Compare classifiers in scikit-learn library.

Introduction

This report analyzes the application of Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Kernel PCA (KPCA) for dimensionality reduction on the Iris and MNIST datasets. The performance of these techniques was evaluated by feeding the reduced data into a decision tree classifier, calculating accuracy, execution time , explained variance ration and running classification report.

Methodology

Datasets:

- Iris dataset shape is (150, 4).
- The MNIST dataset shape is (70000, 784). We extract a subset of the dataset by using the train-test-split method with the stratify option (stratify=yes) and then split MNIST subset to X_train and X_test for the calculations. The MNIST subset shape is(2100, 784)
- Both datasets are split into 70% training and 30% testing sets, with stratification to maintain class distribution.

Evaluation Metrics:

The reduction techniques were implemented in Python using the scikit-learn library and evaluated based on the following metrics.

- Accuracy
- Execution Time
- Explained Variance Ratio
- Classification Report

Results and Analysis

Iris Dataset:

- PCA:
    - Accuracy: 0.8222 for both (criterion=entropy, max_depth=6) and (criterion=gini, max_depth=6)
    - Precision, Recall, and F1-Score: The model performed well for Class 0 (precision=1.00, recall=1.00, f1-score=1.00) but struggled with Classes 1 and 2 (precision, recall and f1-score = 0.73).
    - PCA Execution Time: 0.0039 seconds
    - Decision Tree Execution Time: 0.0055-0.0114 seconds (depending on criterion and max depth)
- LDA
    - Accuracy: 0.9556 for both (criterion=entropy, max_depth=6) and (criterion=gini, max_depth=6)
    - Precision, Recall, and F1-Score: LDA achieved high performance across all classes, with Class 0(precision=1.00, recall=1.00, f1-score=1.00) and for Classes 1 and 2 close to 1.00 for all 3.
    - LDA Execution Time: 0.0036 seconds
- KPCA
    - Accuracy: 0.9111 for both (criterion=entropy, max_depth=6) and (criterion=gini, max_depth=6)
    - Precision, Recall, and F1-Score: LDA achieved high performance across all classes, with Class 0(precision=1.00, recall=1.00, f1-score=1.00) and for Classes 1 and 2 slightly lower performance for all 3.
    - Kernel PCA Execution Time: 0.0125 seconds

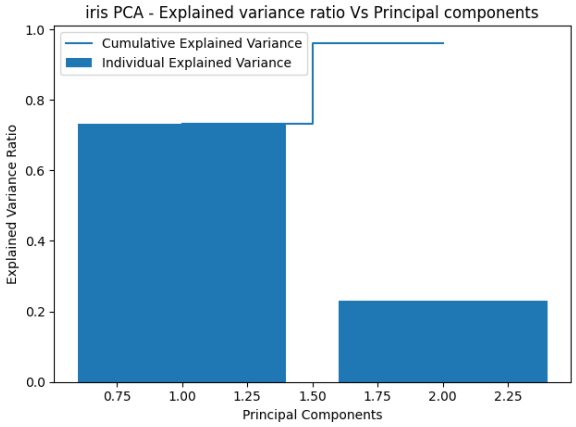
- Decision Tree Execution Time: (criterion=entropy, max_depth=6) Execution Time: 0.0114 and (criterion=gini, max_depth=6) Execution Time: 0.0055.
- Explained variance ratio: [0.73121106 0.23053026].

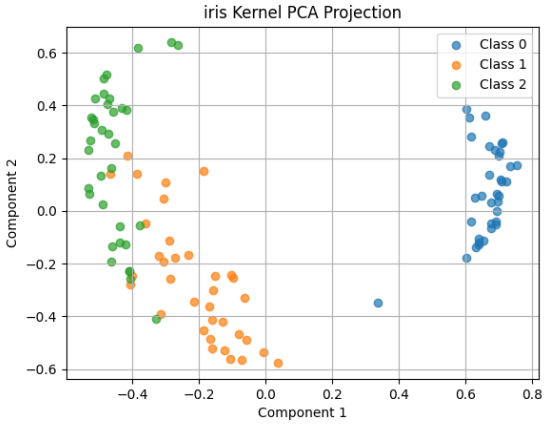
MNIST Dataset:

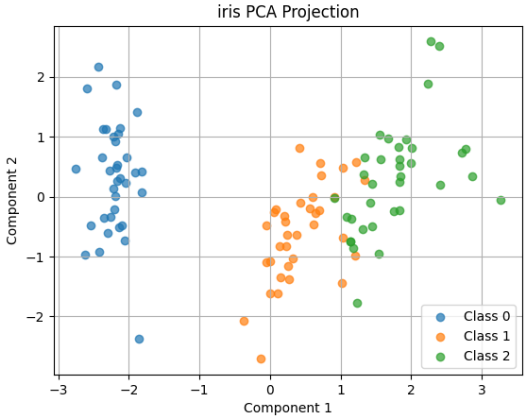
- PCA:
    - Accuracy: 0.4175 for (criterion=entropy, max_depth=6) and 0.3714 for (criterion=gini, max_depth=6).
    - Precision, Recall, and F1-Score: The model performed poorly for all classes. Class 1 had the best performance (precision=0.88, recall=0.86, f1-score=0.87).
    - PCA Execution Time: 0.1062 seconds
- LDA
    - Accuracy: 0.4222 for (criterion=entropy, max_depth=6) and 0.4190 for (criterion=gini, max_depth=6)
    - Precision, Recall, and F1-Score: The model performed poorly for all classes. Class 1 had the best performance (precision=0.74, recall=0.77, f1-score=0.76).
    - LDA Execution Time: 0.6072 seconds
- KPCA
    - Accuracy: 0.1127 for (criterion=entropy, max_depth=6) & (criterion=gini, max_depth=6)
    - Precision, Recall, and F1-Score: The model performed very poorly for all classes. Class 1 had the best performance (precision=0.11, recall=1.00, f1-score=0.20).
    - Kernel PCA Execution Time: 0.2715 seconds
- Decision Tree Execution Time: (criterion=gini, max_depth=6) Execution Time: 0.0424 and (criterion=entropy, max_depth=6) Execution Time: 0.0292
- Explained variance ratio: [0.06528144 0.04586195]

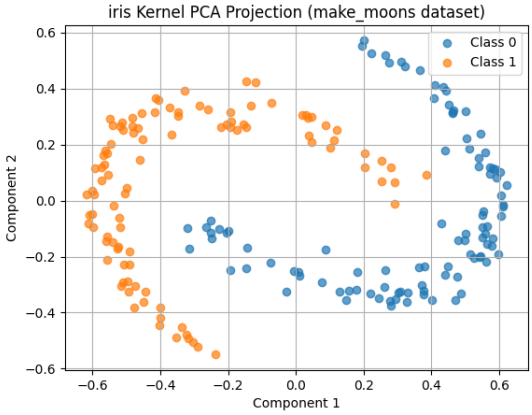
Plots

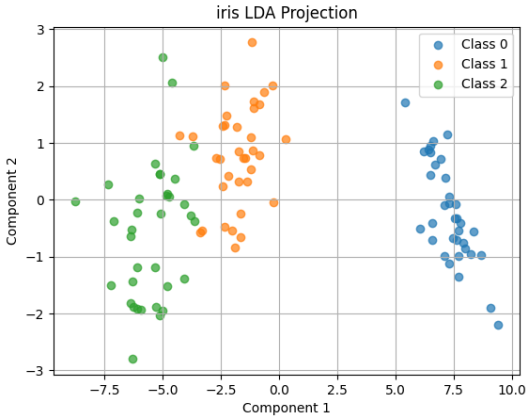
Below are a few of the plots that will be generated by the program.

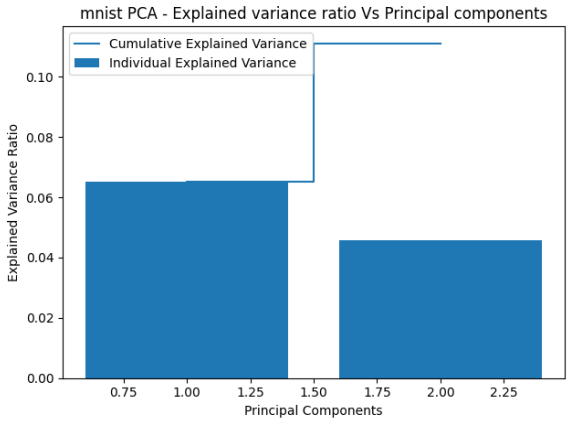


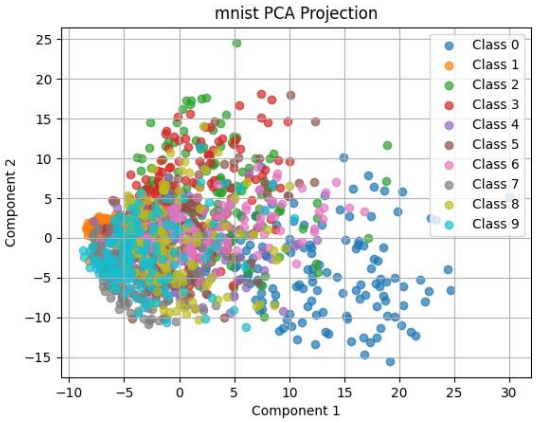


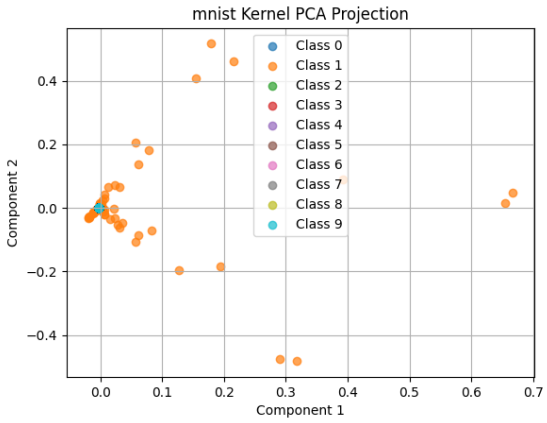


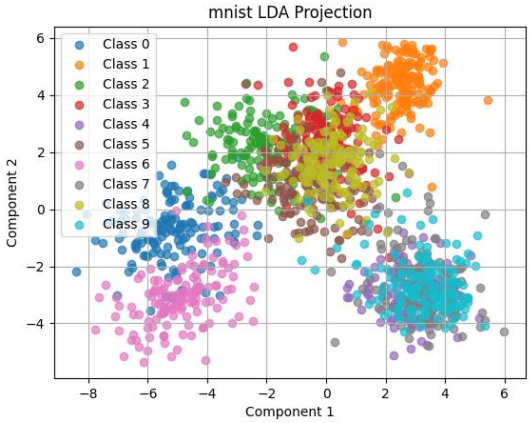


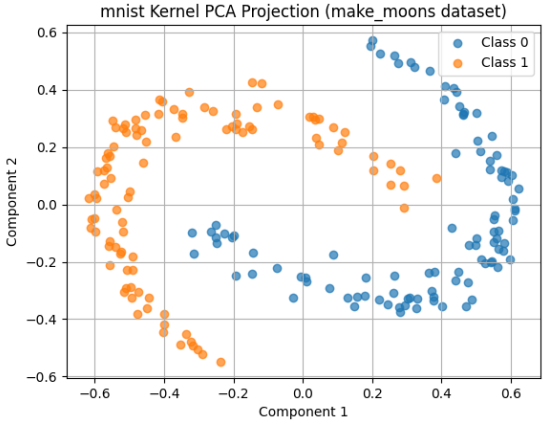





mnist PCA Projection
Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
Class 9
Component 1
Component 2


mnist Kernel PCA Projection
Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
Class 9
Component 1
Component 2


mnist LDA Projection
Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
Class 9
Component 1
Component 2


mnist Kernel PCA Projection (make_moons dataset)
Class 0
Class 1
Component 1
Component 2