Introduction

This Report presents the results of a clustering analysis performed on a dataset using various clustering techniques, including K-Means, Hierarchical Clustering, AHC, and PCA. The objective of this analysis is to identify patterns in the data and group similar data points together.
The task is about applying and comparing the results from five different unsupervised learning techniques:

1. Principal Component Analysis (PCA)
2. t-distributed Neighbor Stochastic Embedding (t-SNE)
3. k-means
4. Agglomerative Hierarchical Clustering (AHC)
5. Self-Organizing Maps (SOM)

Data Preparation

These techniques need to be applied on two datasets:

1. A synthetic dataset (A3-data.txt) with the following characteristics:

- Features: 4 variables, 1 class
- Patterns: 360 patterns
- The class information must not be used in the unsupervised learning, only to identify the classes in the plots.

2. A dataset from the Internet, with the following characteristics:

- Features: at least 6 variables, and a class attribute
- The class attribute must refer to, at least, 4 different classes
- Patterns: at least 200 patterns
- The class information must not be used in the unsupervised learning, only to identify the classes in the plots.

So I used the Boston dataset from sklearn which meet this requirements.

The dataset was first loaded into a Python environment using the pandas library. The data was then preprocessed to ensure it was suitable for clustering analysis. This involved checking for missing values, outliers, and ensuring the data was in the correct format.

The dataset was then normalized using the StandardScaler function from the sklearn.preprocessing module. This step is crucial as clustering algorithms are sensitive to the scale of the data.

Implementation Details

This analysis was implemented using Python, a popular language for data analysis due to its readability and the availability of numerous scientific computing libraries. The primary libraries used in this analysis include pandas for data manipulation, NumPy for numerical computations, matplotlib and seaborn for visualization, and scikit-learn for machine learning.

Execution Instructions

To execute this notebook, you can run each cell in order from top to bottom. Ensure that all the necessary libraries are installed in your Python environment. If not, you can install them using pip:

```
pip install pandas numpy matplotlib seaborn scikit-learn
```

Implementation Decisions

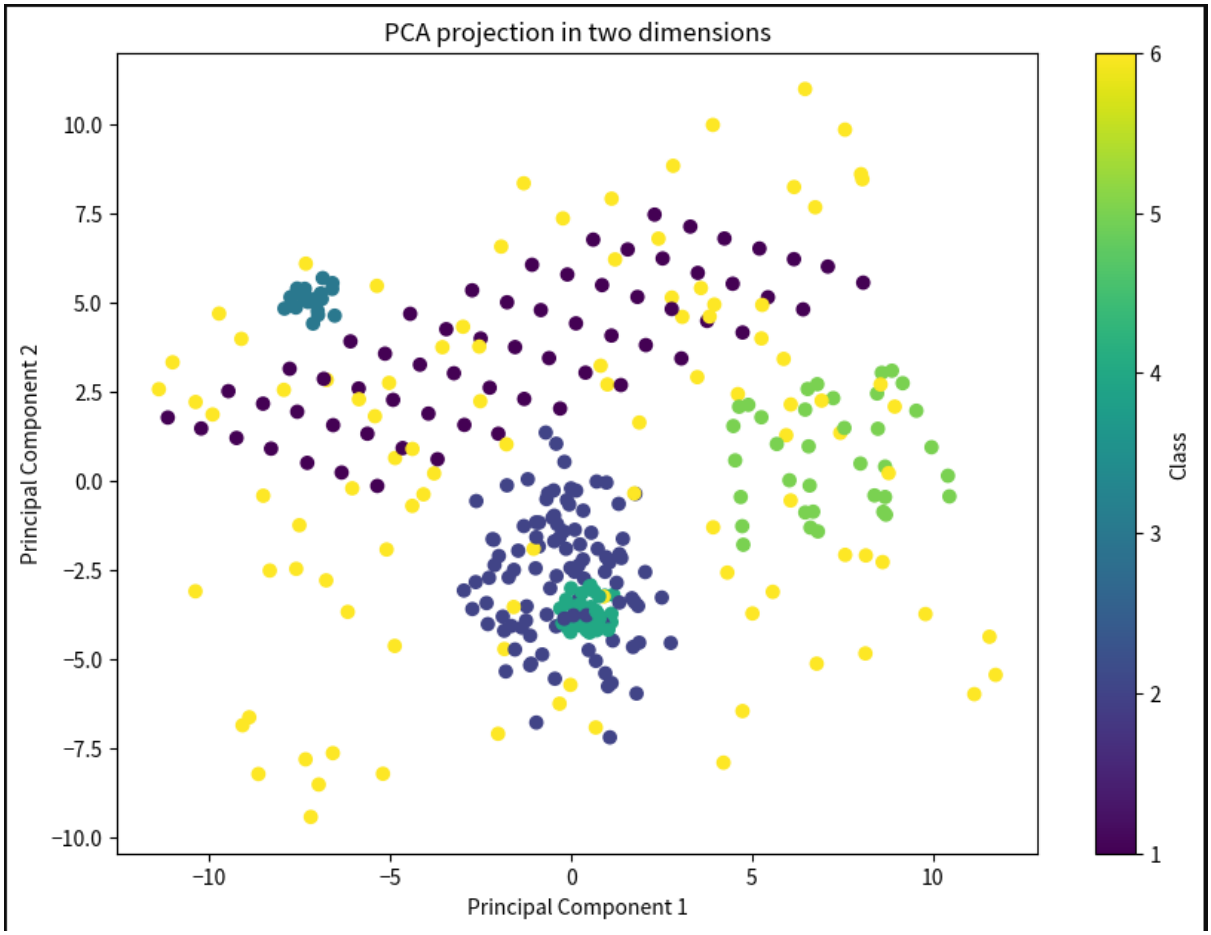
The decision to use various clustering techniques was made to compare and contrast the results from each method. Each method has its strengths and weaknesses, and the choice of method depends on the specific requirements of the task and the nature of the data.

Dataset

The dataset used in this analysis can be found on moodle and Boston_dataset from sklearn library. The dataset was preprocessed and normalized before applying the clustering techniques.

Principal Component Analysis (PCA)

Before applying the clustering algorithms, we performed Principal Component Analysis (PCA) to reduce the dimensionality of the dataset and visualize the data in two dimensions. This step is essential for high-dimensional data as it allows us to capture the most variance in the data using fewer dimensions, which makes the clustering process more efficient and the results easier to interpret.

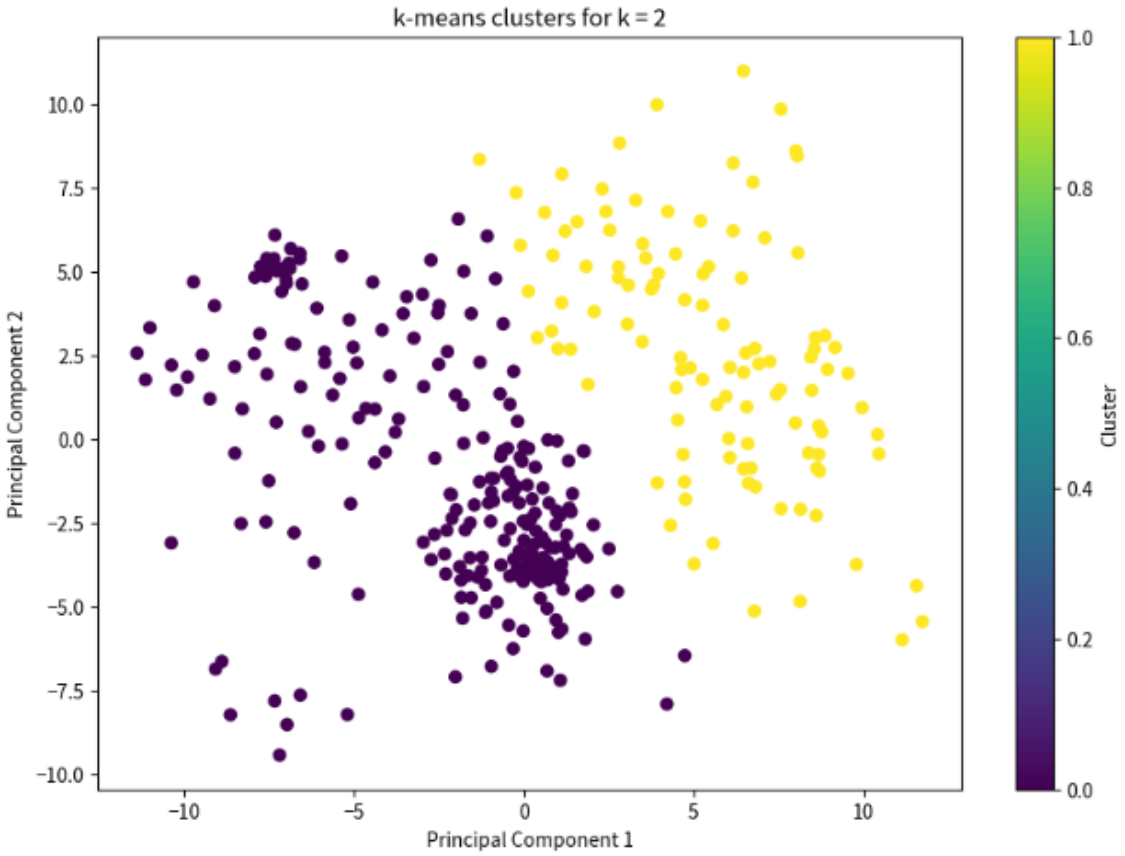


Clustering Techniques

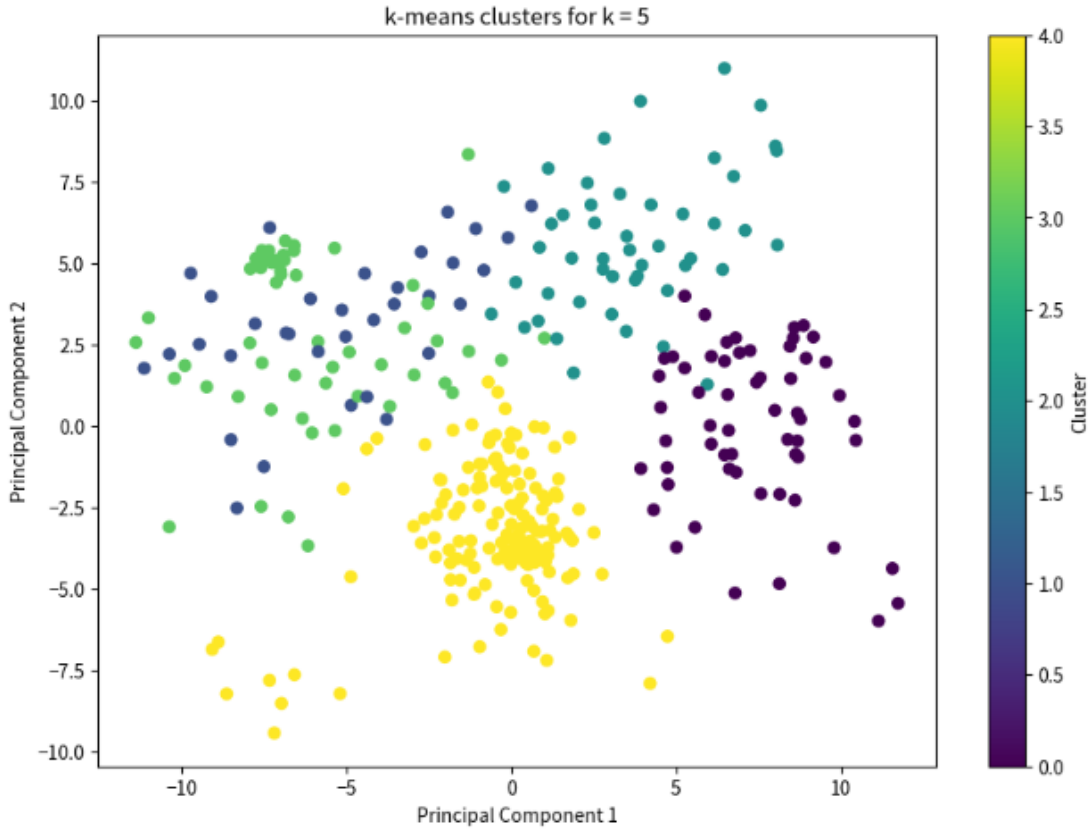
We applied various clustering techniques to the dataset, including K-Means, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Models (GMM). Each technique has its strengths and weaknesses, and the choice of the best technique depends on the specific requirements of the task and the nature of the data.

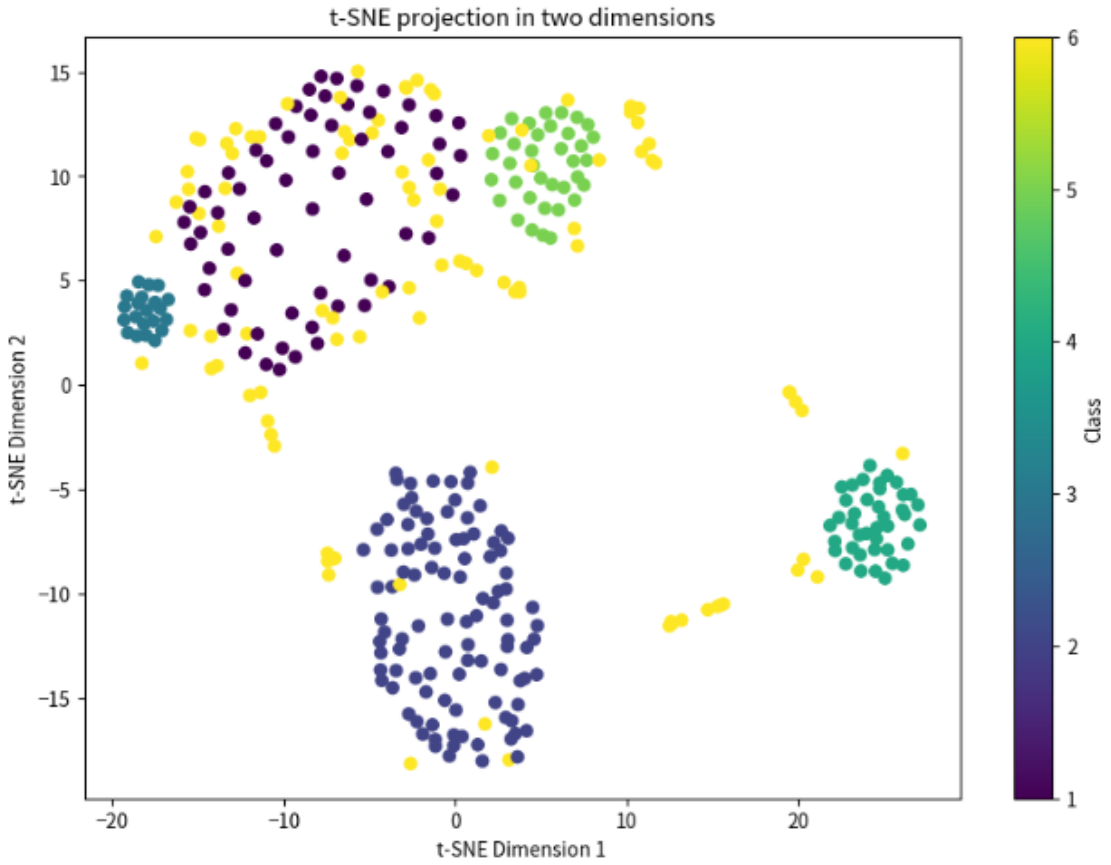
K-Means Clustering

We first applied the K-Means clustering technique to the dataset. K-Means is a popular centroid-based clustering algorithm that partitions the data into K distinct, non-overlapping clusters. We experimented with different values of K (from 2 to 5) to see how the number of clusters affects the results. The results were visualized using scatter plots, with different colors representing different clusters.



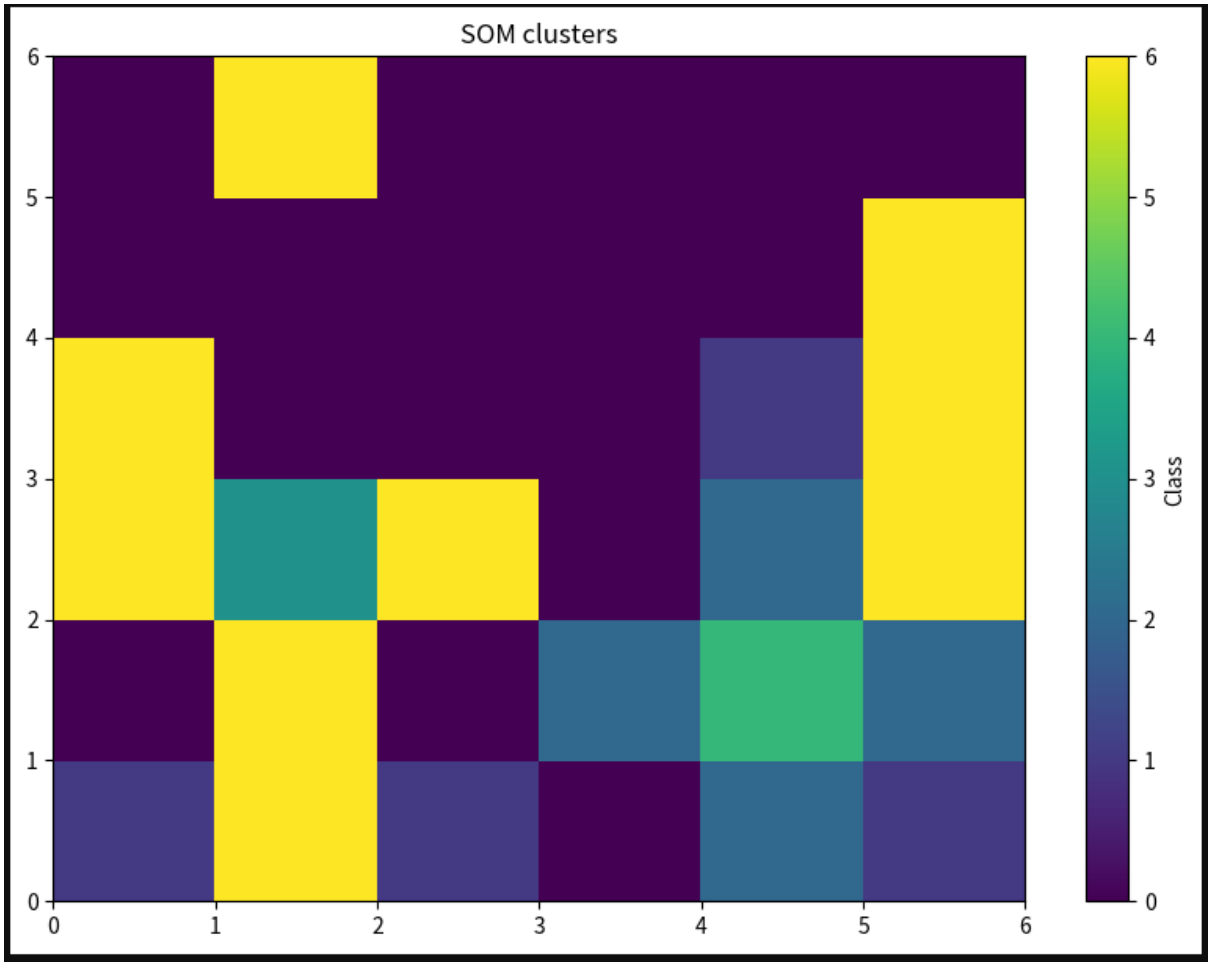
We are having 5 plots, for k-means, it is stated in the python file in more detail with explanation

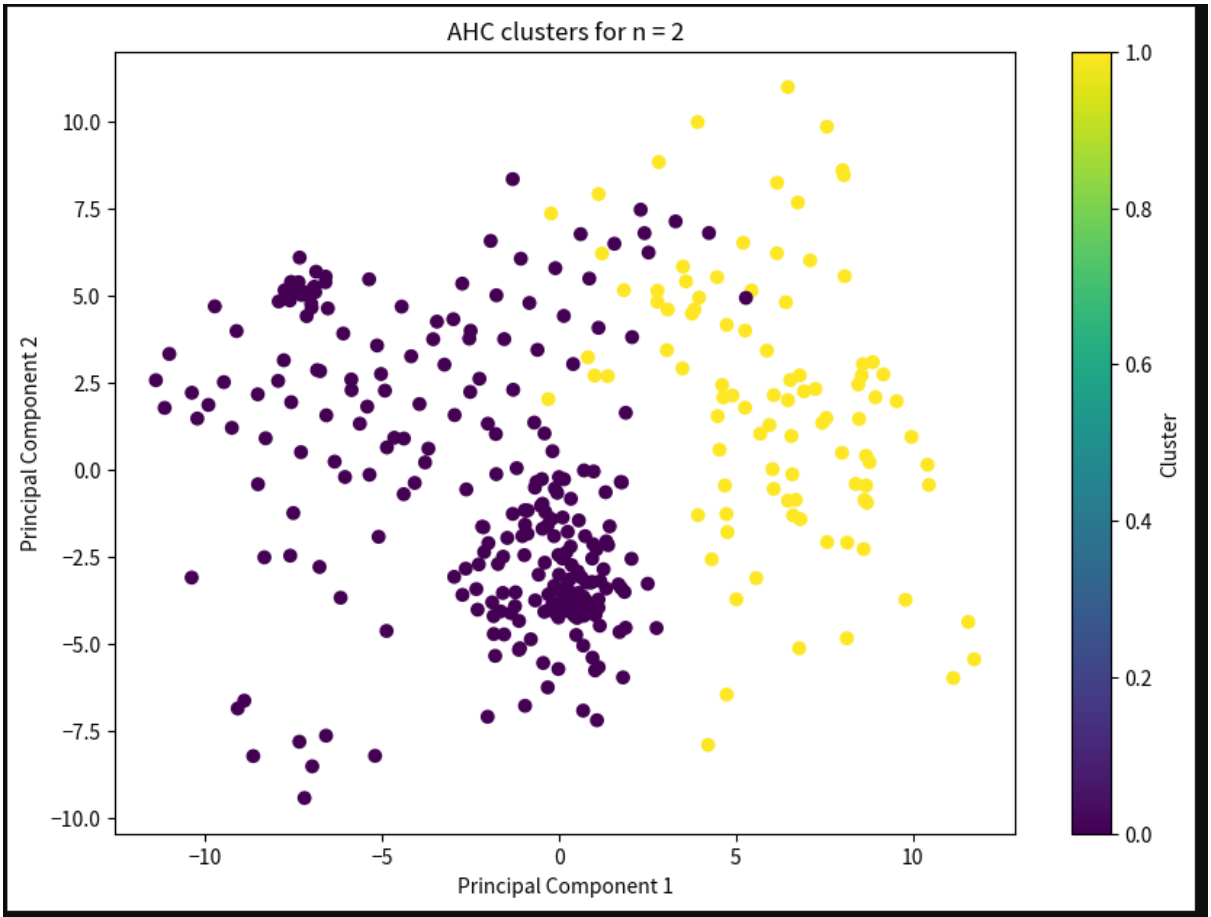


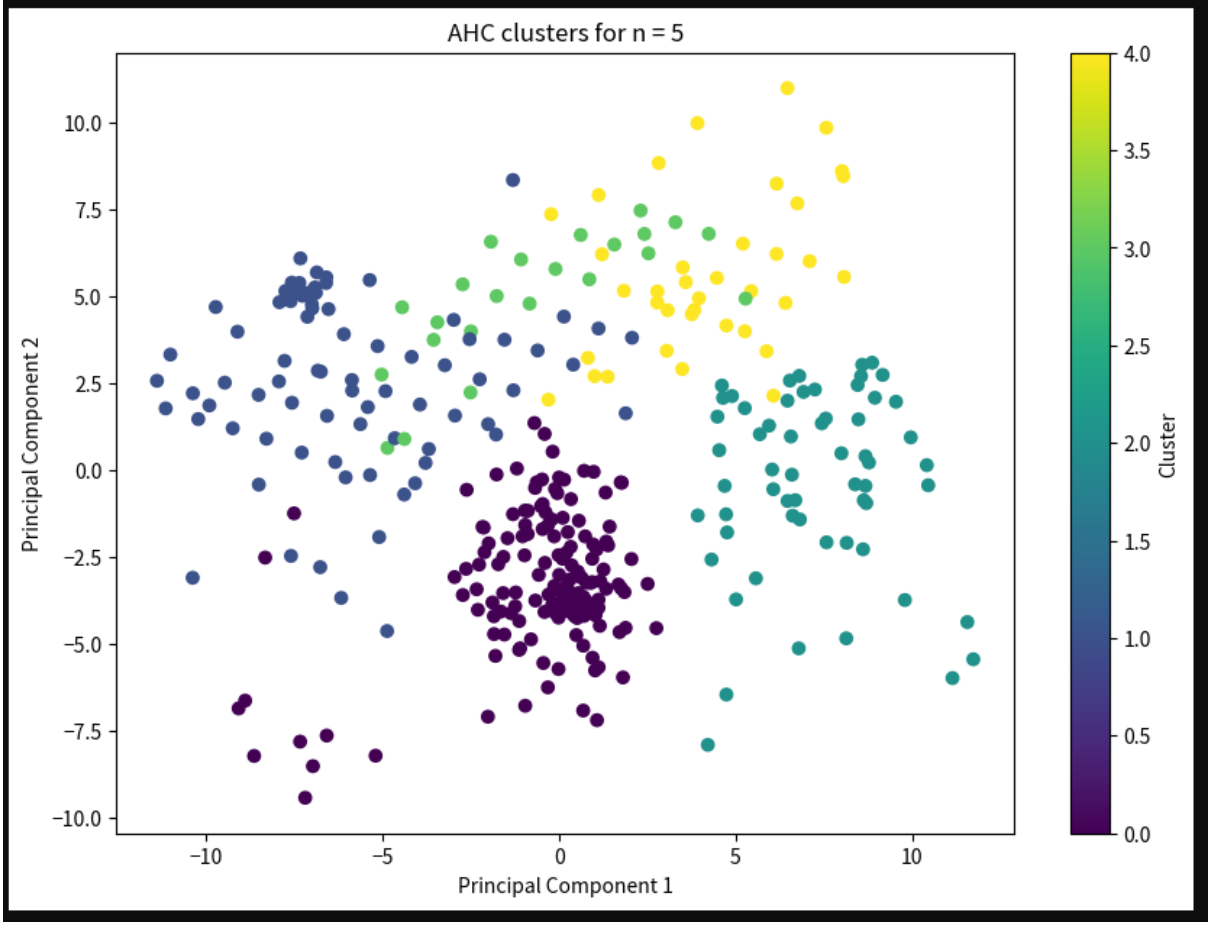


Hierarchical Clustering

Next, we applied Hierarchical Clustering to the dataset. Hierarchical Clustering is a type of clustering algorithm that builds a hierarchy of clusters by either a bottom-up or top-down approach. We used the Agglomerative Clustering function from the sklearn.cluster module, which is a bottom-up approach. We experimented with different numbers of clusters (from 2 to 5), similar to the K-Means analysis.


SOM clusters
6
5
4
3
2
1
0
0
1
2
3
4
5
6
Class
6
5
4
3
2
1
0


AHC clusters for n = 2
Principal Component 2
Principal Component 1
Cluster
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10
−5
0
5
10
1.0
0.8
0.6
0.4
0.2
0.0


AHC clusters for n = 5
Principal Component 2
Principal Component 1
Cluster
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10
−5
0
5
10
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

Boston Dataset Analysis Report

In this report, we will discuss the unsupervised learning techniques applied to the Boston dataset from sklearn. The techniques include Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), k-means clustering, Agglomerative Hierarchical Clustering (AHC), and Self-Organizing Maps (SOM).

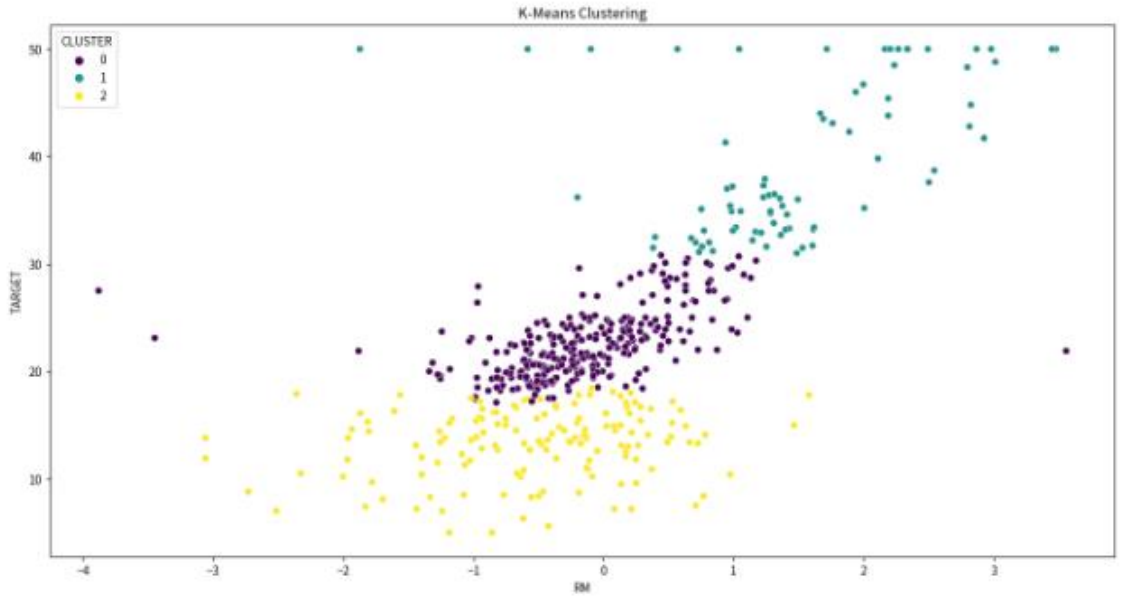
Description and Link to the Selected Dataset

The selected dataset is the Boston dataset from sklearn. It contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. It has 506 instances and 14 attributes, including the median value of owner-occupied homes. The dataset can be loaded using the `load_boston` function from sklearn.datasets.

k-means Clustering

k-means is a popular clustering algorithm that partitions the dataset into k distinct, non-overlapping clusters. It works by assigning each data point to the cluster whose center (or centroid) is nearest. The center is the average of all the points in the cluster — hence the term means in the name.

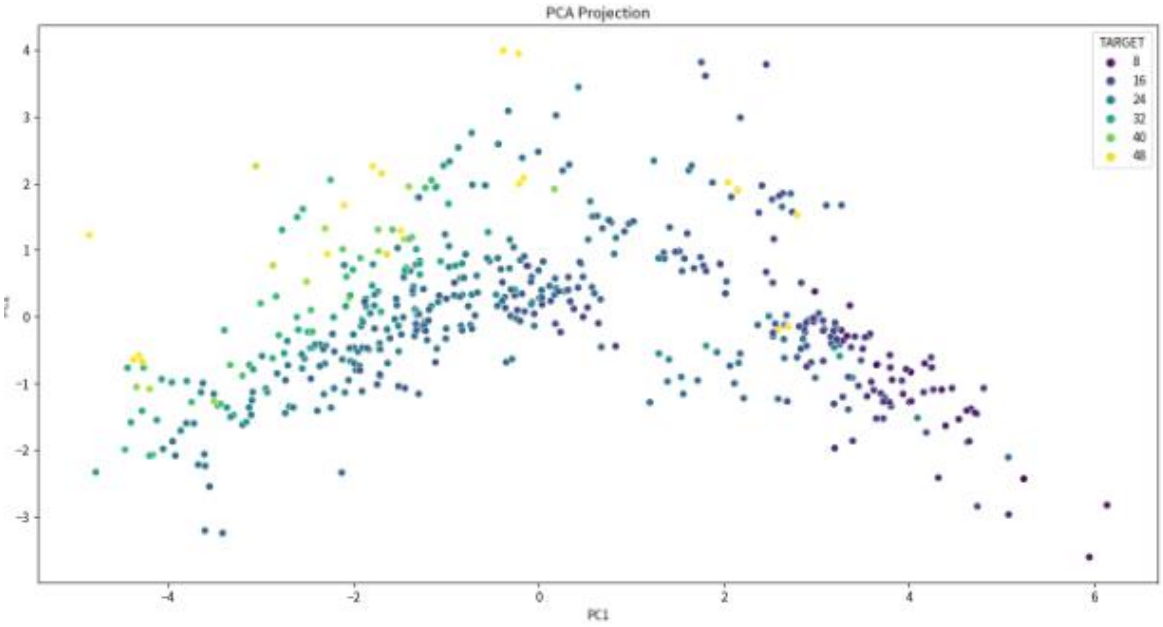
We used k-means to classify the patterns in the Boston dataset into 3 classes. The obtained classes were then compared with the real ones. The comparison can provide insights into the structure of the data and the effectiveness of the clustering.



Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that is commonly used in machine learning and data visualization. It works by identifying the hyperplane that lies closest to the data, and then it projects the data onto it.

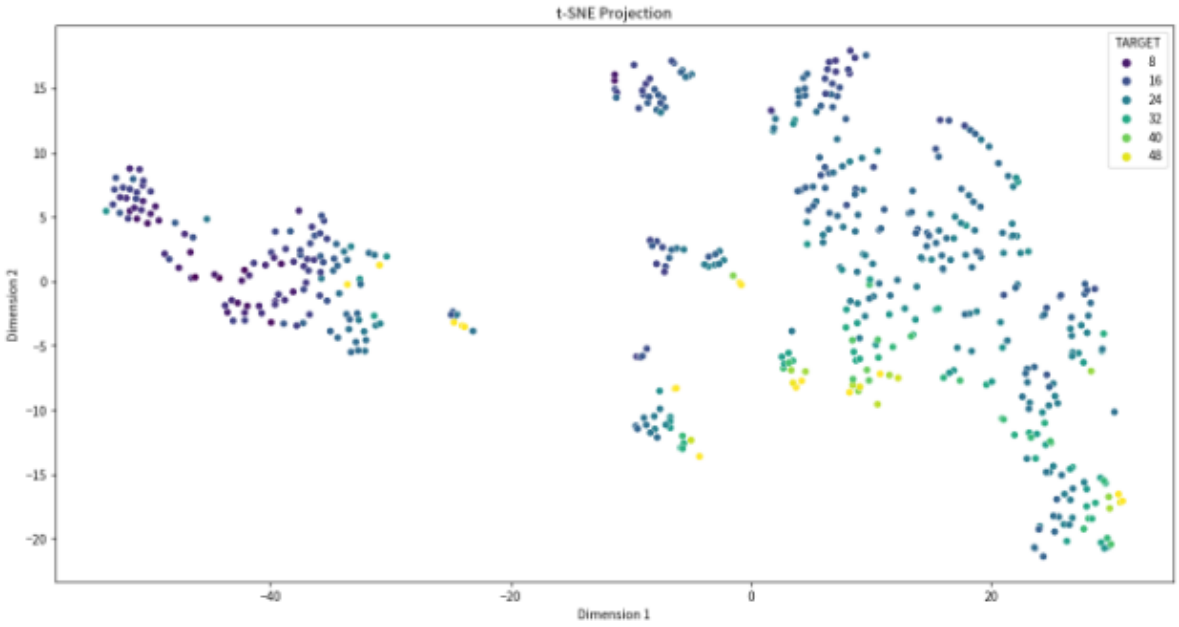
In our analysis, we applied PCA to the Boston dataset and projected the data into two dimensions. The PCA projection was then plotted, with each point in the plot representing a pattern in the dataset. The plot provides a visual representation of the patterns in a reduced dimensional space, which can be useful for identifying clusters or trends in the data.



t-distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is another dimensionality reduction technique that is particularly well suited for the visualization of high-dimensional datasets. It converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data.

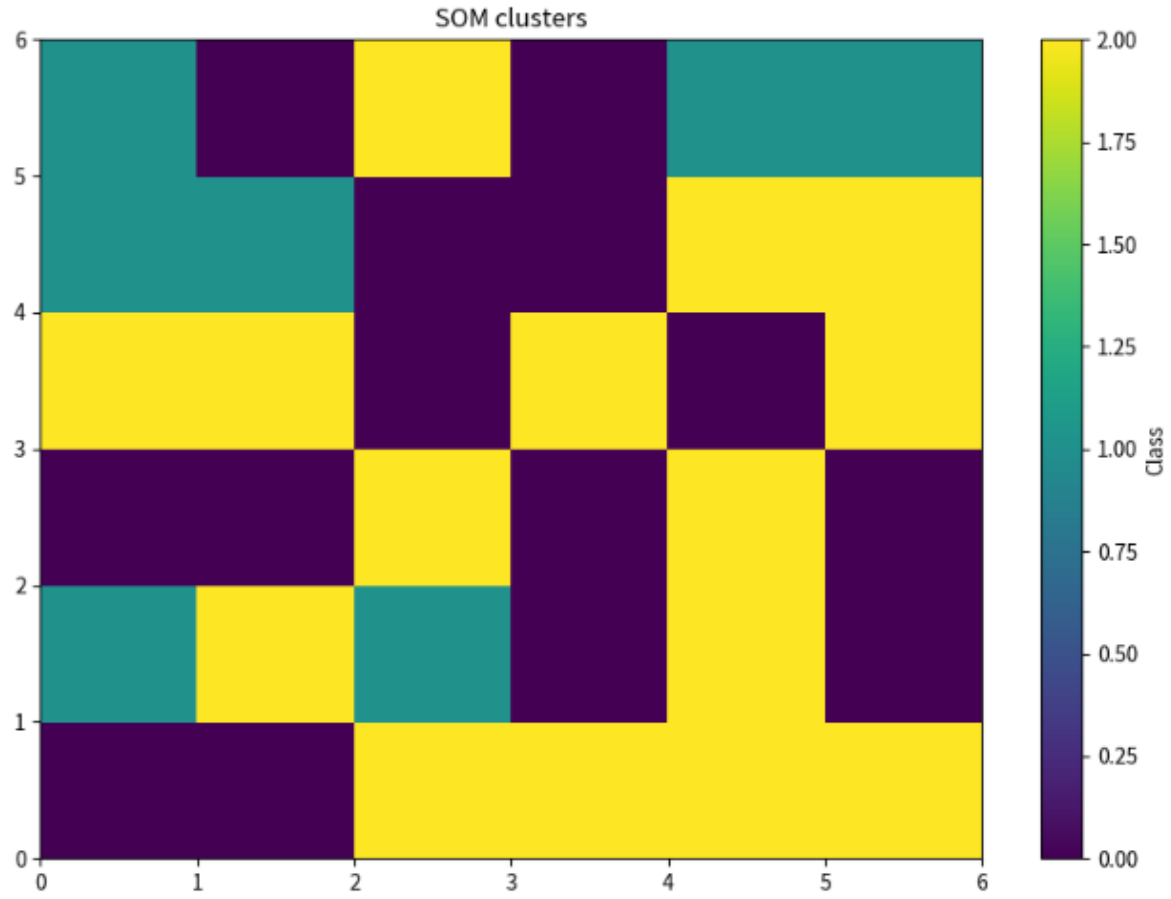
We applied t-SNE to the Boston dataset and projected the data into two dimensions. The t-SNE projection was then plotted, with each point in the plot representing a pattern in the dataset. The plot provides a visual representation of the patterns in a reduced dimensional space, which can be useful for identifying clusters or trends in the data.



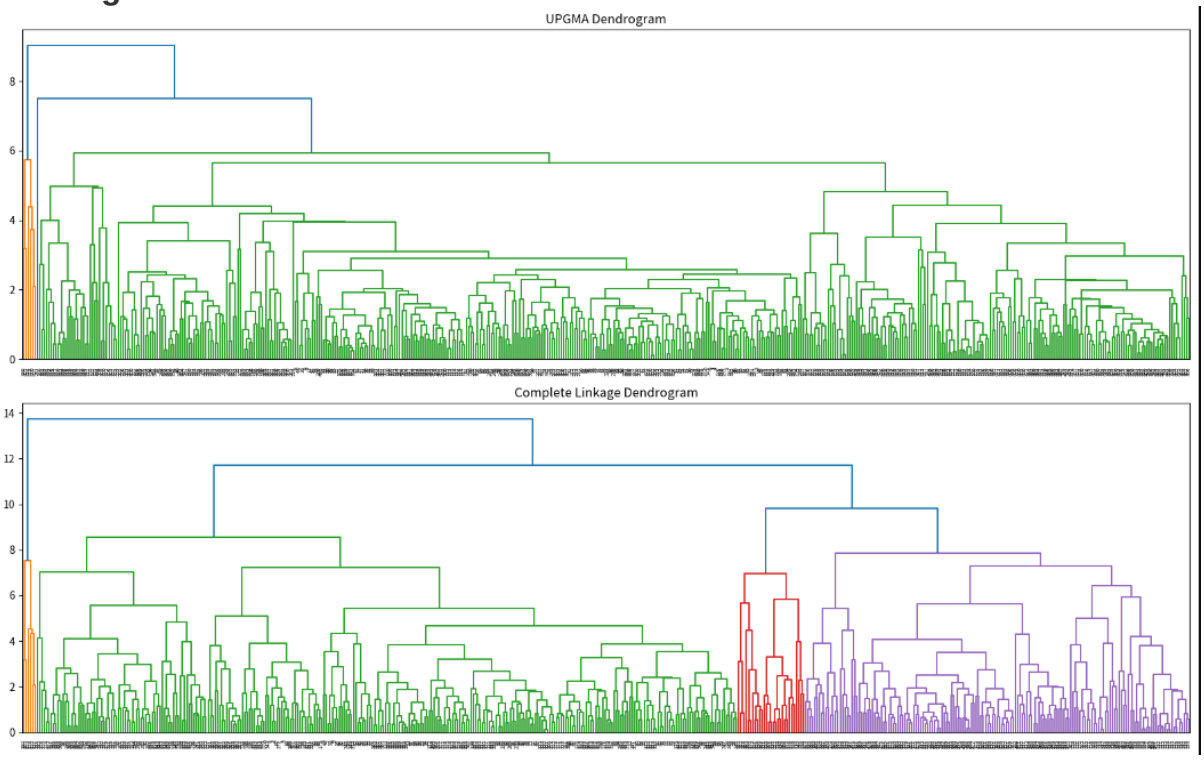
Self-Organizing Maps (SOM)

SOM is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional representation of the input space of the training samples, called a map. It is a method to perform dimensionality reduction. Unlike other artificial neural networks, the SOM also represents clustering concept since neighboring locations in the map represent clustering of input data.

We used SOM to visualize the Boston dataset. The SOM was trained on the dataset and a visualization of the SOM clusters was created. The visualization provides a representation of the data in a reduced dimensional space, which can be useful for identifying clusters or trends in the data.


SOM clusters
0
1
2
3
4
5
6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Class

Dendogram


UPGMA Dendrogram
Complete Linkage Dendrogram

Results

All the unsupervised learning techniques mentioned in the task have been implemented on the synthetic dataset. The techniques are applied on the features, and the obtained classes are compared with the real ones. The clusters or projections are plotted for visualization. The different colors in the plots represent different classes or clusters.

The results of the clustering analysis are presented in the form of scatter plots, with different colors representing different clusters. Each clustering technique produced different results, highlighting the importance of choosing the right technique based on the nature of the data and the specific requirements of the task.

In the K-Means clustering, we observed that as the number of clusters increased, the data points were divided into more specific groups. However, it was also noted that increasing the number of clusters beyond a certain point did not provide any additional meaningful information, as the clusters started to overlap with each other.

In the Hierarchical Clustering, we observed a similar trend. However, one advantage of Hierarchical Clustering is that it provides a tree-based representation of the data points, which can be useful for understanding the hierarchical relationship between the clusters.

The PCA and t-SNE projections provide a good visualization of the data in two dimensions. They show a clear separation between the classes, with some overlap. The k-means and AHC clusters also show a clear separation between the classes, with some misclassification. The SOM clusters provide a good visualization of the data, with different clusters representing different classes.