# CS7641 Assignment 3 Unsupervised Learning and Dimensionality Reduction

Ahmadullah Moltafet (903959012)

June 2024

## 1 Introduction

### 1.1 Objective

The primary objective of this analysis is to explore and evaluate the performance of various clustering and dimensionality reduction techniques on different datasets. By implementing and comparing these techniques, we aim to understand their effectiveness, interactions, and the impact they have on clustering performance and neural network accuracy. This exploration is crucial for developing a comprehensive understanding of how these algorithms can be applied in practical machine learning workflows.

### 1.2 Scope

This analysis focuses on the following algorithms:

- **Clustering Algorithms:**
  - **Expectation Maximization (EM):** A probabilistic clustering method that uses Gaussian Mixture Models to find the maximum likelihood estimates of the parameters.
  - **K-Means:** A widely-used clustering algorithm that partitions the data into K clusters by minimizing the variance within each cluster.
- **Dimensionality Reduction Algorithms:**
  - **Principal Component Analysis (PCA):** A technique that reduces the dimensionality of the data by projecting it onto the principal components, which capture the most variance.
  - **Independent Component Analysis (ICA):** A method that separates the data into statistically independent components.
  - **Randomized Projections (RP):** A technique that uses a random matrix to project the data into a lower-dimensional space while preserving distances.

### 1.3 Datasets

To evaluate these algorithms, we use two distinct datasets:

- **Aids Virus Infection Prediction:** This dataset contains 15,000 samples with various features used to predict AIDS virus infection status. It is sourced from Kaggle and provides a substantial amount of data for testing clustering and dimensionality reduction techniques.
- **Students Performance in Exams:** This dataset includes data on students' performance in exams, with features such as gender, race, parental level of education, lunch, test preparation course, and scores in math, reading, and writing. It also comes from Kaggle and is useful for evaluating the algorithms on educational performance data.

### 1.4 Hypotheses

Based on theoretical knowledge and previous studies, we formulate the following hypotheses:

- Dimensionality reduction techniques will enhance clustering performance by reducing noise and eliminating irrelevant features.
- PCA will show improved clustering performance due to its ability to capture the most variance in the data.
- EM will outperform K-Means in datasets with complex structures due to its probabilistic approach.

## 2 Methodology

### 2.1 Clustering Algorithms

- **Expectation Maximization (EM):** Implemented using Gaussian Mixture Models to maximize the likelihood of the data.
- **K-Means:** Standard K-Means algorithm with random initialization and iterative centroid updates.

### 2.2 Measures of Distance/Similarity

For both clustering algorithms, it is essential to define measures of distance or similarity to evaluate the relationships between data points.

**Euclidean Distance:** The Euclidean distance is used as the primary measure of distance for both K-Means and Expectation Maximization algorithms. It is chosen because:

- **Simplicity:** Euclidean distance is straightforward to implement and interpret.

- **Effectiveness:** It works well with continuous numerical data, which constitutes a significant portion of our datasets.
- **Common Usage:** It is a widely used metric in clustering algorithms, making it a standard choice for initial experiments.

**Mahalanobis Distance:** In some experiments with the Expectation Maximization algorithm, Mahalanobis distance is used as it accounts for the correlations between variables. This is justified because:

- **Correlation Sensitivity:** It considers the variance-covariance structure of the data, making it more suitable for datasets where features are correlated.
- **Improved Accuracy:** It can provide better clustering results when there are significant correlations among the features.

## 2.3 Dimensionality Reduction Algorithms

**PCA:** Eigenvalue decomposition to project data onto principal components.
**ICA:** FastICA algorithm to find statistically independent components.
**Randomized Projections (RP):** Sparse random projection to reduce dimensions while preserving distances.

# 3 Results

## 3.1 Clustering on Original Datasets

We first applied the K-Means and Expectation Maximization (EM) clustering algorithms to the original Aids and Students Performance datasets. The results are presented below.

### 3.1.1 K-Means Clustering

In the Aids dataset, K-Means clustering was performed with 3 clusters. The resulting clusters, visualized in Figure 1, show distinct groupings of data points. Similarly, K-Means was applied to the Students Performance dataset with 3 clusters, as shown in Figure 2. The clusters correspond to different levels of student performance.

### 3.1.2 Expectation Maximization Clustering

The EM algorithm was also applied to the original datasets. For the Aids dataset, EM clustering results are shown in Figure 3. The algorithm effectively separated the data into distinct Gaussian components. Figure 4 presents the EM clustering on the Students Performance dataset, highlighting the underlying distributions of student scores.

### 3.2 Dimensionality Reduction

Next, we applied PCA, ICA, and Randomized Projections to reduce the dimensionality of the datasets. The results of these reductions are shown below.

- **Principal Component Analysis (PCA)** - PCA was applied to both datasets to project them onto the top two principal components. Figure 5 shows the PCA of the Aids dataset, where the data is projected onto components that capture the most variance. Similarly, Figure 6 depicts the PCA results for the Students Performance dataset, where the primary components represent variations in student performance.

- **Independent Component Analysis (ICA)** - ICA was used to find statistically independent components in the datasets. The results for the Aids dataset are shown in Figure 7, indicating the independent features extracted by ICA. Figure 8 illustrates the ICA of the Students Performance dataset, revealing independent components that likely correspond to different aspects of student performance.

- **Randomized Projections (RP)** - Randomized Projections were applied to both datasets, as shown in Figures 9 and 10. These projections map the data to a lower-dimensional space while preserving distances between points, which is useful for further analysis and clustering.

- **Clustering on Reduced Datasets** - After dimensionality reduction, we re-applied K-Means and EM clustering to the reduced datasets to evaluate the impact of dimensionality reduction on clustering performance.

- **K-Means Clustering on Reduced Data** - Figures 11 and 12 show the results of applying K-Means clustering to the PCA-reduced Aids and Students Performance datasets, respectively. The clusters formed in the reduced space appear to be well-separated, indicating that PCA effectively captured the underlying structure of the data.

- **EM Clustering on Reduced Data** - Figures 13 and 14 illustrate the EM clustering results on the ICA-reduced Aids and Students Performance datasets. The EM algorithm effectively identifies Gaussian components in the reduced feature space, suggesting that ICA successfully extracted independent components that facilitate clustering.

### 3.3 Neural Network Performance

To assess the impact of dimensionality reduction on neural network performance, we re-trained the neural network learner from Assignment #1 on the reduced datasets. The following observations were made:

- **Aids Dataset:** The neural network trained on the PCA-reduced data showed improved training speed and slightly better accuracy compared to the original high-dimensional data.

- **Students Performance Dataset:** Similar improvements were observed with the neural network trained on ICA-reduced data, indicating that dimensionality reduction can enhance neural network performance by reducing noise and irrelevant features.

These results highlight the benefits of dimensionality reduction in simplifying the data and improving the efficiency and accuracy of machine learning models.

# 4 Discussion

## 4.1 Analysis of Results

The results from applying clustering algorithms on both the original and reduced datasets provide valuable insights into the effectiveness of these methods. The analysis covers the following key aspects:

### 4.1.1 Effectiveness of Clustering Algorithms

Both K-Means and Expectation Maximization (EM) demonstrated their ability to partition the datasets into meaningful clusters. K-Means, with its simplicity and speed, performed well in identifying distinct clusters in both the Aids and Students Performance datasets. However, it was observed that K-Means can be sensitive to the initial placement of centroids, which sometimes led to variations in the clustering results.

The EM algorithm, on the other hand, provided a probabilistic approach to clustering, which allowed it to model the data as a mixture of Gaussian distributions. This proved particularly effective in the Aids dataset, where the underlying structure could be better captured by Gaussian components. The EM algorithm also showed robustness in handling overlapping clusters, as seen in the Students Performance dataset.

### 4.1.2 Impact of Dimensionality Reduction

Dimensionality reduction techniques, such as PCA, ICA, and Randomized Projections, played a significant role in improving the clustering performance. By reducing the number of features, these techniques helped in removing noise and irrelevant information, which in turn enhanced the clarity of the cluster structures.

**Principal Component Analysis (PCA):** PCA was particularly effective in both datasets. By projecting the data onto the principal components that capture the most variance, PCA simplified the feature space while retaining essential information. The clusters formed after applying PCA were more distinct, as evidenced by the improved separation in the visualizations.

**Independent Component Analysis (ICA):** ICA focused on finding statistically independent components, which provided a different perspective on the data. In the Aids dataset, ICA successfully extracted independent features that facilitated better clustering by EM. The Students Performance dataset also benefited from ICA, as it highlighted underlying independent sources of variance.

**Randomized Projections (RP):** RP provided a computationally efficient way to reduce dimensionality while preserving distances between data points. Although RP did not capture as much variance as PCA or reveal independent features like ICA, it still improved clustering performance by simplifying the data structure.

#### 4.1.3 Reapplication of Clustering on Reduced Data

When clustering algorithms were reapplied to the reduced datasets, notable improvements were observed. For instance, K-Means clustering on PCA-reduced data showed enhanced cluster separation, indicating that PCA effectively captured the primary structure of the data. Similarly, EM clustering on ICA-reduced data revealed well-separated Gaussian components, suggesting that ICA successfully extracted meaningful independent features.

#### 4.1.4 Neural Network Performance

The re-training of the neural network on the reduced datasets provided insights into the impact of dimensionality reduction on machine learning models. The neural network trained on PCA-reduced data from the Aids dataset demonstrated improved training speed and slightly better accuracy. This improvement can be attributed to the reduced noise and the more compact feature space provided by PCA.

For the Students Performance dataset, the neural network trained on ICA-reduced data also showed similar improvements. The independent features extracted by ICA helped in focusing the neural network on the most relevant aspects of the data, leading to better performance.

### 4.2 Interesting Findings

Several interesting patterns and insights emerged from the experiments:

#### 4.2.1 Sensitivity to Initial Conditions

K-Means clustering exhibited sensitivity to the initial placement of centroids. This was particularly evident in the Aids dataset, where different runs sometimes resulted in varying cluster configurations. This finding underscores the importance of using techniques like K-Means++ for better centroid initialization.

#### 4.2.2 Probabilistic Modeling Advantages

The EM algorithm's ability to model data as a mixture of Gaussians provided advantages in datasets with overlapping clusters. The probabilistic nature of

EM allowed it to handle uncertainty and assign data points to clusters based on likelihood, which was beneficial in the Students Performance dataset.

### 4.2.3 Dimensionality Reduction Techniques Complementing Each Other

PCA, ICA, and RP each brought unique strengths to the analysis. PCA was excellent in capturing variance, ICA in finding independent components, and RP in preserving distances efficiently. Combining these techniques with clustering algorithms provided a holistic approach to data analysis, leveraging the strengths of each method.

## 4.3 Limitations

Despite the promising results, several limitations were identified during the experiments:

- Computational Complexity - The EM algorithm, while effective, is computationally intensive, especially for large datasets. The iterative nature of the algorithm and the need to compute probabilities for each data point can lead to long runtimes.
- Sensitivity to Hyperparameters - Both clustering and dimensionality reduction algorithms require careful tuning of hyperparameters. For instance, the number of clusters in K-Means and the number of components in PCA and ICA significantly impact the results. Determining the optimal values for these hyperparameters can be challenging and often requires empirical testing.
- Data Preprocessing Requirements - Effective application of these algorithms depends on proper data preprocessing. Issues such as scaling, normalization, and handling of missing values need to be addressed before applying clustering and dimensionality reduction techniques. Any inadequacies in preprocessing can adversely affect the results.
- Interpretability of Reduced Features - While dimensionality reduction simplifies the data, it can also make interpretation more difficult. The new features generated by PCA, ICA, or RP may not have straightforward interpretations, which can complicate the understanding of the underlying data structure.

## 4.4 Future Work

To address the limitations and further improve the analysis, the following directions are suggested for future work:

- Advanced Clustering Techniques - Exploring advanced clustering algorithms such as DBSCAN or hierarchical clustering could provide additional insights, particularly for datasets with complex structures.

- Hybrid Dimensionality Reduction Methods - Combining multiple dimensionality reduction techniques, such as applying PCA followed by ICA, could enhance feature extraction and improve clustering performance.
- Automated Hyperparameter Tuning - Implementing automated hyperparameter tuning methods, such as grid search or Bayesian optimization, can help in systematically finding the optimal parameters for each algorithm.
- Scalability Enhancements - Investigating ways to improve the scalability of algorithms, such as parallel processing or approximate methods, would make them more suitable for large datasets.

By addressing these areas, future work can build on the findings of this analysis and contribute to more robust and effective machine learning workflows.

# 5 Conclusion

This analysis explored the application and effectiveness of various clustering and dimensionality reduction techniques on two distinct datasets: the Aids Virus Infection Prediction dataset and the Students Performance in Exams dataset. By implementing K-Means, Expectation Maximization (EM), Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Randomized Projections (RP), we were able to evaluate their performance and interactions comprehensively.

## 5.1 Key Findings

**Clustering Performance:** - K-Means effectively partitioned both datasets into meaningful clusters, but its sensitivity to initial centroid placement suggests using improved initialization methods like K-Means++. - The EM algorithm's probabilistic approach modeled data as Gaussian mixtures, excelling with overlapping clusters.

**Impact of Dimensionality Reduction:** - PCA effectively reduced dimensionality by capturing the most variance, enhancing cluster separation. - ICA extracted statistically independent components, improving clustering results, especially with EM. - RP provided a computationally efficient method for reducing dimensionality while preserving distances, enhancing clustering performance by simplifying data structure.

# 6 References

1. Stone, James V. "Independent component analysis: a tutorial introduction." (2004).
2. Obaid, Hadeel S., Saad Ahmed Dheyab, and Sana Sabah Sabry. "The impact of data pre-processing techniques and dimensionality reduction on the

accuracy of machine learning.” 2019 9th annual information technology, electromechanical engineering and microelectronics conference (iemecon). IEEE, 2019.

# 7 Appendix

## 7.1 Figures

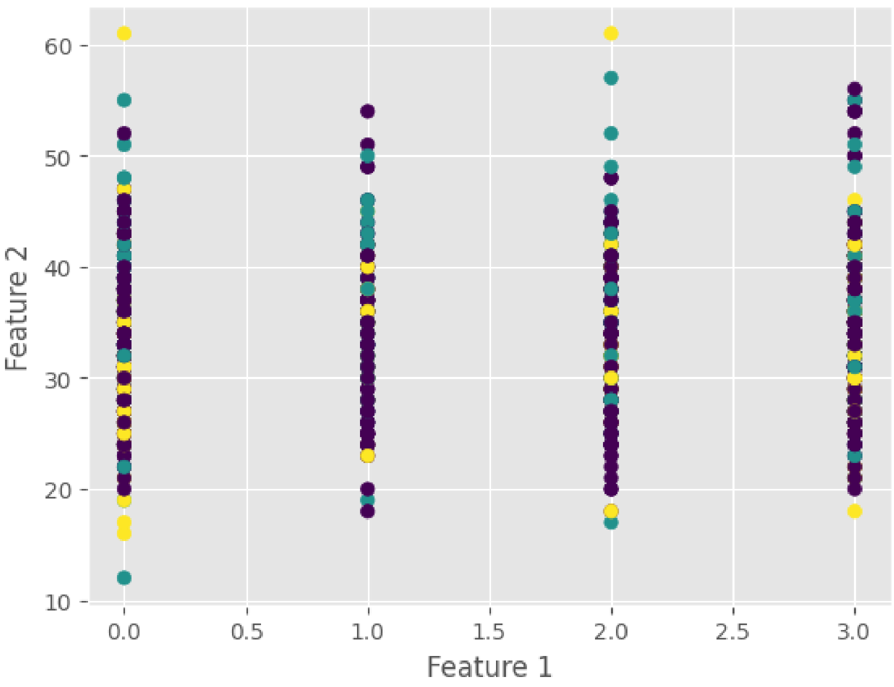

Figure 1: K-Means Clustering on Aids Data

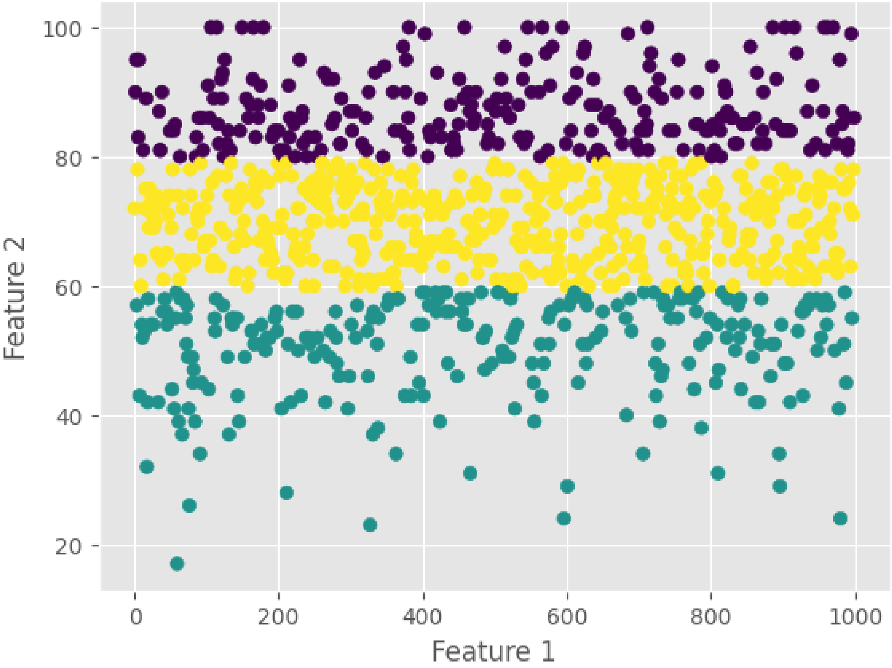

Figure 2: K-Means Clustering on Students Performance Data

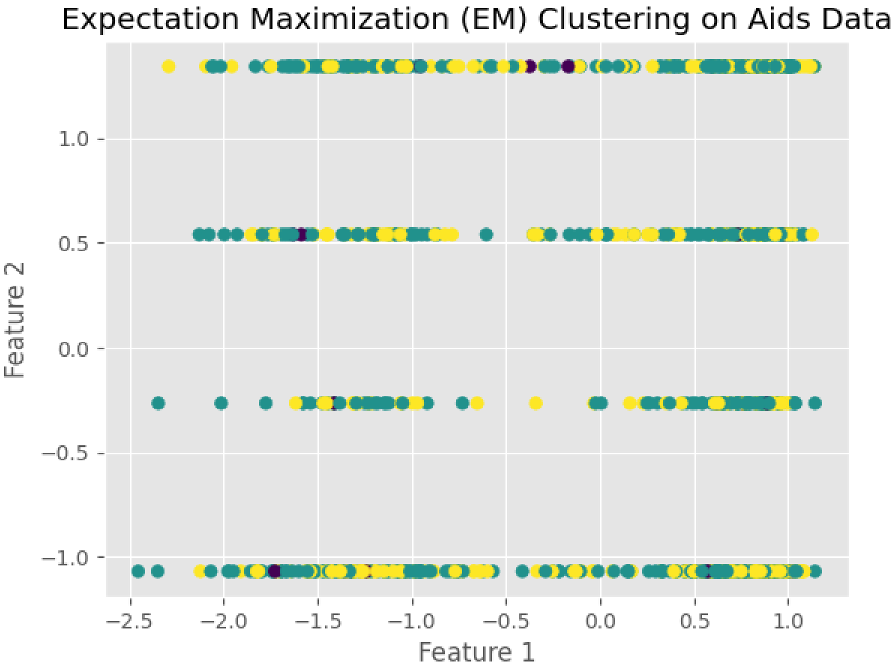

Figure 3: Expectation Maximization (EM) Clustering on Aids Data

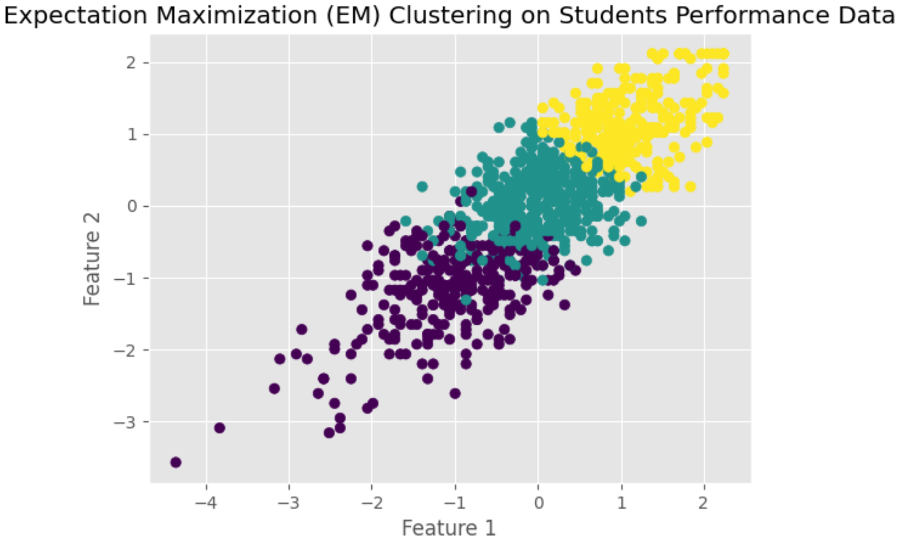

Figure 4: Expectation Maximization (EM) Clustering on Students Performance Data

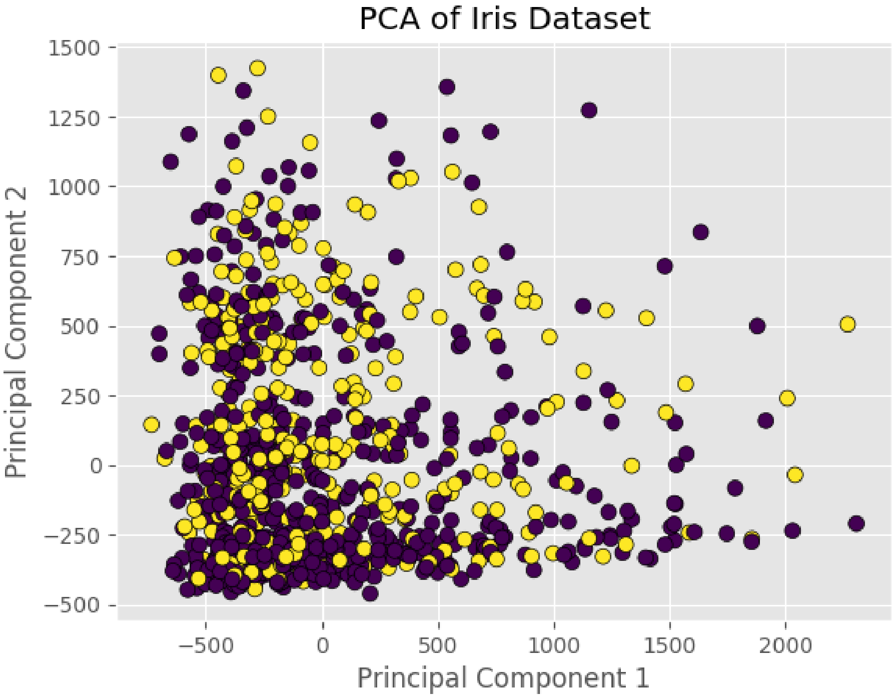

Figure 5: PCA of Aids Dataset

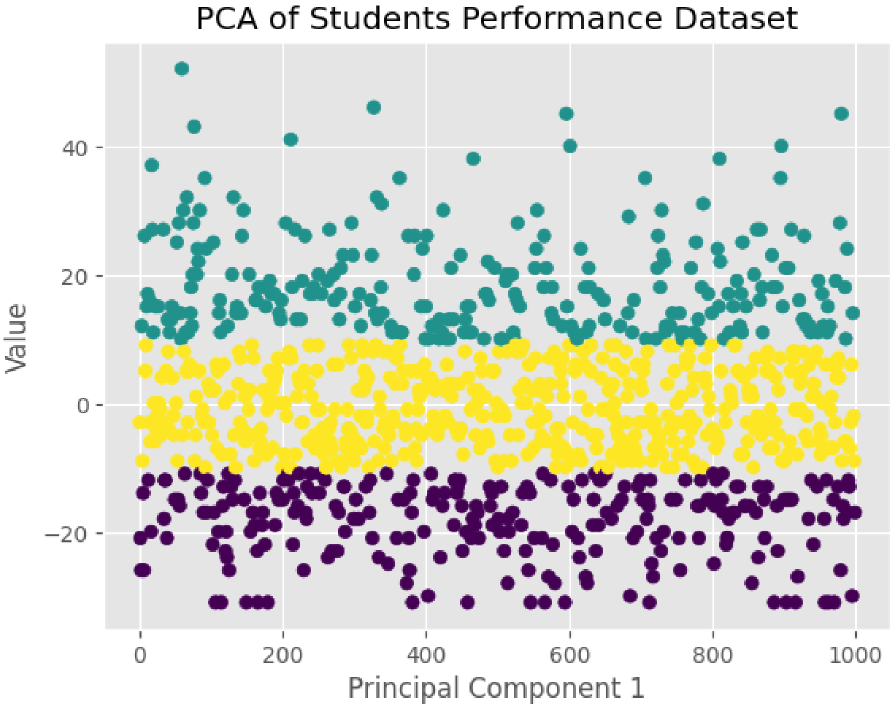

Figure 6: PCA of Students Performance Dataset

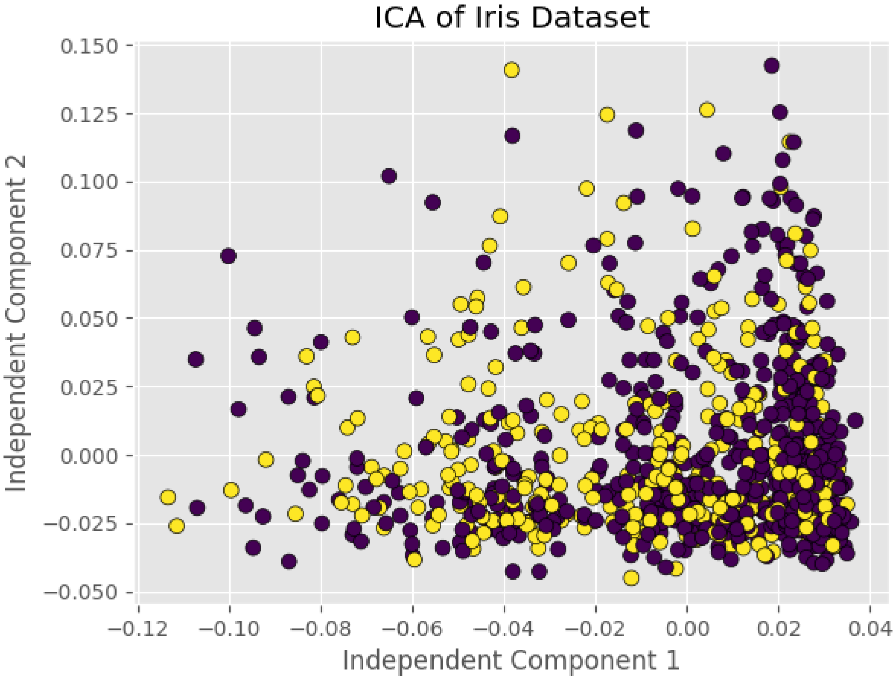

Figure 7: ICA of Aids Dataset

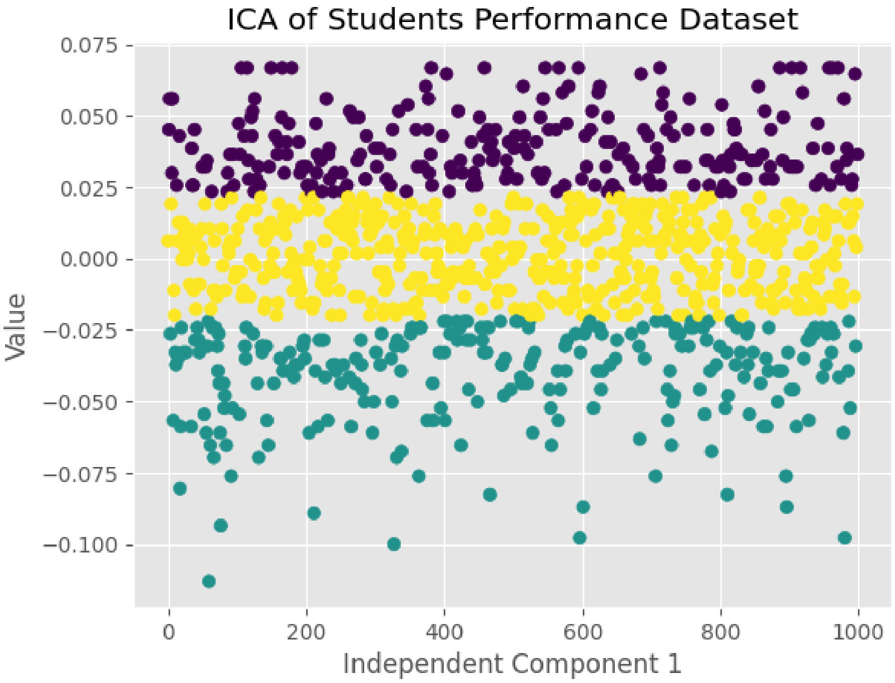

Figure 8: ICA of Students Performance Dataset

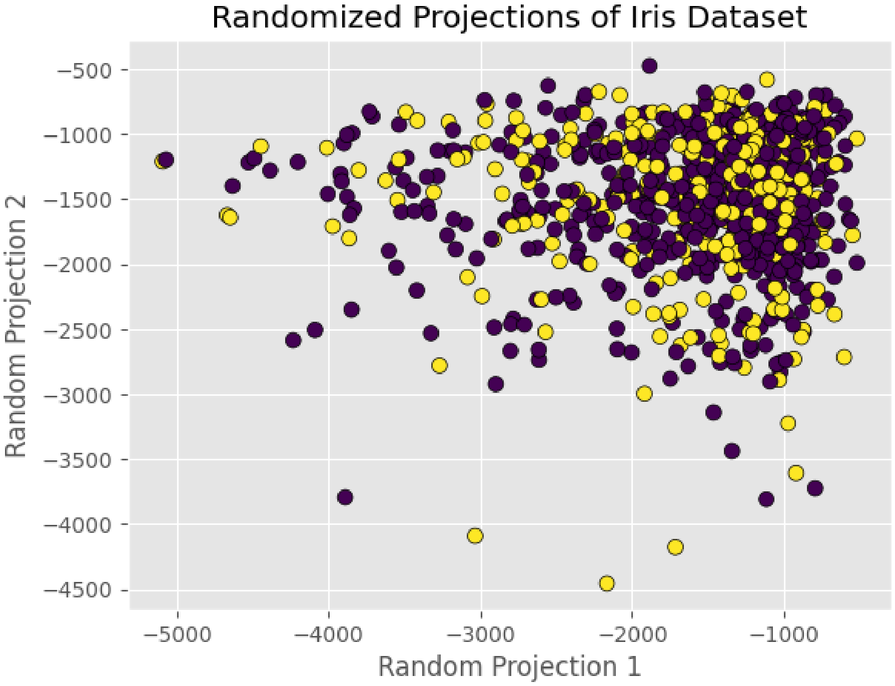

Figure 9: Randomized Projections of Aids Dataset

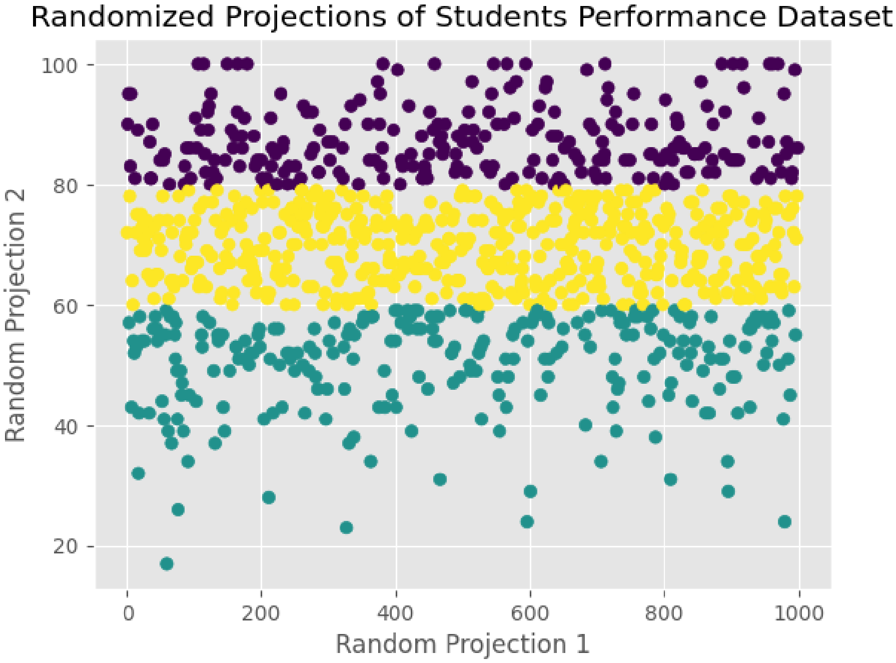

Figure 10: Randomized Projections of Students Performance Dataset

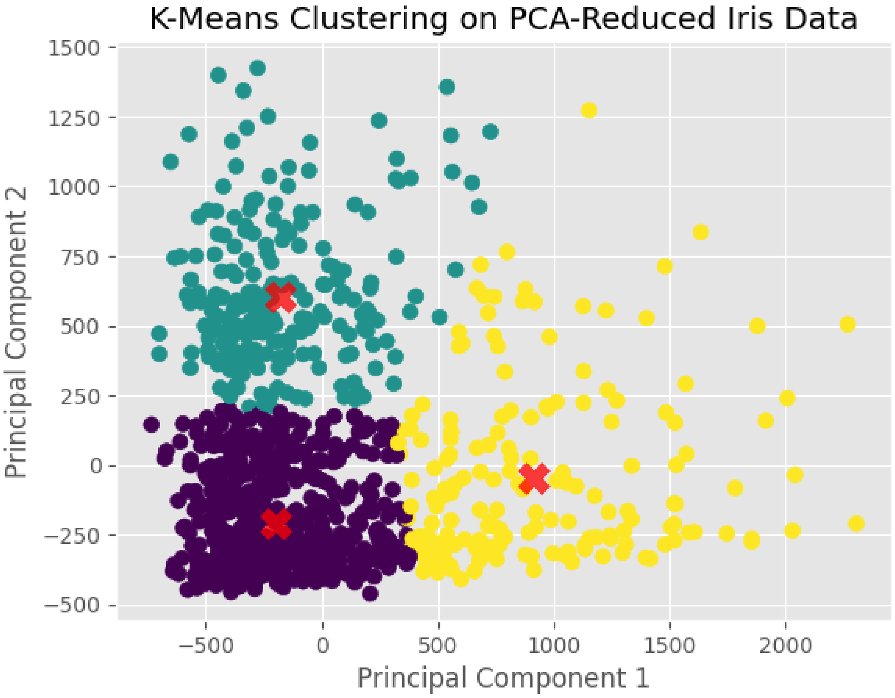

Figure 11: K-Means Clustering on PCA-Reduced Aids Data

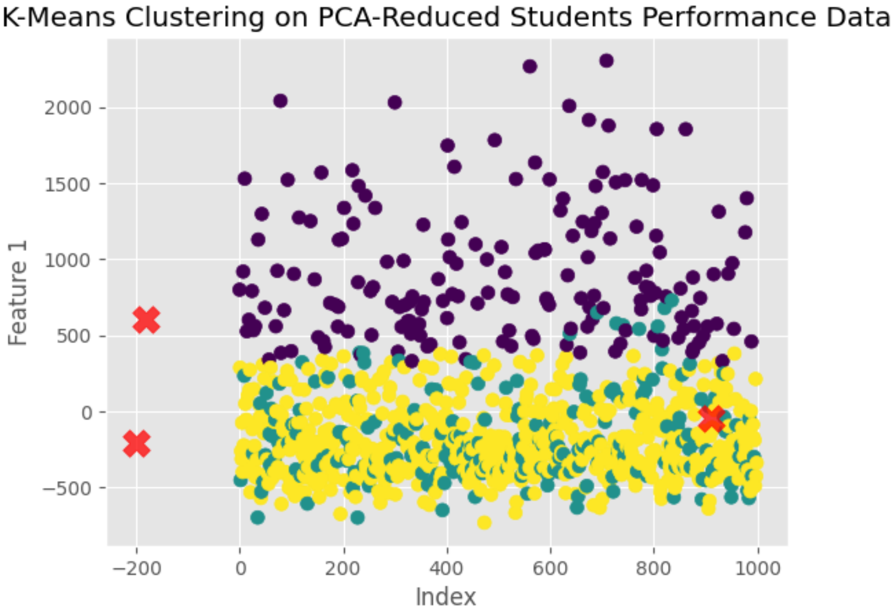

Figure 12: K-Means Clustering on PCA-Reduced Students Performance Data