MATH/CSCI 485 Assignment#2:PCA and Dimensionality Reduction

Objective

Use Principal Component Analysis (PCA) to explore and visualize a high-dimensional dataset and compare its effectiveness with t-SNE for dimensionality reduction.

Dataset

Use the **"Wine Quality"** dataset from the UCI Machine Learning Repository.

Task

1. **Data Preprocessing**
    - Retrieve the dataset (see hints below).
    - Normalize the dataset to ensure all features are on the same scale.
2. **Dimensionality Reduction & Visualization**
    - Apply PCA to reduce the dataset to **2 or 3** principal components.
    - Visualize the transformed data using **scatter plots (2D and 3D)**.
    - Identify the **variance explained** by each principal component and discuss the trade-off between dimensionality reduction and information loss.
3. **Comparison with t-SNE**
    - Apply **t-SNE** on the dataset to obtain a 2D representation.
    - Compare the results with PCA in terms of interpretability and clustering.
    - Discuss how PCA and t-SNE handle high-dimensional data differently.

Hints: How to Get the UCI Wine Quality Dataset

- The dataset is available at: UCI Machine Learning Repository – Wine Quality.
- It consists of two separate CSV files: **winequality-red.csv** and **winequality-white.csv**. You can choose to analyze one or combine both.
- Use `pandas.read_csv()` to load the dataset and `df.info()` to inspect it.
- Be sure to **handle missing values** if any exist.
- Normalize the features using `StandardScaler` from `sklearn.preprocessing`.

Deliverables

- A **Jupyter Notebook** containing:
    - Data loading and preprocessing steps.
    - PCA and t-SNE implementations.
    - Visualizations comparing the two methods.
    - Interpretation of results.
- A **brief report** (Markdown section in the notebook or a separate PDF document) discussing:
    - The trade-offs between PCA and t-SNE.
    - Key observations from the visualizations.

Grading Rubric (Total: 100 points)

- Data Preparation (15 points)
    - Dataset is correctly retrieved and loaded.
    - Data is properly cleaned and normalized.
- PCA Implementation (20 points)
    - PCA is correctly applied to the dataset.
    - Principal components are extracted and interpreted.
    - Explained variance is computed and analyzed.
- PCA Visualization (15 points)
    - 2D and/or 3D scatter plots are created.
    - Plots effectively illustrate PCA results.
- t-SNE Implementation (15 points)
    - t-SNE is correctly applied to the dataset.
    - Visualizations are clear and meaningful.
- Comparison & Discussion (20 points)
    - Thoughtful comparison of PCA and t-SNE.
    - Discussion on trade-offs, strengths, and weaknesses of both methods.
- Clarity & Code Quality (15 points)
    - Code is well-structured and easy to follow.
    - Proper documentation and comments are provided.

Submission Requirements

- Create a Github project that includes the following:
    - Name of your repository: MATH/CSCI485_Spring25_<Firstname>_<Lastname>
    - Within your repository, create project: Assignment_2
    - Within your project, include the following:
        - Source code (either python code, or python code in jupyter notebook)
        - A readme file describes how to use your code

        - A screenshot of how your code is successfully executed and generating required outputs, graphs and tables etc.
        - A PDF file of the report (Or embed your report in the Markdown sections within your Jupyter Notebook)
- Submission:
    - On Canvas Assignments, PDF report and link to your Github repo/projects
    - Make sure you have invited me to your github repo/projects

Additional Notes:

For submission of github links, you need to either make your github repo public, or add me as a collaborator. Here is an instruction on adding people as collaborators.

https://docs.github.com/en/account-and-profile/setting-up-and-managing-your-personal-account-on-github/managing-access-to-your-personal-repositories/inviting-collaborators-to-a-personal-repository

You can find me on github using my email: bshen2@csuchico.edu or my github account name: boshen-csuchico.