t-SNE and UMAP Dimensionality Reduction Analysis Report

# Introduction

High-dimensional datasets are often difficult to visualize and interpret directly. Dimensionality reduction techniques such as t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) have emerged as powerful tools to project high-dimensional data into lower-dimensional spaces (typically 2D or 3D) while preserving meaningful structure.  
  
This report presents a comparative analysis of t-SNE and UMAP applied to a dataset, highlighting how each technique performs in terms of clustering quality, preservation of local/global structures, and interpretability.

The primary objectives of this analysis are:  
1. Data Preparation & Analysis – Scaling, preprocessing, and preparing the dataset for dimensionality reduction.  
2. Dimensionality Reduction – Applying t-SNE and UMAP to visualize cluster structures.  
3. Method Comparison – Analyzing strengths, limitations, and practical use cases of both methods.

# Problem Statement

The task is to reduce high-dimensional data into a two-dimensional space using t-SNE and UMAP, compare their performance, and evaluate their ability to preserve local and global data structures.  
  
Key Questions:  
- How well do the methods separate clusters?  
- Which method better preserves local neighborhood relationships?  
- Which method is more computationally efficient?

# Task 1: Data Preparation and Analysis

Dataset Overview:  
- Input data: Preprocessed and standardized before applying dimensionality reduction.  
- Scaling: StandardScaler applied to normalize features.  
- Labels: A categorical target variable (used for coloring visualization plots).  
  
Preprocessing Steps:  
- Normalization using StandardScaler.  
- Handling categorical labels for visualization.  
- No significant missing values were reported in the notebook.

# Task 2: Dimensionality Reduction

## Method 1: t-SNE

Parameters Used:  
- Perplexity = 30  
- Learning Rate = 200  
- Iterations = 1000  
  
Results:  
- t-SNE produced clear local clusters with strong separation between groups.  
- However, the global distances between clusters are not reliable.  
  
Visualization Insight:  
- Excellent for detecting small, well-separated groups.  
- Computationally expensive for large datasets.

## Method 2: UMAP

Parameters Used:  
- n\_neighbors = 15  
- min\_dist = 0.1  
- random\_state = 42  
  
Results:  
- UMAP preserved both local clusters and global distances better than t-SNE.  
- Clusters were well-defined and reflected their true relationships in high-dimensional space.  
  
Visualization Insight:  
- Suitable for both clustering and downstream machine learning tasks.  
- Computationally more efficient than t-SNE.

# Task 3: Comparative Analysis

|  |  |  |
| --- | --- | --- |
| Aspect | t-SNE | UMAP |
| Local Structure | Strong preservation of local neighborhoods | Strong preservation |
| Global Structure | Weak (distances between clusters not reliable) | Better global preservation |
| Computation | Slower, especially on large datasets | Faster and scalable |
| Parameter Sensitivity | Highly sensitive to perplexity/learning rate | More stable across parameters |
| Best Use Case | Visualization of small, tight clusters | Visualization + downstream tasks |

# Challenges Faced

- Parameter Tuning:  
 \* t-SNE required careful adjustment of perplexity and learning rate.  
 \* UMAP was less sensitive but still required fine-tuning of n\_neighbors and min\_dist.  
  
- Interpretability:  
 \* Both methods provide excellent visualizations but do not directly give feature importance.

Conclusion and Recommendations

- t-SNE: Best suited for visualizing small datasets with complex local structures. Provides clearer cluster boundaries but sacrifices global interpretability.  
- UMAP: Recommended for larger datasets and when both local and global preservation are important. More computationally efficient and versatile.  
Final Recommendation: For general-purpose dimensionality reduction and visualization, UMAP is preferred due to its balance between local/global structure preservation and efficiency. t-SNE can be used when the primary focus is on revealing fine-grained local clusters.