Handwritten Digit Recognition using Classical Machine Learning

1. Introduction

This assignment focuses on building a handwritten digit recognition system using classical machine learning techniques. The MNIST dataset in CSV format was used, where each image is represented as a flattened 28×28 grayscale image with 784 pixel features. The objective was to design an end-to-end machine learning pipeline including data preprocessing, model training, evaluation, and performance comparison, without using any pre-trained models or neural networks.

2. Dataset and Preprocessing

The dataset consists of labeled handwritten digits from 0 to 9. Initial data exploration was performed to analyze the number of samples, class distribution, and sample images. The dataset was found to be clean with no missing values. Pixel values were normalized to the range 0–1 to improve model performance. Dimensionality reduction using Principal Component Analysis (PCA) was applied to reduce computational cost and improve efficiency, especially for SVM.

3. Model Implementation

Three classical machine learning models were implemented:

- K-Nearest Neighbors (KNN) with tuning of the number of neighbors.
- Support Vector Machine (SVM) using the **RBF kernel** with tuning of **C and gamma parameters.**
- Decision Tree with tuning of maximum depth and minimum samples split.

Hyperparameters were tuned manually to ensure simplicity and transparency.

4. Model Evaluation

The models were evaluated using accuracy and confusion matrices. **SVM achieved the highest accuracy**, followed by KNN and Decision Tree. Confusion matrix analysis showed that **SVM made fewer misclassifications across most digit classes.** Misclassified images were visualized to understand error patterns.

5. Ensemble Learning

A voting ensemble combining KNN, SVM, and Decision Tree was implemented using hard voting. The ensemble achieved comparable or slightly improved accuracy compared to individual models, demonstrating improved robustness by combining multiple classifiers.

6. Observations on Misclassification

Most misclassifications occurred between visually similar digits such as 3 and 5, 4 and 9, and 7 and 1. These errors were mainly due to variations in handwriting styles, unclear strokes, and overlapping pixel patterns.

7. Conclusion and Future Improvements

Among the evaluated models, **SVM proved to be the most effective for handwritten digit recognition due to its ability to handle high-dimensional data and non-linear patterns**. KNN performed reasonably well but was computationally expensive during prediction, while Decision Tree showed lower generalization performance.

Future improvements could include better feature extraction, increased training data, dimensionality reduction techniques, and ensemble methods. Overall, this assignment demonstrates a complete and systematic implementation of classical machine learning techniques for image classification.