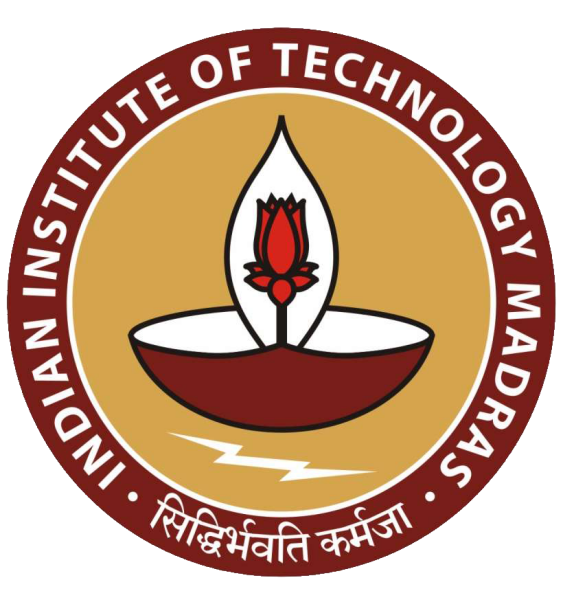
ID5002W: Industrial AI Laboratory

Assignment Report



ASSIGNMENT - 6

Dimensionality Reduction

SOUMYA MUKHERJEE

CH24M571

### 1.0 Preface

We have been given a single cell RNA Sequencing dataset with expression

values for different genes in different cells. The aim is to find the SVM classifier model that fits the data . In the second question , we have the objective to predict the survival rate based on the tumor data provided .

### 2.0 Data Pre-processing

We investigated the dataset for null and missing values . The dataset contains 25010 columns of features and 2 categorical information . This report summarizes the analysis performed on a gene expression dataset using various dimensionality reduction techniques. The primary goals were to:

- Preprocess and filter the data for analysis.

- Apply Principal Component Analysis (PCA) to reduce dimensionality and identify principal components.

- Explore Non-negative Matrix Factorization (NMF) to identify latent features.

- Visualize the data using t-distributed Stochastic Neighbor Embedding (t-SNE).

- Visualize the data using Uniform Manifold Approximation and Projection (UMAP).

Data Preprocessing

The initial dataset contained missing values and categorical features. The following preprocessing steps were performed:

- Removal of the 'Published\_Sample\_Name' column.

- Genes with the lowest 10% variance were removed to reduce noise.

- The remaining data was scaled using StandardScaler and MinMaxScaler for PCA and NMF, respectively.MinMaxScaler was used since NMF cannot use negative values , which are produced by StandardScalar .

**Principal Component Analysis (PCA)**

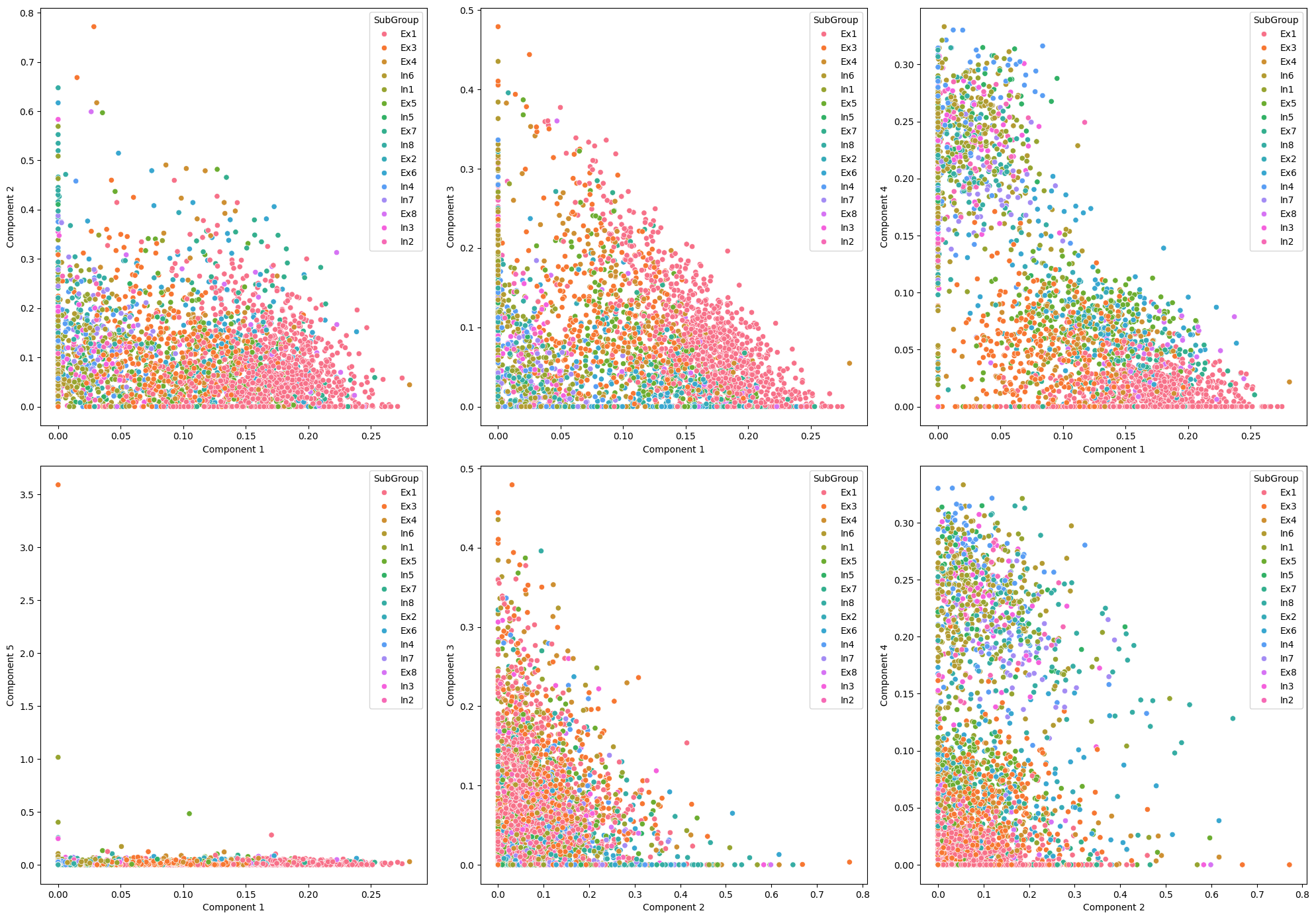
PCA was applied to the standardized data, reducing the dimensionality to 100 principal components. The analysis showed that the top 100 components captured a significant portion of the variance in the dataset.

The scatter plot of the first two principal components colored by ‘SubGroup’ and ‘BA’ indicates some degree of separation between different subgroups. However, there is also overlap, suggesting that these two components alone may not be sufficient to completely distinguish all subgroups. There might be a few outlier samples that are distant from the main clusters in the PCA plot .



**Non-negative Matrix Factorization (NMF)**

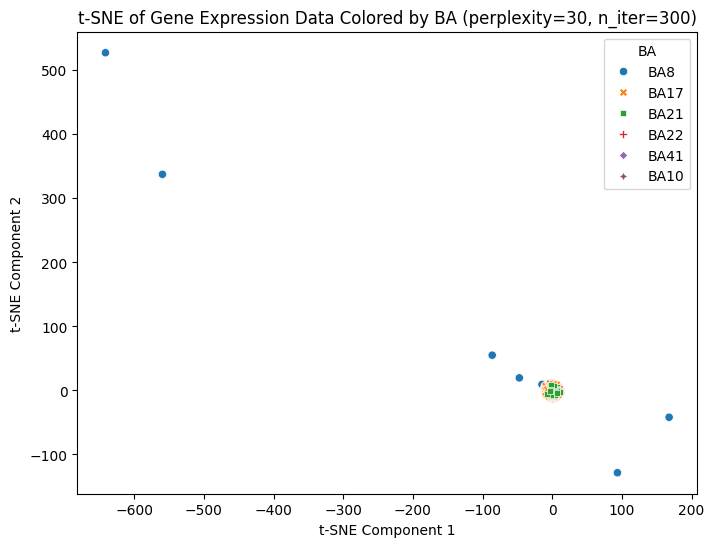
NMF was applied to the data scaled using MinMaxScaler. Five components were extracted, and scatter plots of component pairs were generated. These plots suggested potential clusters, and the top contributing features for each component were identified.

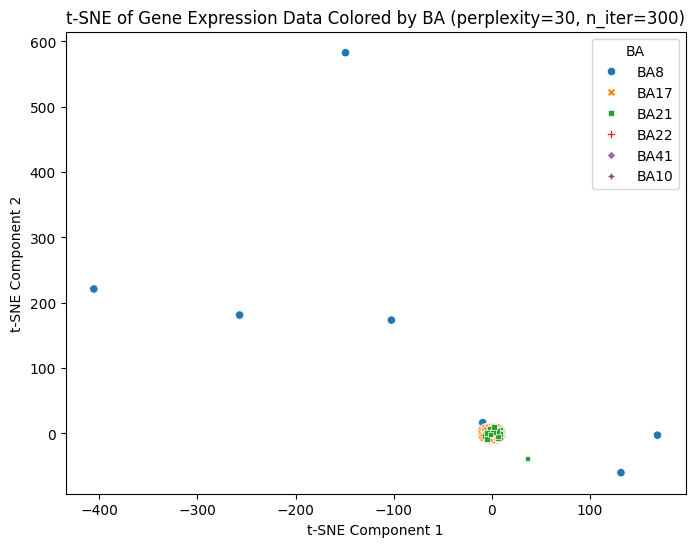


The scatter plots of NMF components against each other reveal patterns in the data that may not be captured by PCA. Some component combinations might show clearer distinctions between subgroups than others, suggesting specific components are more informative for certain subgroup comparisons.

**t-distributed Stochastic Neighbor Embedding (t-SNE)**

t-SNE was applied with different perplexity values and random states to visualize the data in two dimensions. The resulting scatter plots showed varying degrees of clustering, with the data points colored by the 'BA' variable. Some clear separation was observed, implying the presence of distinct groups.





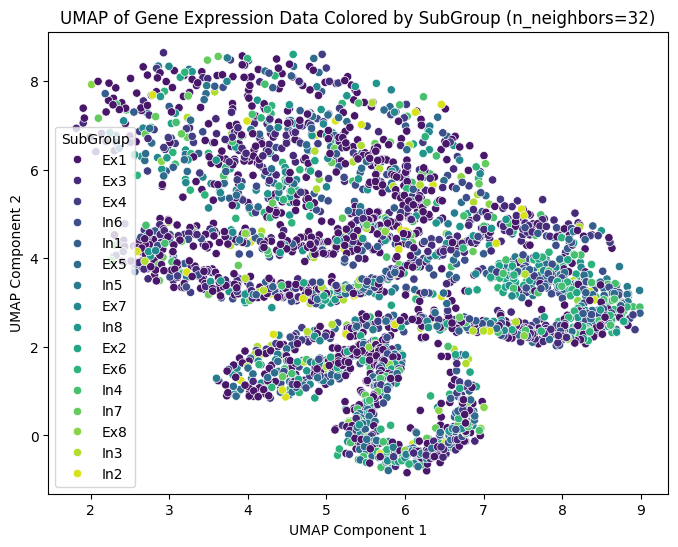
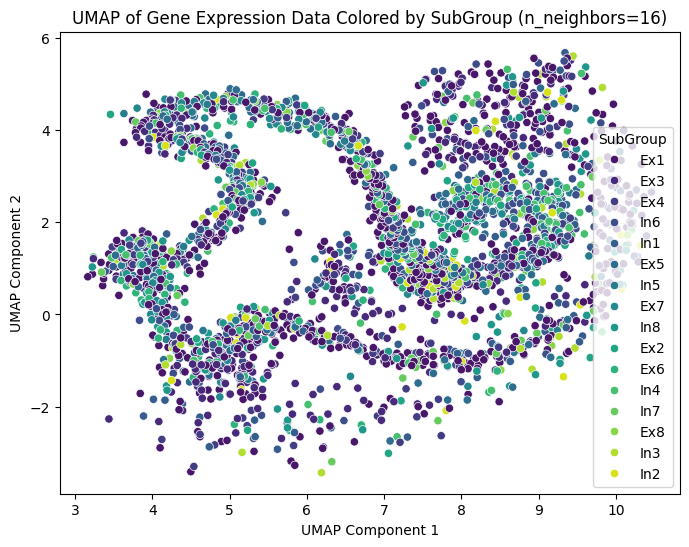
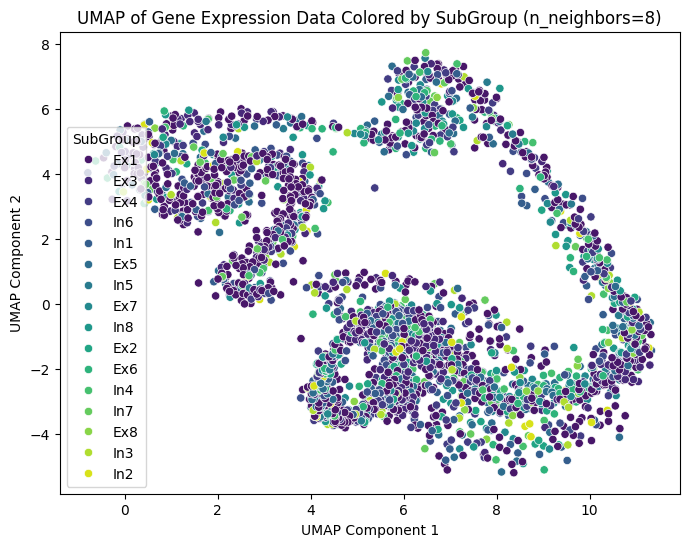
Cluster Visualization: t-SNE aims to preserve local neighborhood structures, making it effective for visualizing clusters in the data. The scatter plots colored by 'BA' can reveal how samples with different BA levels cluster together.

Impact of Random State: The plots generated using different random states might show variations in the exact arrangement of clusters. This highlights the sensitivity of t-SNE to initialization and emphasizes the importance of running it with multiple random states to ensure robustness of the results.

Potential for Subgroup Exploration: t-SNE visualizations can help explore potential subgroups within the data based on their proximity in the lower-dimensional space.

**Uniform Manifold Approximation and Projection (UMAP)**

UMAP was applied with different numbers of neighbors to visualize the data. The plots colored by 'SubGroup' demonstrated clusters, similar to t-SNE. The influence of the 'n\_neighbors' parameter on the visualization was also explored.



Cluster Visualization: Similar to t-SNE, UMAP aims to preserve local and global structures, making it suitable for visualizing clusters in the data. The scatter plots colored by 'SubGroup' can reveal how samples with different subgroups cluster together.

Impact of n\_neighbors: The plots generated using different 'n\_neighbors' values can show how this parameter affects the visualization. Larger 'n\_neighbors' values tend to emphasize global structure, while smaller values focus more on local structures.

Potential for Subgroup Exploration: UMAP visualizations can help explore potential subgroups within the data based on their proximity in the lower-dimensional space.

**Conclusions**

The analysis revealed that:

- Dimensionality reduction techniques such as PCA, NMF, t-SNE, and UMAP were effective in reducing the complexity of the gene expression data and revealed underlying patterns.

- PCA showed that the dataset could be represented by a smaller number of principal components while preserving a significant amount of variance.

- NMF identified latent features and their contributing genes.

- t-SNE and UMAP visualizations revealed clustering and relationships between samples based on 'SubGroup' and 'BA' variables.