HW-4 DIMENSIONALITY REDUCTION TECHNIQUES

INTRODUCTION:

Dimensionality reduction is a crucial machine learning preprocessing step that optimizes model performance by decreasing noise, expediting training durations, and simplifying models. This study focuses on two datasets: the huge, high-dimensional MNIST dataset and the tiny, low-dimensional Iris dataset to investigate the performance of various algorithms on datasets with distinct features.

METHODOLOGY:

Using the dataset loading utilities provided by sci-kit-learn, the Iris dataset—which has 150 instances with four features—and a portion of the MNIST dataset—which has 5000 instances of digit images—were loaded.

PREPROCESSING:

The feature values for MNIST were scaled to a range of [0, 1]. A 70-30 ratio was used to divide the two datasets into training and test sets**.**

DIMENSIONALITY REDUCTION TECHNIQUES:

1. PCA is a linear technique that reduces dimensions by retaining the components with the highest variance after transforming the data into a new coordinate system.

2. A supervised technique called LDA (exclusive to Iris) aims to maximize class separability. It was only used on the Iris dataset due to class limitations.

3. PCA extended with kernel functions that can be used to capture non-linear interactions is called kernel PCA. We conducted our trials using the RBF kernel.

CLASSIFICATION ALGORITHMS:

Decision Tree Classifier (Iris): Designed for smaller datasets, the Decision Tree Classifier (Iris) is a straightforward and efficient classifier that is favored for its interpretability.

Support Vector Machine (SVM, MNIST): An RBF kernel equipped SVM, which is well-known for working well in situations like this, was used because of MNIST's high dimensionality.

RESULTS:

	Accuracy		F1 Score		Training Time	
Dataset	Iris	MNIST	Iris	MNIST	Iris	MNIST
Method						
KernelPCA	1.0	0.921333	1.0	0.921079	0.204380	0.635451
LDA	1.0	0.931333	1.0	0.931202	0.151579	0.247583
PCA	1.0	0.952667	1.0	0.952614	0.150352	0.631832

CONCLUSION:

Our research indicates that non-linear dimensionality reduction methods like Kernel PCA are better suited to complicated datasets like MNIST, yet LDA and PCA perform extraordinarily well on simpler datasets like Iris. Computational efficiency suffers as a result, though.