By Kamilla Ten

Description of Implementation and selected Dataset

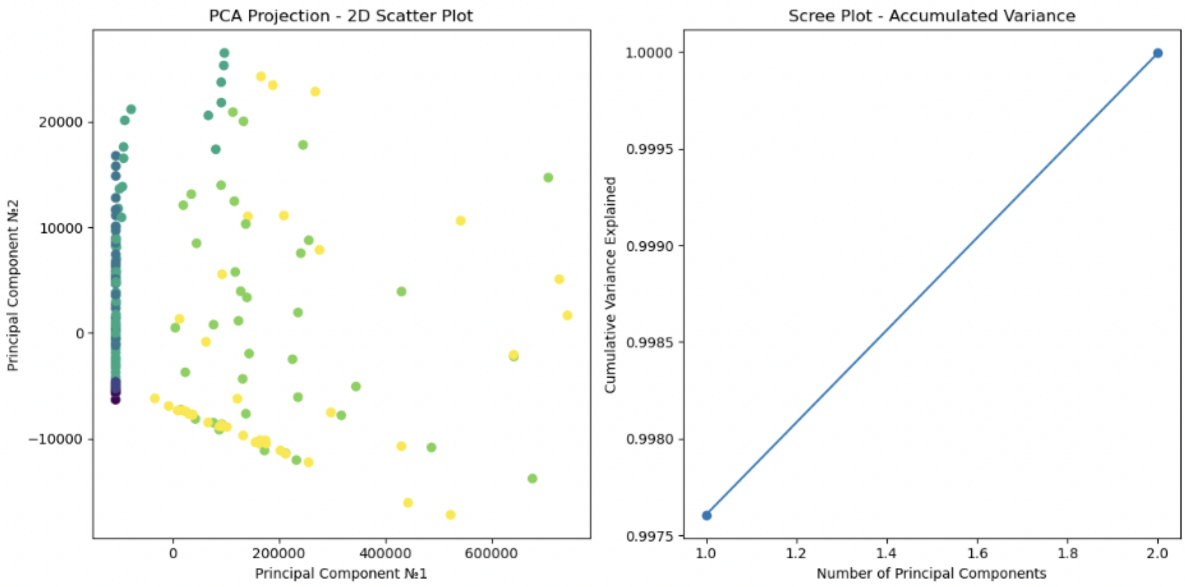
The analysis was conducted on two datasets. The first, a synthetic dataset named A3-data.txt, consists of four variables and one class across 360 patterns, where the class attribute was used solely for visualization purposes and not in the learning process. The second dataset features over six variables and a class attribute with at least four categories, encompassing more than 200 patterns. Data preprocessing steps included normalization and handling of missing values, preparing the dataset for the unsupervised learning techniques.

Link to the new dataset: https://www.kaggle.com/datasets/deepu1109/star-dataset

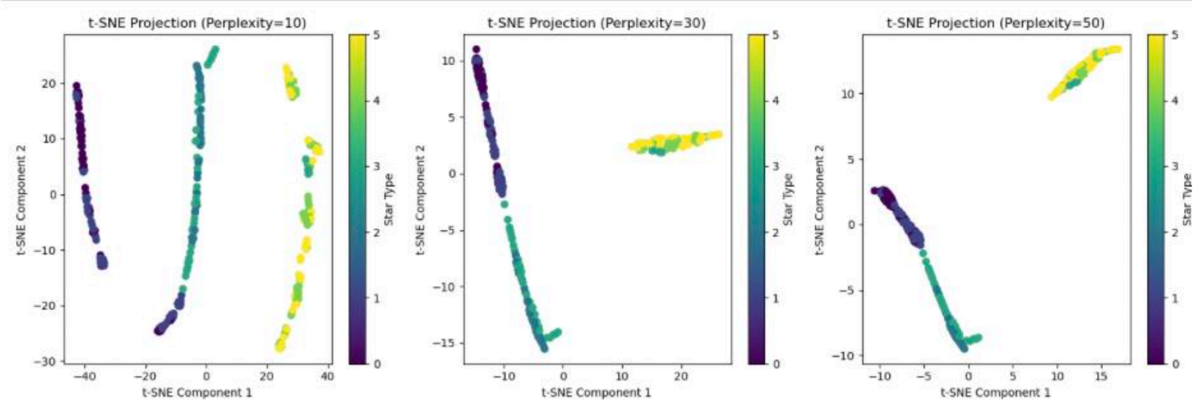
The Principal Component Method (PCA) is a powerful tool in the field of machine learning and statistics, used to reduce the dimensionality of data and highlight the most important characteristics, while preserving the maximum amount of information. PCA has a wide range of applications and is often used for data visualization, image compression, highlighting the most significant features, etc. The main purpose of PCA is to find new axes, called principal components, in the direction of maximum variation of the data. These new axes are constructed in such a way that the first main component explains the largest variance of the data, and the subsequent components explain the largest remaining variance, considering the orthogonality from the previous components. The PCA algorithm starts by calculating the covariance matrix or correlation matrix and further decomposing it into eigenvectors and eigenvalues. The main components are these eigenvectors, and the eigenvalues reflect the explained variance of the data in the appropriate directions.

Evaluation of the results

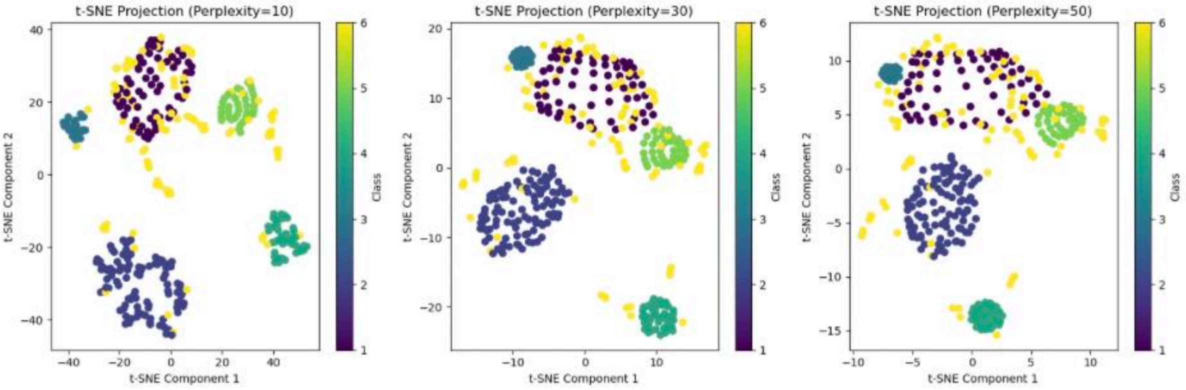
Our data analysis revealed distinctive clustering patterns when applying PCA and t-SNE to the synthetic and real-world datasets. For the synthetic dataset, PCA exposed clear groupings, allowing us to infer potential relationships between the different classes. The PCA plots indicate a variance in data distribution which could be pivotal for further classification tasks.



The t-SNE visualizations of the synthetic dataset, with varying perplexity values, showed that the choice of perplexity significantly impacts the data separation. Lower perplexity tended to generate more distinct clusters, which became less discernible with higher perplexity values. This insight is critical for understanding the local structure of the data and suggests that parameter tuning is essential for optimal cluster representation.

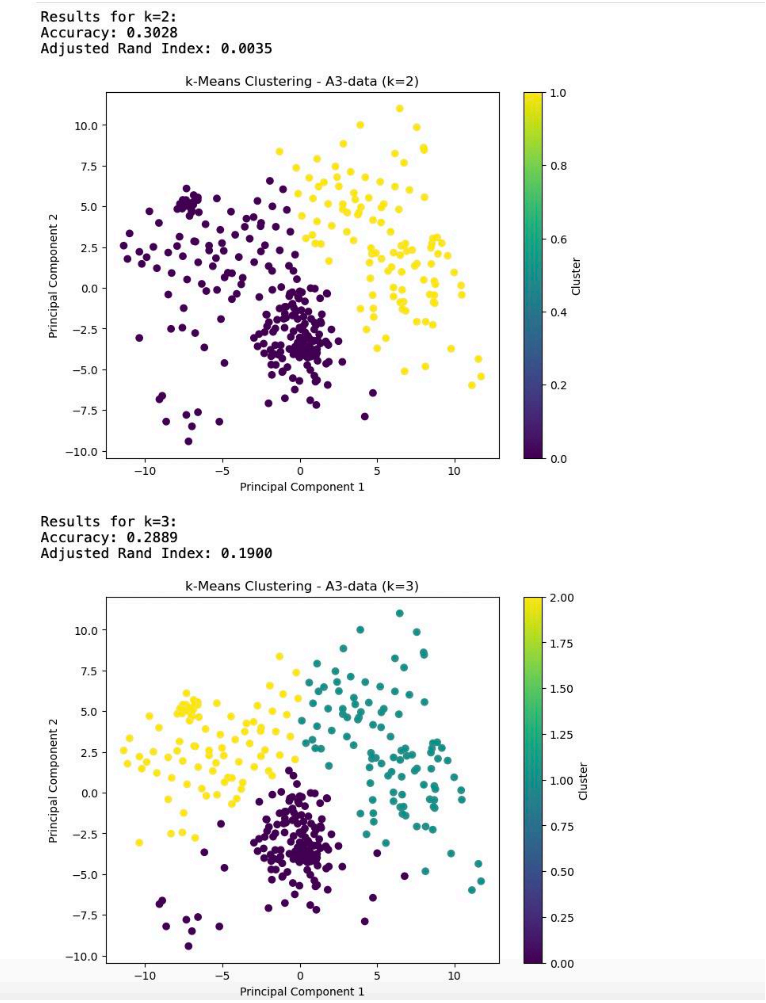


For the real-world dataset, t-SNE projections revealed a stark separation of star types. As perplexity increased, the clusters became more spread out, indicating that different star types have unique properties that can be distinguished in a reduced-dimensional space. This aligns with the known astrophysical characteristics that differentiate star types.



The application of k-Means clustering to the synthetic dataset, with various k values, highlighted the impact of cluster count on the classification accuracy and the adjusted Rand index

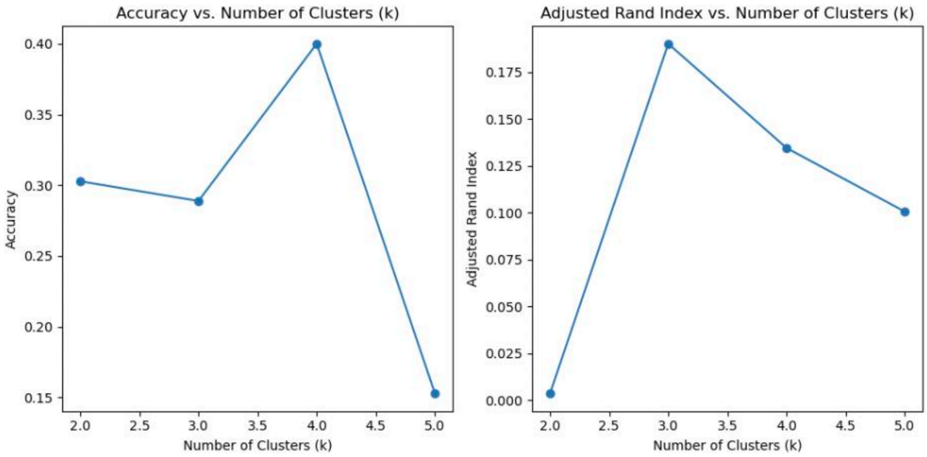
For k=2, we observed a moderate accuracy of 0.3028 and a low adjusted Rand index of 0.0035.



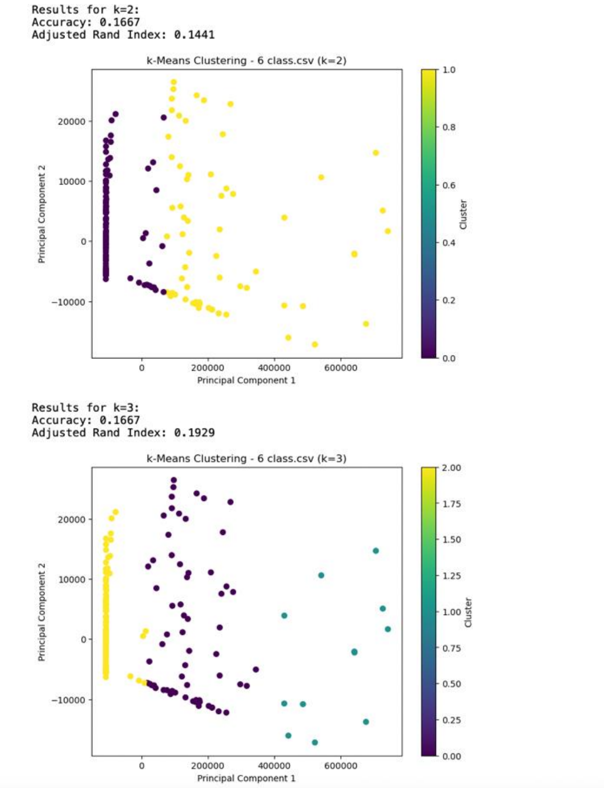
```
Results for k=2:
Accuracy: 0.3028
Adjusted Rand Index: 0.0035
```

```
Results for k=3:
Accuracy: 0.2889
Adjusted Rand Index: 0.1900
```

Increasing k to 3 improved the accuracy slightly to 0.2889, with an adjusted Rand index of 0.1900. The clusters became more defined, as seen in the scatter plot, but the confusion matrix still reflected substantial misclassification.

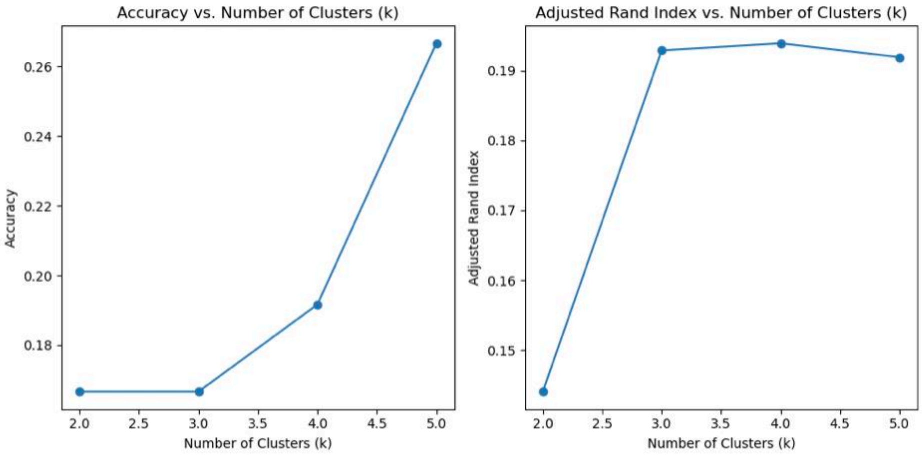


The application of k-Means clustering to the internet dataset, designed to categorize star types, demonstrates significant variability in performance metrics across different values of k. For k=2, we observe limited accuracy 0.1667. As we increment k to 3 and beyond, thereis no consistent improvement in accuracy 0.1667.

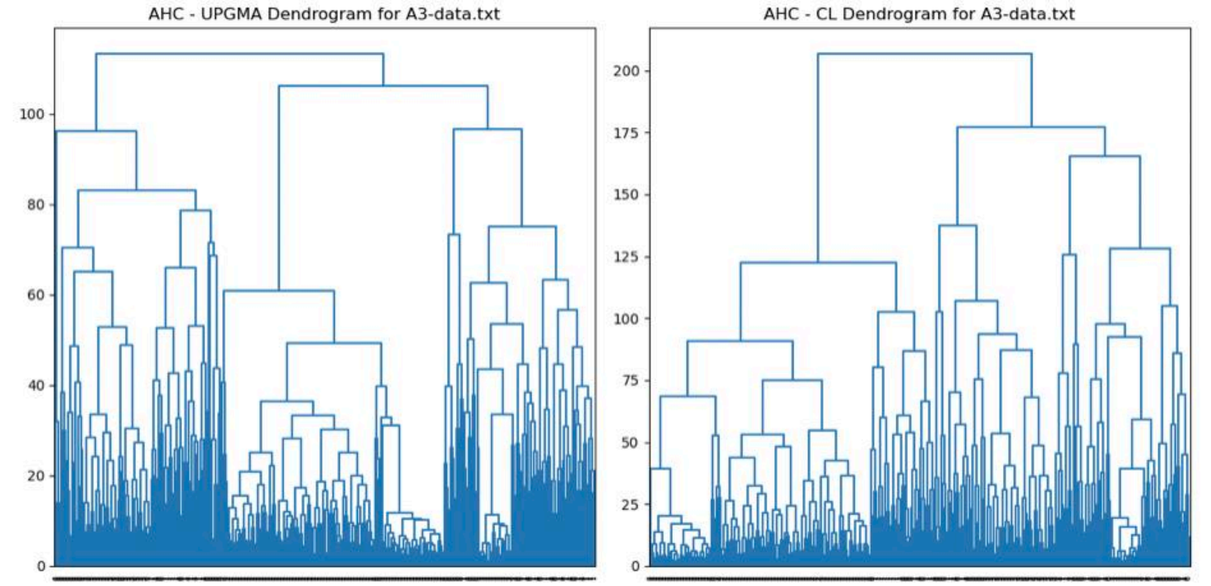


```
Results for k=2:
Accuracy: 0.1667
Adjusted Rand Index: 0.1441
```

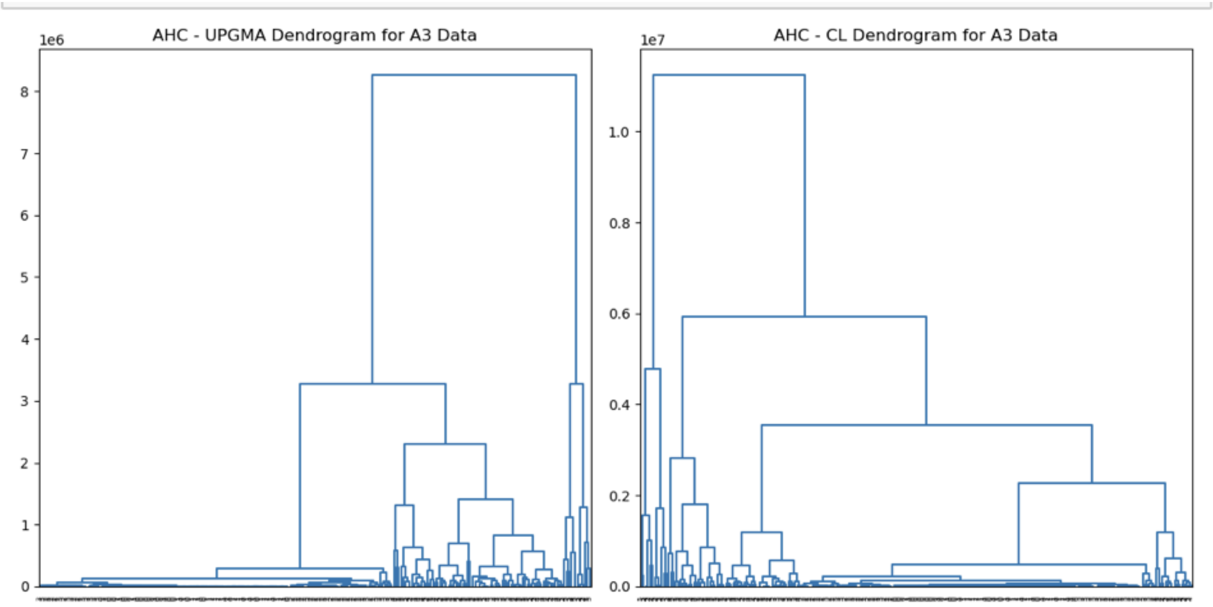
```
Results for k=3:
Accuracy: 0.1667
Adjusted Rand Index: 0.1929
```



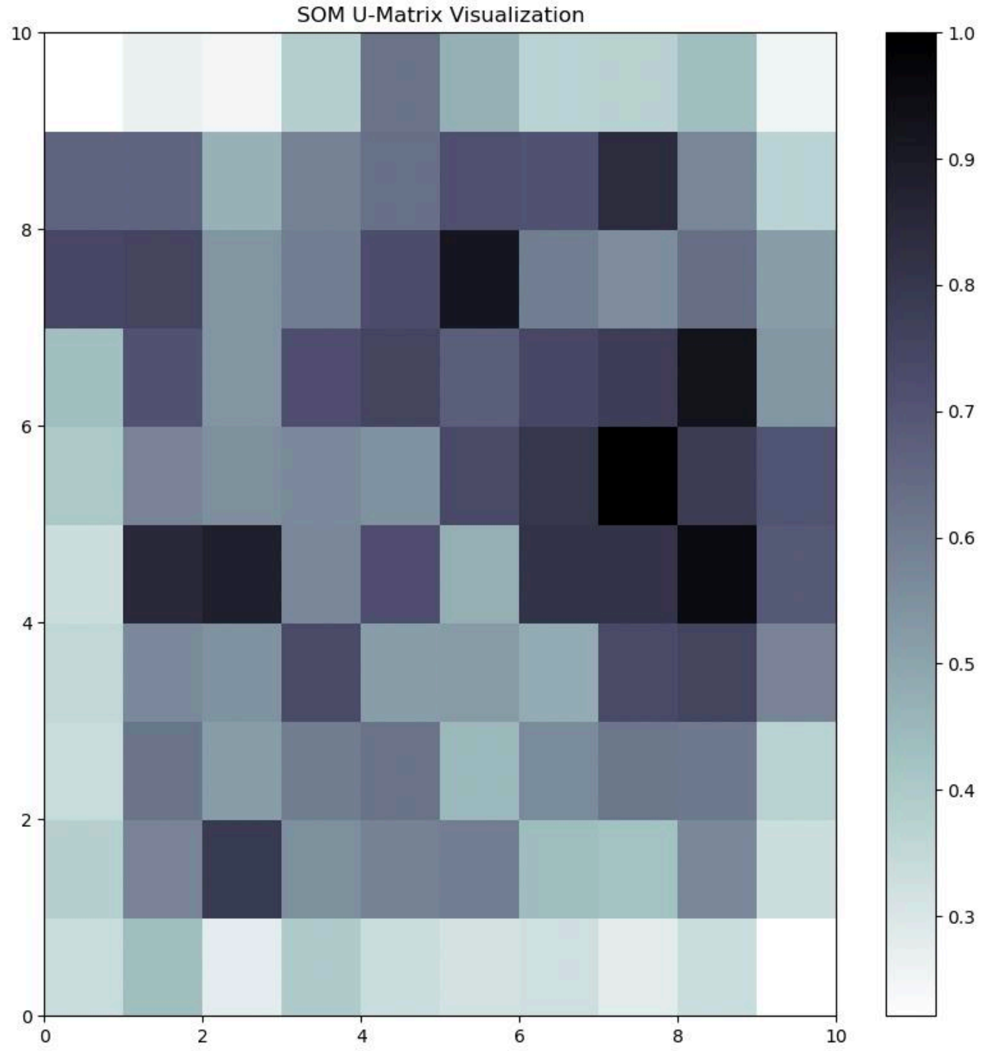
The dendrograms represent the agglomerative clustering process where each step merges the two nearest clusters.For the A3-data, the dendrogram suggests a complex structure with numerous clusters, as depicted by the extensive branching. This complexity indicates a diverse dataset where multiple levels of similarity exist among the data points.

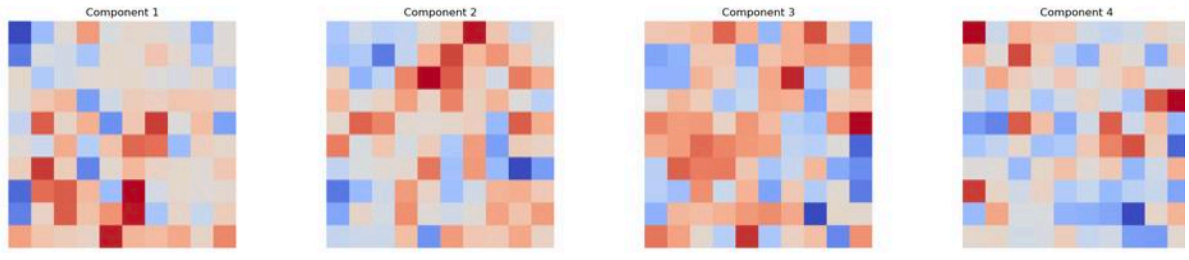


Conversely, the dendrogram for the 6-class csv dataset presents a clearer separation between clusters, especially at higher distances. This suggests more defined groupings within the data, potentially corresponding to different star types.

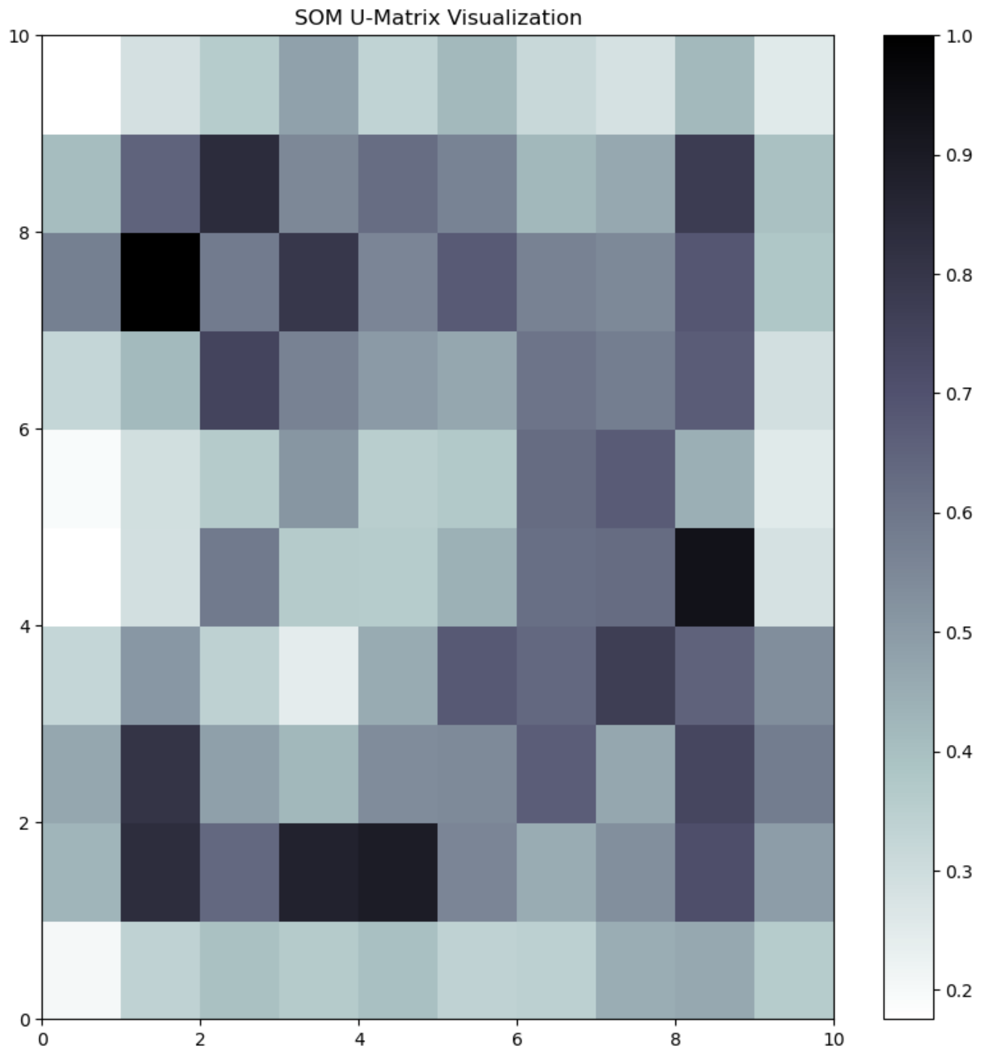


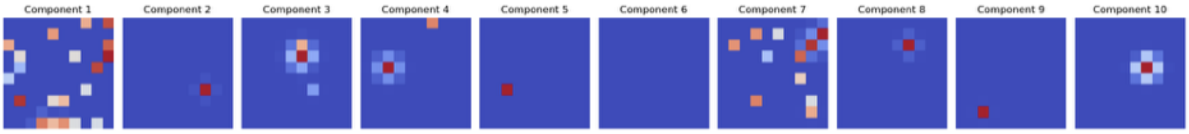
Based on the Self-Organizing Map (SOM) U-Matrix visualizations for both datasets, it's evident that the synthetic A3-data presents a more complex structure with several distinct high-distance regions, indicating varied data densities and potentially more clusters. On the other hand, the 6-class csv.csv dataset shows a clear, centralized high-distance region, suggesting a more concise cluster formation.





These SOM visualizations allow us to identify topological relationships within the datasets, with darker areas on the U-Matrix representing potential cluster borders. The contrast between the two datasets' visualizations could imply differences in data homogeneity and cluster separation.





Conclusion

The study validates the effectiveness of the chosen unsupervised learning techniques in extracting meaningful patterns and groupings from complex datasets. The application of PCA and t-SNE to both synthetic and real-world datasets has provided valuable insights into their structure. The observed clustering patterns confirm that unsupervised learning is a powerful tool for feature reduction and exploration of data. FCluster analysis via k-means provided a quantifiable means to ascertain the data's natural partitions. These methodologies, collectively, offer a robust framework for tackling unsupervised learning challenges in diverse datasets.