figure 16) in PCA highlights a steep drop-off in variance contribution across components, with only a few components holding substantial information. As more components are added, the eigenvalue scores decrease, meaning they contribute less proportionally.

Comparing these results to results from RP in section IV-C, PCA does indeed perform significantly better by reducing the dimensionality of components from **68** to **55** while still maximizing over **95%** of the variance.

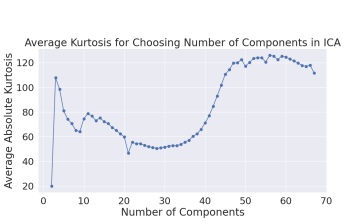

Fig. 17. ICA: Kurtosis

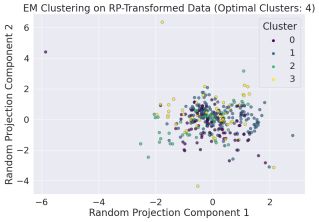

Fig. 18. EM Clustering: RP

F. *Independent Component Analysis: Dataset 2*

ICA implementation revealed that **56** maximizes the average kurtosis (Figure 17). This aligns closely with PCA analysis results, where **55** components captured **95%** of explained variance. The similarity suggests that both methods identified a similar dimensional structure, although PCA focuses on maximizing variance, while ICA maximizes non-Gaussianity, or kurtosis. The selection of **56** components highlights independent features that may capture unique patterns in the data. These findings are consistent with our hypothesis that ICA would be able to effectively reduce dimensionality by maximizing kurtosis

G. *EM Clustering on RP Transformed Dataset 2*

The RP algorithm was ran on dataset 2 after determining an optimal number of Clusters based on a BIC value of **4**. Additionally, the component count of **67** determined in Section IV-A was used in the implementation.

Cluster analysis (Figure 18) reveals feature insights through various grouping patterns. Cluster **0** has a high number of casualties, moderate transaction amounts, and occasional donation spikes, suggesting support from a mid-sized pro-gun advocates. Cluster **1** represents substantial casualties with high donation amounts and numerous pro-gun donations, suggesting a more intense response from pro-gun groups. Cluster **2** has fewer casualties, lower donation count, and infrequent donation spikes.

The results aligned fairly well to the hypotheses. As anticipated, RP clusters show less distinction than expected, supporting the hypothesis that Random Projection may capture broader trends but struggles with distinct feature separation in complex data.

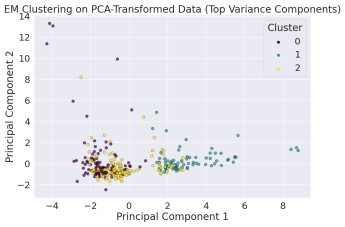

Fig. 19. EM Clustering: PCA

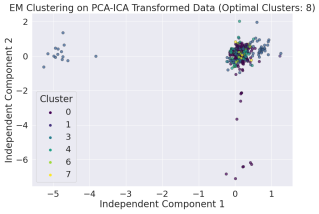

Fig. 20. EM Clustering: ICA

H. *EM Clustering on PCA Transformed Dataset 2*

Applying EM to Dataset 2 after PCA reduced it to **55** components (capturing over **95%** of the variance) identified **3**

optimal clusters based on BIC and AIC scores. Selecting fewer clusters aligns with BIC's preference for model simplicity.

Cluster 0 represents smaller donation amounts with moderate pro-gun affiliation and fewer donation spikes, while **Cluster 1** shows a significant spike in large donations but fewer casualty incidents, suggesting high value but low frequency responses. **Cluster 2** has moderate donation levels with a higher number of incidents and donation spikes, indicating a steady response to events.

These results support the hypothesis that PCA combined with EM would yield interpretable clusters. PCA effectively captured donation patterns, producing well-defined clusters and distinctive behaviors.

I. EM Clustering on ICA Transformed Dataset 2

After applying ICA on PCA reduced data with **55** components for both algorithms, EM identified **8** clusters as optimal based on BIC scores. Notably, results from running this implementation in dataset 1 revealed that plotting the 2 components with the highest kurtosis values did not produce distinct clusters. Therefore the clustering (Figure 20) represents components that maximize variance.

The results support the hypothesis that ICA following PCA might reveal unique feature combinations event though cluster separation remains somewhat limited. This outcome suggests that while ICA may still slightly struggle to uncover distinct patterns due to the high feature independence.

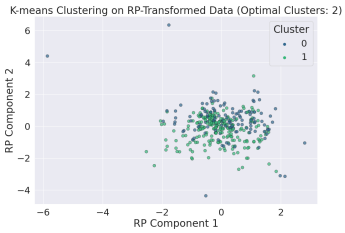

Fig. 21. K-means Clustering: RP

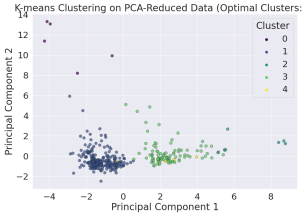

Fig. 22. K-means Clustering: PCA

J. K-means Clustering on RP Transformed Dataset 2

Applying K-means on the RP transformed data with **67** (calculated in Section IV-C) components yielded an optimal cluster count of **2** based on Silhouette score. Results of the clustering are displayed in Figure 21. Of note, **Cluster 0** includes incidents with moderate casualties but significantly higher average transaction amounts and a notable pro-gun affiliation. Conversely, **Cluster 1** has higher casualties, lower transaction amounts, and fewer donation spikes.

Although results only produced **2** clusters, K-means produced broad groupings with distinct donation characteristics, which is slightly consistent with the hypothesis. A large amount of complex features may still be slightly preventing high clarity separable clustering.

K. K-means Clustering on PCA Transformed Dataset 2

Applying K-means on PCA transformed data of **55** components, which was calculated in Section 15, resulted in an optimal cluster count of **5** based on the Silhouette score.

Notably, **Cluster 2** shows a unique number of high transaction amounts with relatively low casualties, suggesting fewer, larger donations. **Cluster 3** captures the most substantial pro-gun affiliation and highest donation spikes, which could explain highly publicized incidents driving donations. This clustering approach produced more linearly separable groups then previous implementations, aligning well with the hypothesis that PCA would enhance cluster separability by reducing dimensional noise.

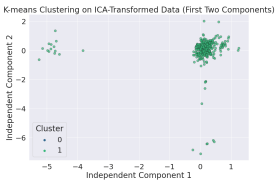

Fig. 23. K-means: ICA

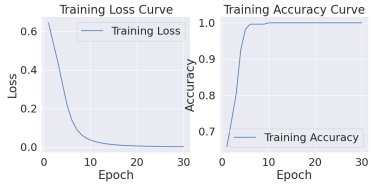

Fig. 24. NN: Training Loss/Accuracy

L. K-means Clustering on ICA Transformed Dataset 2

Applying K-means on ICA transformed dataset, which was initially reduced using PCA with **55** components, resulted in an optimal cluster count of **2** based on the Silhouette score. Notably, unlike previous ICA trials in **Dataset 1**, graphed components (Figure 23) were not selected based on kurtosis, as those trials yielded clusters with weak separation.

Contrary to the hypothesis, K-means produced more linearly separable clusters than EM in this ICA transformed dataset. This may be because EM assumes Gaussian distributions that may not align as well with the sparsity and independence introduced by ICA, which led to less distinct clustering.

Table II and figure 24 display the results of the original NN implementation after tuning hyperparameters. These figures can be used as a metric for comparison and analysis.

TABLE II
NEURAL NETWORK: ORIGINAL RESULTS

Class	Precision	Recall	F1-score	Support
No Donation Spike	0.92	0.98	0.95	48
Donation Spike	0.95	0.83	0.88	23
Accuracy			0.93	71
Macro Avg	0.94	0.90	0.92	71
Weighted Avg	0.93	0.93	0.93	71

V. DR TECHNIQUES ON NEURAL NETWORK LEARNER

In this section, the optimal number of components is determined by maximizing test accuracy, as opposed to relying on reconstruction error, variance, or kurtosis. Since this is a supervised learning task, the data is divided into training and test sets with an 80/20 split. Next, hyperparameters are tuned, and back-propagated into the NN while applying various

DR algorithms with the goal of producing components that enhance test accuracy.

A. *Hypothesis: Neural Network Implementation*

Implementing DR algorithms in the NN should reduce computational cost due to fewer features, and render the training process more efficient in comparison to original clustering (Figures 12 and 13), and in previous NN implementations. However, given the already high test accuracy of **92.96%**, significant improvement in accuracy is unlikely. The introduction of DR will introduce slight improvements in accuracy, but overall performance should remain competitive. However, Clustering with EM and K-means will refine feature patterns.

B. *RP: Neural Network Implementation*

The RP algorithm reduced the feature space from **68** to **20** components based on maximizing test accuracy. Due to the inherent randomness in projections, RP yielded varying results across runs, with test accuracy fluctuating between **85.92%** and **92.96%**. Some runs showed significant oscillations and struggled with convergence in the training loss and accuracy curves, suggesting that RP occasionally removed essential components needed for stable training. This variability reflects how random projections can lead to information loss and overall performance.

EM-RP Clustering 25: Running EM over **10** clusters (derived from BIC), RP reduced components from **68** to **20** to maximize test accuracy. This produced less distinct, linearly separable clusters compared to the initial EM implementation (Figure 12). Despite dimension reduction, the stochastic nature of RP may have omitted meaningful features, leading to incomplete cluster separation. This partial separation highlights RP's limitations in fully capturing feature complexity when dimensions are significantly reduced.

K-means Clustering 26:K-means clustering with **10** clusters on RP-reduced data showed compact but less distinct clusters, likely due to variance lost through random projection. In contrast, clusters in the original implementation (Figure 13) were better defined and more linearly separable. While RP reduced computational costs, it compromised clustering precision by removing feature distinctions, underscoring the trade-off between DR and cluster separability.

Consistent to the hypothesis, RP reduced computational cost but did not significantly improve accuracy. The variability in test accuracy and reduced cluster separability confirm that random projections can impact model stability and clustering precision. While RP improved training efficiency, this came at the cost of some clustering quality.

C. *PCA: Neural Network Implementation*

Applying PCA reduced the number of components from **68** to **10**, chosen to maximize test accuracy. This approach led to clear improvements, with test accuracy increasing to **98.59%** from the original NN test accuracy of **92.96%** (Table III) without DR.

The training loss and accuracy curves mirror those from the original implementation (Figure 24), maintaining a nearly identical trajectory throughout the epochs. he network achieved **100%** training accuracy by epoch **20**, with wall clock training time remaining nearly the same at **0.77** seconds, compared to **0.76** seconds previously.

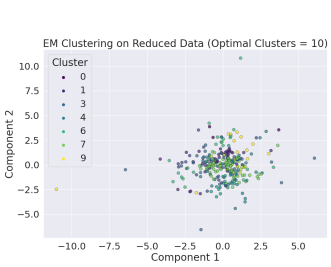

Fig. 25. EM Clustering: RP

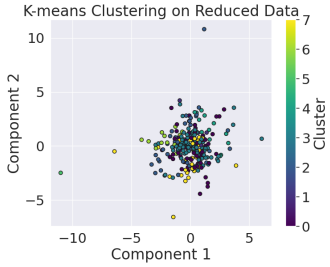

Fig. 26. K-means Clustering: RP

TABLE III
NEURAL NETWORK: PCA REDUCTION

Class	Precision	Recall	F1-score	Support
No Donation Spike	1.00	0.98	0.99	48
Donation Spike	0.96	1.00	0.98	23
Accuracy			0.99	71
Macro Avg	0.99	0.99	0.98	71
Weighted Avg	0.99	0.99	0.99	71

EM-PCA Clustering 27: Results from EM clustering reduced the feature space from **68** to **10** components, leading to faster wall clock time woth **10** clusters (based on BIC). This reduction produced separable clusters but required longer runtime, highlighting PCA's ability to reduce noise and improve clustering efficiency in a lower-dimensional space. Notably, the initial EM clustering without DR 13) yielded similarly separable clusters with a longer runtime.

K-means clustering on PCA-reduced data with **10** clusters (chosen by silhouette score) produced clear, linearly separable clusters. The PCA reduction from **68** to **10** components removed noise, resulting in cohesive clusters with minimal overlap. In contrast, initial K-means clustering without PCA (13) displayed less distinct cluster separation, suggesting that PCA enhanced clustering by simplifying the feature space.

The results align well with the hypothesis. PCA effectively reduced the feature space, which improved test accuracy from the original implementation and improved computational efficiency. As predicted, PCA's DR enhanced clustering precision by reducing noise. The clustering results further confirmed the hypothesis, as both EM and K-means produced more distinct clusters with PCA than without, supporting the idea that PCA would enhance feature separability.

D. *ICA: Neural Network Implementation*

ICA reduced the feature space from **68** to **5** components while maintaining a high test accuracy of **97.18%** (Table IV). The training loss and accuracy curves closely match those of the original NN without DR (Figure 24), indicating that ICA preserved critical patterns for model performance. The training

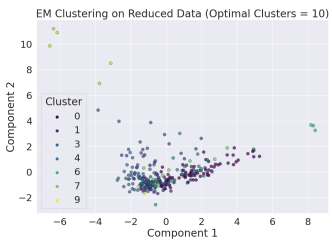

Fig. 27. EM Clustering: PCA

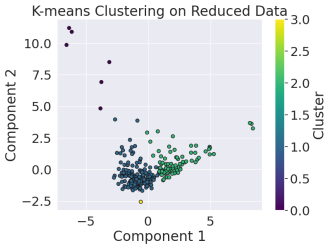

Fig. 28. K-means Clustering: PCA

time decreased slightly to **0.62 seconds**, a modest improvement from **0.76**. This suggests that with larger datasets, reducing from **68** to **5** components yields computational efficiency.

TABLE IV
NEURAL NETWORK: ICA REDUCTION

Class	Precision	Recall	F1-score	Support
No Donation Spike	0.96	0.98	0.98	48
Donation Spike	0.96	0.96	0.96	23
Accuracy			0.97	71
Macro Avg	0.97	0.97	0.97	71
Weighted Avg	0.97	0.97	0.97	71

EM-ICA Clustering 29: Clustering on ICA transformed data with **5** components reveals a tradeoff between test accuracy and cluster linear separability. While ICA achieved high test accuracy in previous NN implementations, the clusters lack clear linear separation. ICA maximizes statistical independence, which may retain components enhancing predictive power but not necessarily suitable for clustering. This contrasts with prior EM clustering on aggregated data (Figure 12), which showed more distinct separation. [2]

Similar to EM clustering results, K-means on ICA-transformed data also struggled to form distinct, linearly separable clusters. The ICA components do not align well with K-means' distance-based partitioning approach, making it difficult for the algorithm to establish meaningful clusters in a reduced space lacking natural group structure.

K-means-ICA Clustering 30: Similar to EM clustering results, K-means on ICA reduced data also struggled to form distinctly separable clusters. The ICA components do not align well with K-means' distance-based partitioning approach, making it difficult for the algorithm to establish meaningful clusters in a reduced space lacking natural group structure.

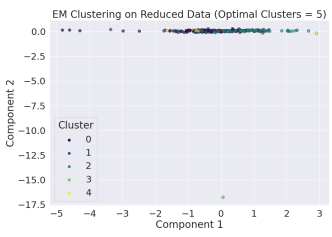

Fig. 29. EM-ICA Clustering

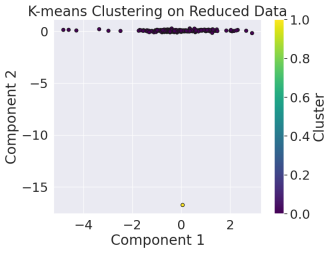

Fig. 30. K-means-ICA Clustering

I expected DR with ICA to enhance clustering without sacrificing cluster separability, but high test accuracy (**97.18%**) came at the expense of poorly defined clusters. ICA's focus on statistical independence preserved essential features for prediction. However, this did not translate to clear clustering structure, highlighting a tradeoff between maximizing predictive performance and achieving distinct clusters.

CONCLUSION

Unsupervised clustering and DR techniques were applied to two datasets—one focusing on political contributions and the other on school shootings and donation patterns. The goal was to study DR method's (PCA, RP, and ICA) relationship between clustering performance and DR while using EM and K-means algorithms. Also, to understand the role of DR in NN performance.

Overall, PCA emerged as the most effective DR technique across both datasets. In Dataset 1, which contained only **8** features, PCA reduced the feature space to **6**, while RP and ICA reduced it to **7**. K-means clustering notably improved after PCA reduction compared to the original clustering results, whereas EM clustering performed slightly worse with reduced linear separability. In Dataset **2**, containing **68** features, PCA achieved substantial DR to **55** components, outperforming RP **(55)** and ICA **(56)** in clustering. Here, K-means clustering on PCA-reduced data produced the best results, with more linearly separable clusters than EM.

In the NN context, DR techniques were applied to maximize test accuracy on Dataset 2. This led to significant feature reduction, with RP reducing the space from **68** to **20**, PCA to **10**, and ICA to **5** components. Interestingly, while maximizing test accuracy, the DR transformed data did not necessarily yield efficient clustering results. PCA reduced data once again performed the best for both EM and K-means, with K-means providing slightly clearer clustering than EM. ICA, despite achieving substantial DR, resulted in poorly grouped clusters, suggesting that high test accuracy does not always translate to optimal clustering efficiency. Future work could further explore more adaptive DR methods that balance feature reduction with clustering patterns, further bridging the gap in high dimensional socio-political data analysis.

REFERENCES

[1] The Washington Post, "School shootings database," *GitHub Repository*, https://github.com/washingtonpost/data-school-shootings (accessed Sep. 19, 2024).

[2] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997, pp. 191-197.

[3] "sklearn.preprocessing.StandardScaler," *scikit-learn: Machine Learning in Python*. https://scikit-learn.org/dev/modules/generated/sklearn.preprocessing.StandardScaler.html (accessed Nov. 9, 2024).

[4] "No Straight Lines Here! The Wacky World of Non-Linear Manifold Learning," *Georgia Tech OMSCS 7641 Blog*, Mar. 10, 2024. Available: https://sites.gatech.edu/omscs7641/2024/03/10/no-straight-lines-here-the-wacky-world-of-non-linear-manifold-learning/. (Accessed: Nov. 9, 2024).

[5] S. Liu and T. LaGrow, "How to Evaluate Features after Dimensionality Reduction?" *Georgia Tech OMSCS 7641 Blog*, Mar. 7, 2024. Available: https://sites.gatech.edu/omscs7641/2024/03/07/how-to-evaluate-features-after-dimensionality-reduction/. (Accessed: Nov. 9, 2024).