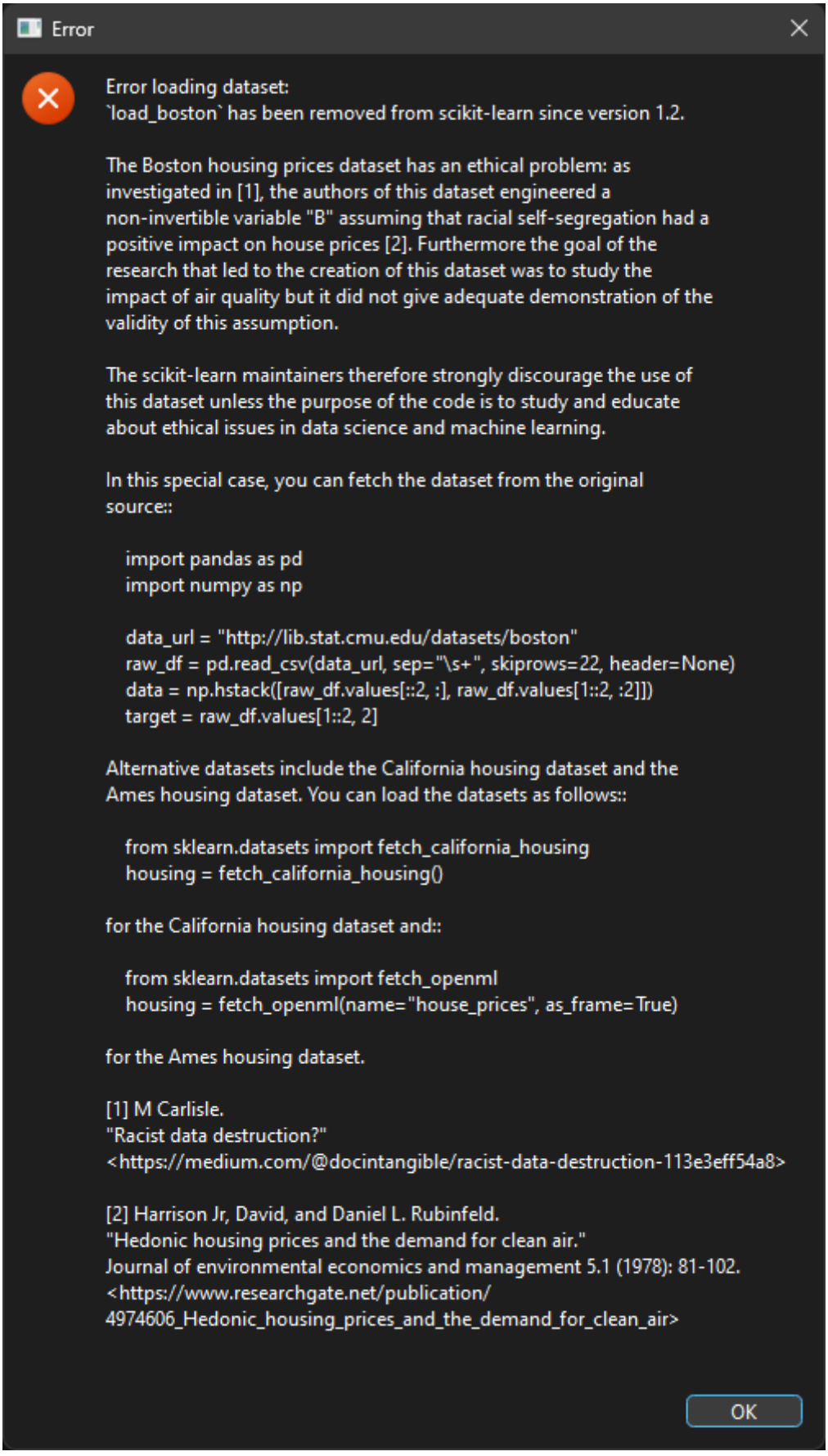
MKT3434 21067011 Furkan Karstarlı

Homework 2

- **Python 3.10 is used.**
- **Examples conducted with datasets as "iris-dataset classification"**
- **DONT FORGET TO LOOK AT README.md FILE!!!**
- **The Boston-Housing dataset examples are erased from the library due to some ethical issues. Here is the error information below:**


Error
Error loading dataset:
`load_boston` has been removed from scikit-learn since version 1.2.
The Boston housing prices dataset has an ethical problem: as
investigated in [1], the authors of this dataset engineered a
non-invertible variable "B" assuming that racial self-segregation had a
positive impact on house prices [2]. Furthermore the goal of the
research that led to the creation of this dataset was to study the
impact of air quality but it did not give adequate demonstration of the
validity of this assumption.
The scikit-learn maintainers therefore strongly discourage the use of
this dataset unless the purpose of the code is to study and educate
about ethical issues in data science and machine learning.
In this special case, you can fetch the dataset from the original
source::
import pandas as pd
import numpy as np
data_url = "http://lib.stat.cmu.edu/datasets/boston"
raw_df = pd.read_csv(data_url, sep="\s+", skiprows=22, header=None)
data = np.hstack([raw_df.values[::2, :], raw_df.values[1::2, :2]])
target = raw_df.values[1::2, 2]
Alternative datasets include the California housing dataset and the
Ames housing dataset. You can load the datasets as follows::
from sklearn.datasets import fetch_california_housing
housing = fetch_california_housing()
for the California housing dataset and::
from sklearn.datasets import fetch_openml
housing = fetch_openml(name="house_prices", as_frame=True)
for the Ames housing dataset.
[1] M Carlisle.
"Racist data destruction?"
<https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8>
[2] Harrison Jr, David, and Daniel L. Rubinfeld.
"Hedonic housing prices and the demand for clean air."
Journal of environmental economics and management 5.1 (1978): 81-102.
<https://www.researchgate.net/publication/
4974606_Hedonic_housing_prices_and_the_demand_for_clean_air>
OK

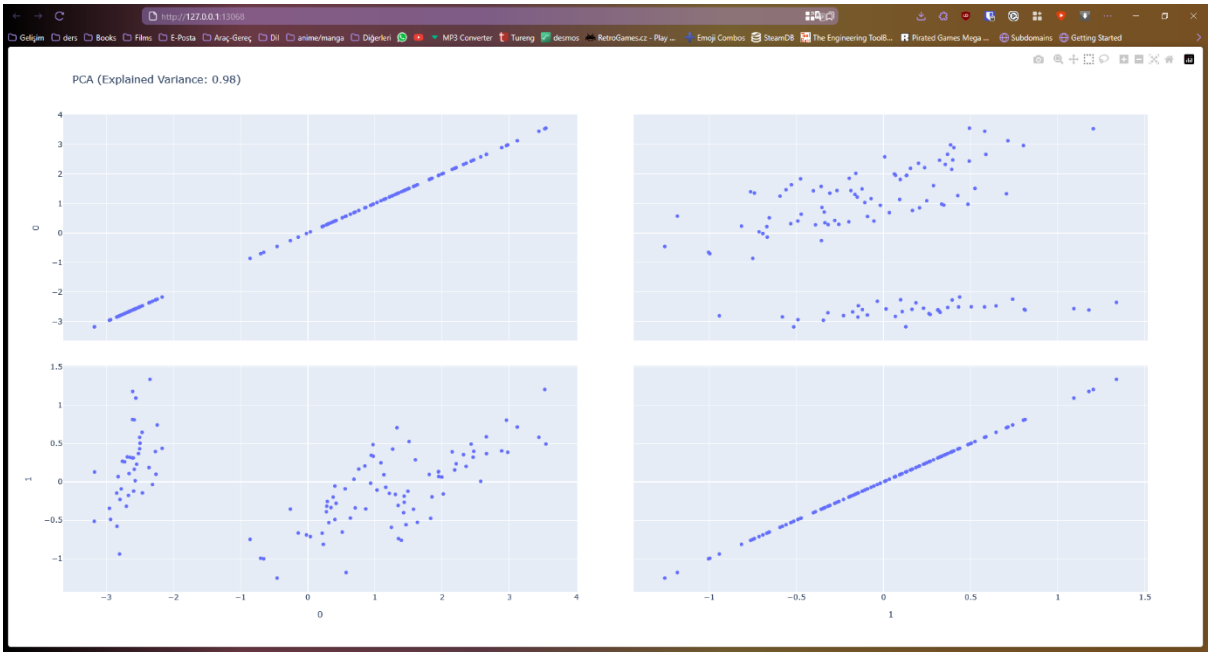
Implementation Summary and Dimensionality Reduction Comparison

This project is a PyQt6-based GUI for exploring machine learning models through an interactive interface. It supports dataset loading, preprocessing (imputation and scaling), and training models for regression, classification, SVM, and Naive Bayes. Evaluation is handled with visualizations (scatter plots, confusion matrices) and metrics like accuracy, MSE, and RMSE. Users can also run K-Fold cross-validation with configurable folds.

For dimensionality reduction, the app offers PCA, LDA, t-SNE, and UMAP with interactive visualization. PCA is a linear method that preserves global structure and is fast and interpretable. t-SNE is nonlinear, better for visualizing clusters, and focuses on preserving local relationships but is slower and more sensitive to parameters. Both methods help users understand and visualize complex data effectively.

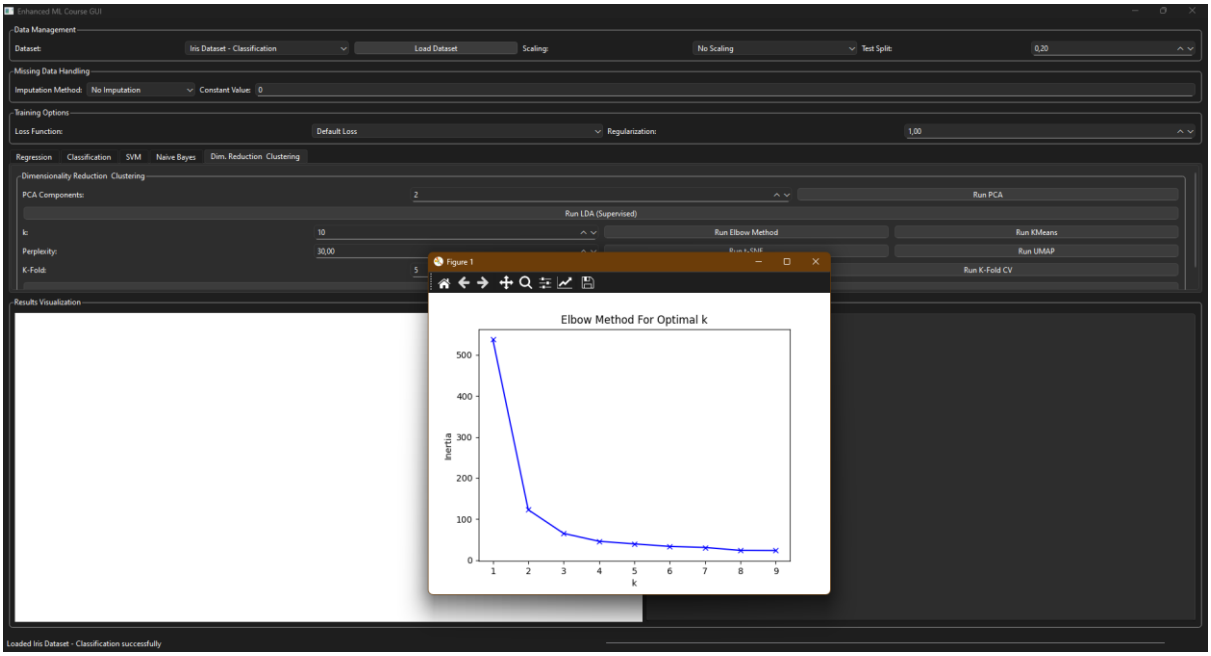
New tab added as “<u>Dimensionality Reduction Clustering</u>”

Running PCA at Component value = 2



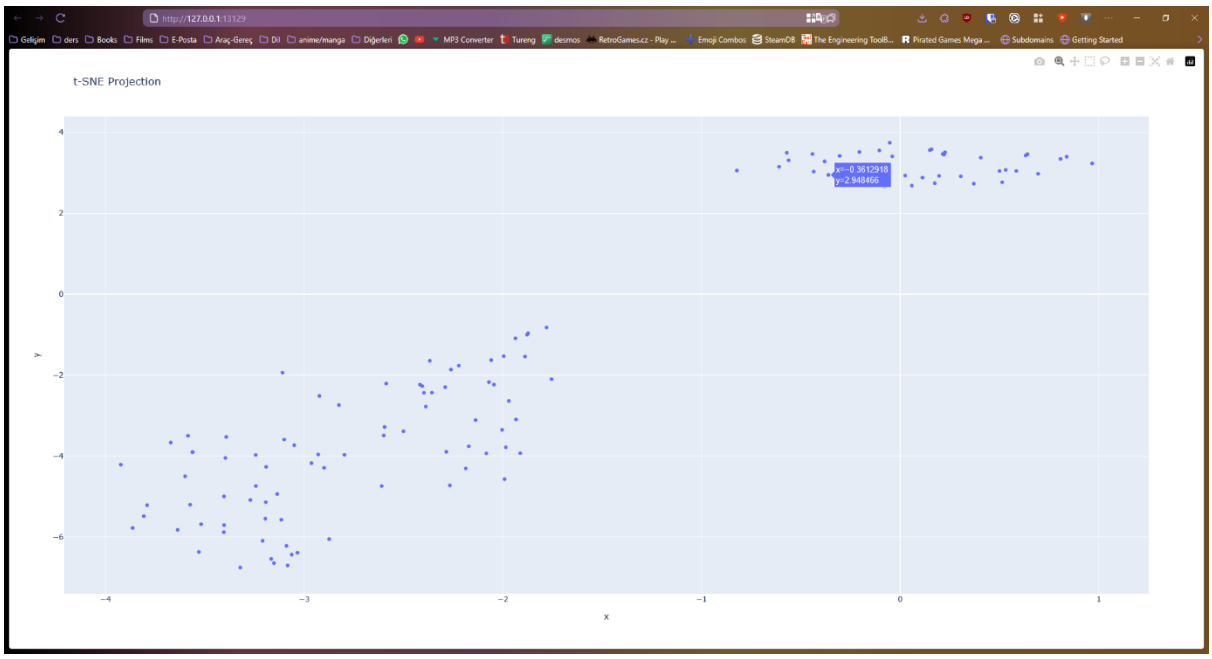
Supervised LDA

Running Elbow Method at k = 10



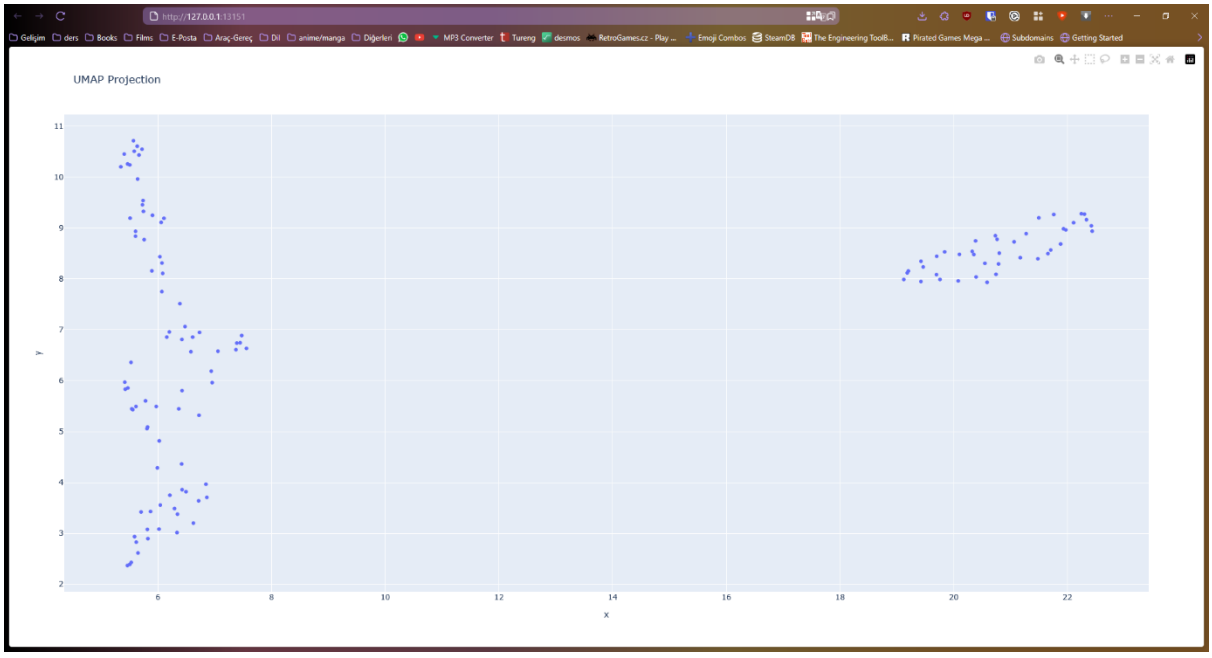
Running K-Means at k=10

Running t -SNE Perplexity at 30

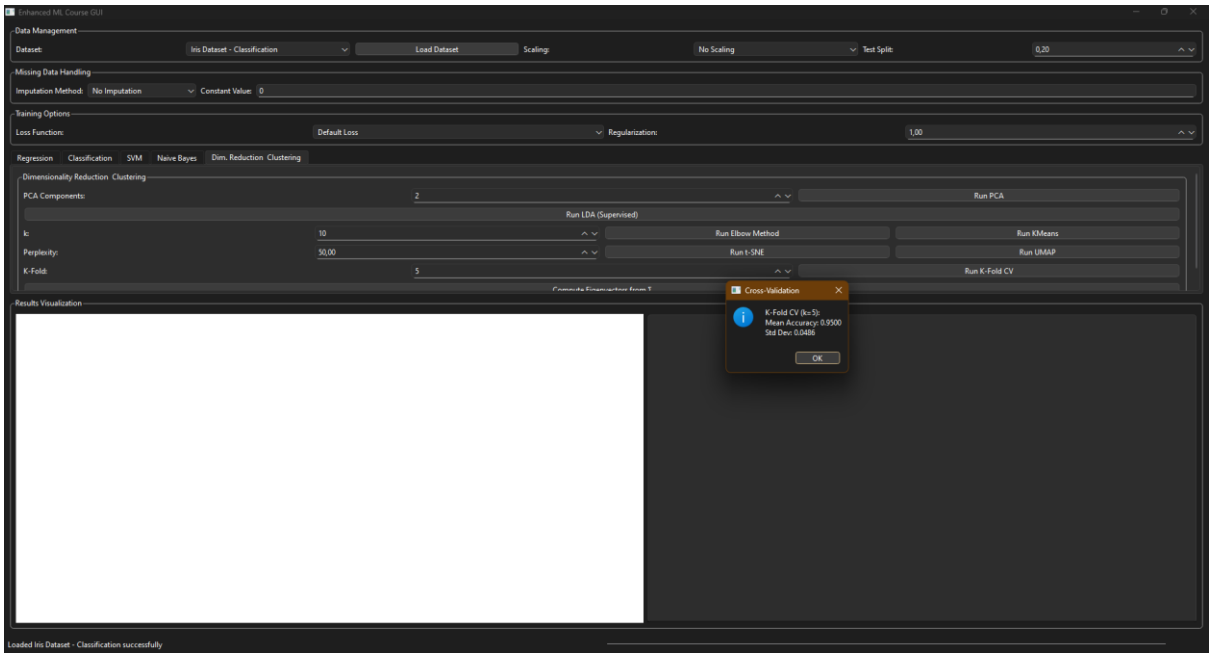


Running UMAP Perplexity at 30

Note: The UMAP function work slowly, please be patient while using that!!!



Running K-Fold at 5



Eigenvector Computation