CSCI 485

Assignment 2: Dimensionality Reduction on Wine Quality Data: A Comparative Analysis of PCA and t-SNE

By: Akshay Aralikatti

Date: 14/02/2025

Introduction

Dimensionality reduction is a critical step when working with high-dimensional data. This report demonstrates how to apply **Principal Component Analysis (PCA)** and **t-Distributed Stochastic Neighbor Embedding (t-SNE)** to a wine quality dataset to visualize and analyze its underlying structure. We compare these two techniques, highlighting their strengths and limitations in capturing global versus local structures.

Dataset Description

The dataset used in this analysis is a wine quality dataset. Each record includes various chemical properties of wine along with a quality score. The CSV file uses semicolons (`;`) as delimiters and contains the following columns:

- `fixed acidity`
- `volatile acidity`
- `citric acid`
- `residual sugar`
- `chlorides`
- `free sulfur dioxide`
- `total sulfur dioxide`
- `density`
- `pH`
- `sulphates`
- `alcohol`
- `quality`

Sample Data

```
"fixed acidity";"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur d
7;0.27;0.36;20.7;0.045;45;170;1.001;3;0.45;8.8;6
6.3;0.3;0.34;1.6;0.049;14;132;0.994;3.3;0.49;9.5;6
8.1;0.28;0.4;6.9;0.05;30;97;0.9951;3.26;0.44;10.1;6
```

Objectives

- **Dimensionality Reduction:** To reduce the high-dimensional wine dataset to two dimensions using PCA and t-SNE.

- **Visualization:** To visualize the data projections and compare PCA's ability to capture global structure against t-SNE's ability to reveal local clusters.

Methodology

Data Loading and Preprocessing

1. **Loading the Data:**
   - The dataset is loaded using `pandas` with a semicolon (`;`) delimiter.
2. **Separating Features and Labels:**
   - All columns except `quality` are used as features.
   - The `quality` column is used as labels for coloring the visualizations.
3. **Standardization:**
   - The features are standardized (mean = 0, variance = 1) to ensure each contributes equally during dimensionality reduction.

Example Code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
from sklearn.preprocessing import StandardScaler

# Load data from CSV with semicolon delimiter
data = pd.read_csv('wine_data.csv', delimiter=';')

# Separate features and labels
features = data.drop('quality', axis=1)
labels = data['quality']

# Standardize features
scaler = StandardScaler()
features_scaled = scaler.fit_transform(features)
```

PCA (Principal Component Analysis)

PCA is a linear technique that reduces dimensionality by projecting data onto directions of maximum variance. In this analysis, the data is reduced to 2 principal components.

Steps:

1. Apply PCA on the standardized features.
2. Print the explained variance ratio to assess how much variance is captured by the principal components.

t-Distributed Stochastic Neighbor Embedding (t-SNE) t-SNE is a non-linear technique that excels at preserving local structure, making it excellent for visualizing clusters within high-dimensional data.

Visualization

PCA 3D Scatter Plot:

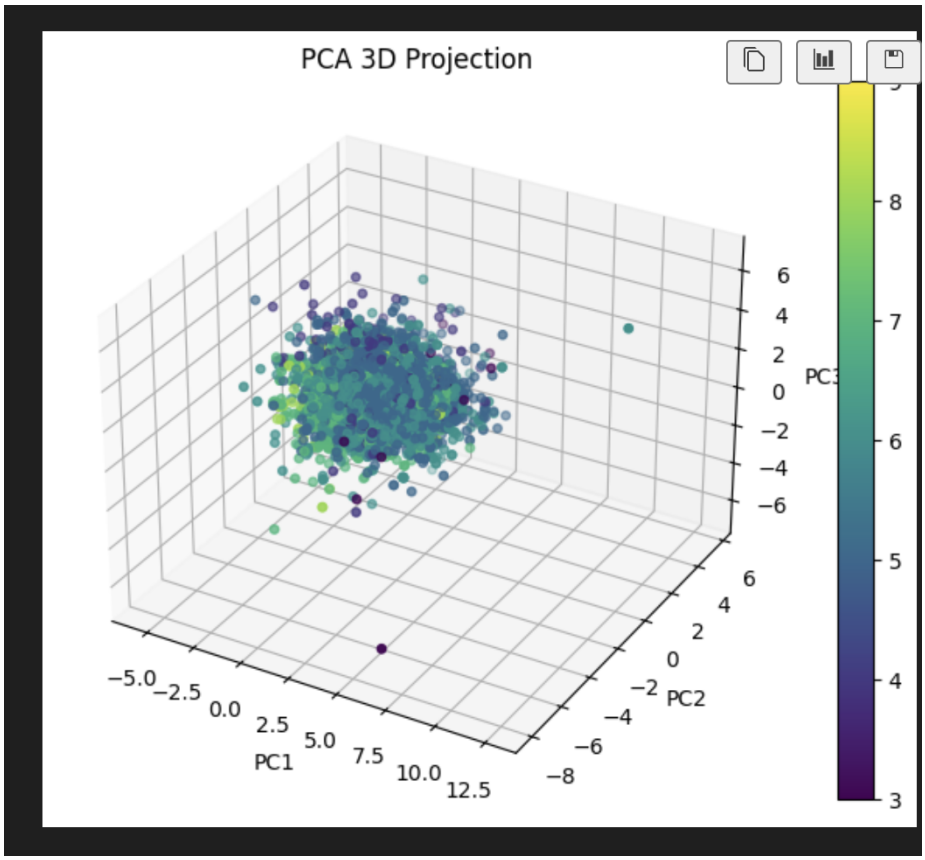

Figure 1: PCA 2D Scatter Plot

PCA vs t-SNE 2D Scatter Plot:

Results and Analysis:

PCA Visualization

- Global Structure: The PCA projection provides insight into the overall spread of the data by capturing the directions of maximum variance. The

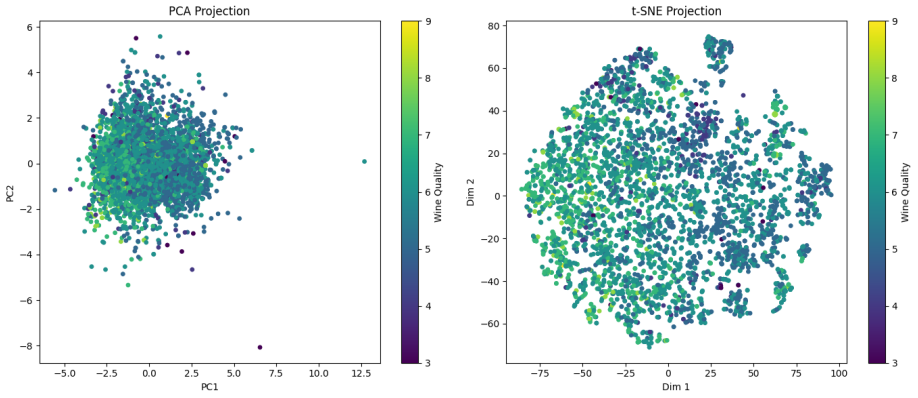

Figure 2: PCA vs t-SNE 2D Scatter Plot

explained variance ratio indicates the proportion of the total variance that is represented by the two principal components.

- Observations: The PCA plot gives a broad overview of how different wine samples are distributed based on their chemical properties. It shows the general relationships between samples but may not capture local clusters or non-linear relationships.

t-SNE Visualization

- Local Structure: The t-SNE projection emphasizes local relationships, making clusters or groupings of similar wine samples more evident. It is useful for identifying patterns that may not be apparent in the original high-dimensional space.
- Observations: The t-SNE plot reveals more distinct clusters compared to PCA, showing how samples with similar quality scores are grouped together based on their chemical properties. This local structure can provide more detailed insights into the data distribution.