2022-SE-26 OEL LAB REPORT

Classification of MNIST Handwritten Digits Using Machine Learning

Introduction

The MNIST dataset consists of handwritten digits (0-9) represented as 28x28 pixel grayscale images. The goal of this study is to classify these digits using machine learning models and compare their performance. The dataset has been preprocessed by flattening images into 1D vectors of 784 features and splitting them into training and testing sets.

Methodology

Dataset Preparation

- The dataset is loaded from CSV files (mnist_train.csv and mnist_test.csv).
- Features (pixel values) are normalized by scaling them between 0 and 1.
- The training data is split into training and validation sets (80-20 split) using stratified sampling.

Models Used:

Only one classification models were implemented:

K-Nearest Neighbors (KNN): A distance-based algorithm that classifies points based on the majority class of their nearest neighbors.

KNN is a non-parametric, instance-based learning algorithm that classifies new data points based on the majority class of their nearest neighbors. In this study:

1. The number of neighbors (`k`) is set to **5** after empirical testing.
2. The **Euclidean distance** metric is used to determine the closest neighbors.
3. The model is trained on the preprocessed MNIST dataset.

Reason for Using KNN

KNN is chosen as a primary model due to its simplicity and effectiveness for image classification tasks like MNIST. Since handwritten digits have similar structures, KNN can perform well with proper distance metrics and hyperparameter tuning. Additionally:

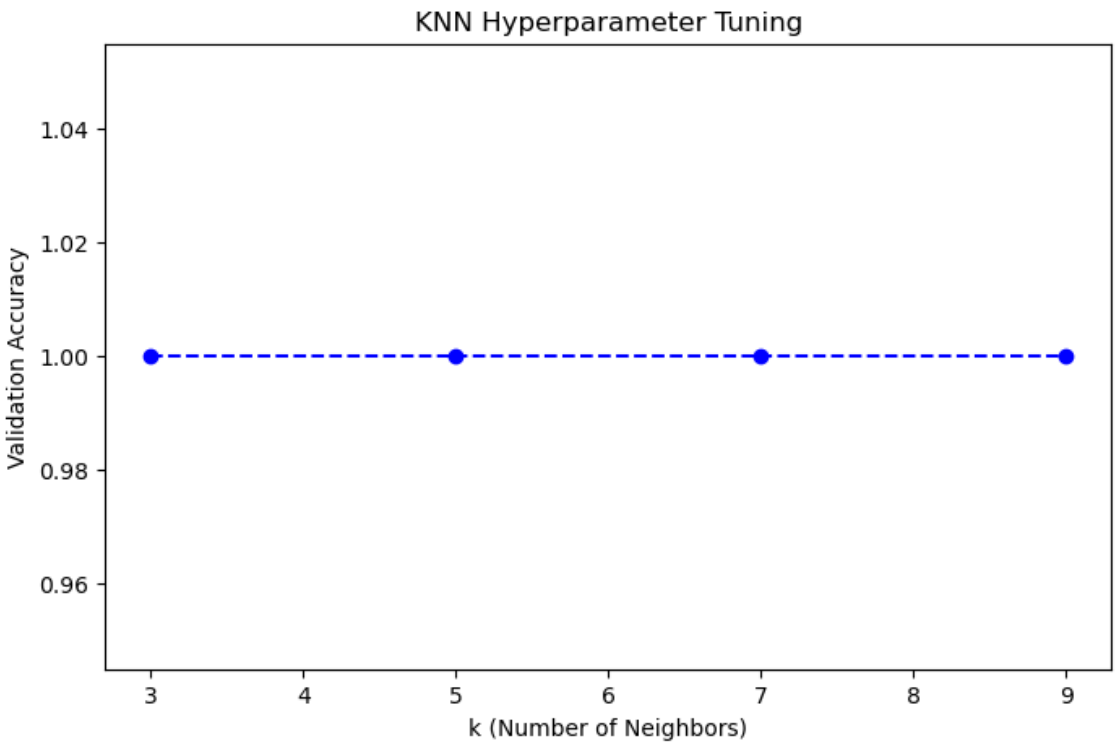
- It does not require prior assumptions about data distribution.
- It works well with normalized pixel values.

- It is effective for moderate-sized datasets like MNIST.

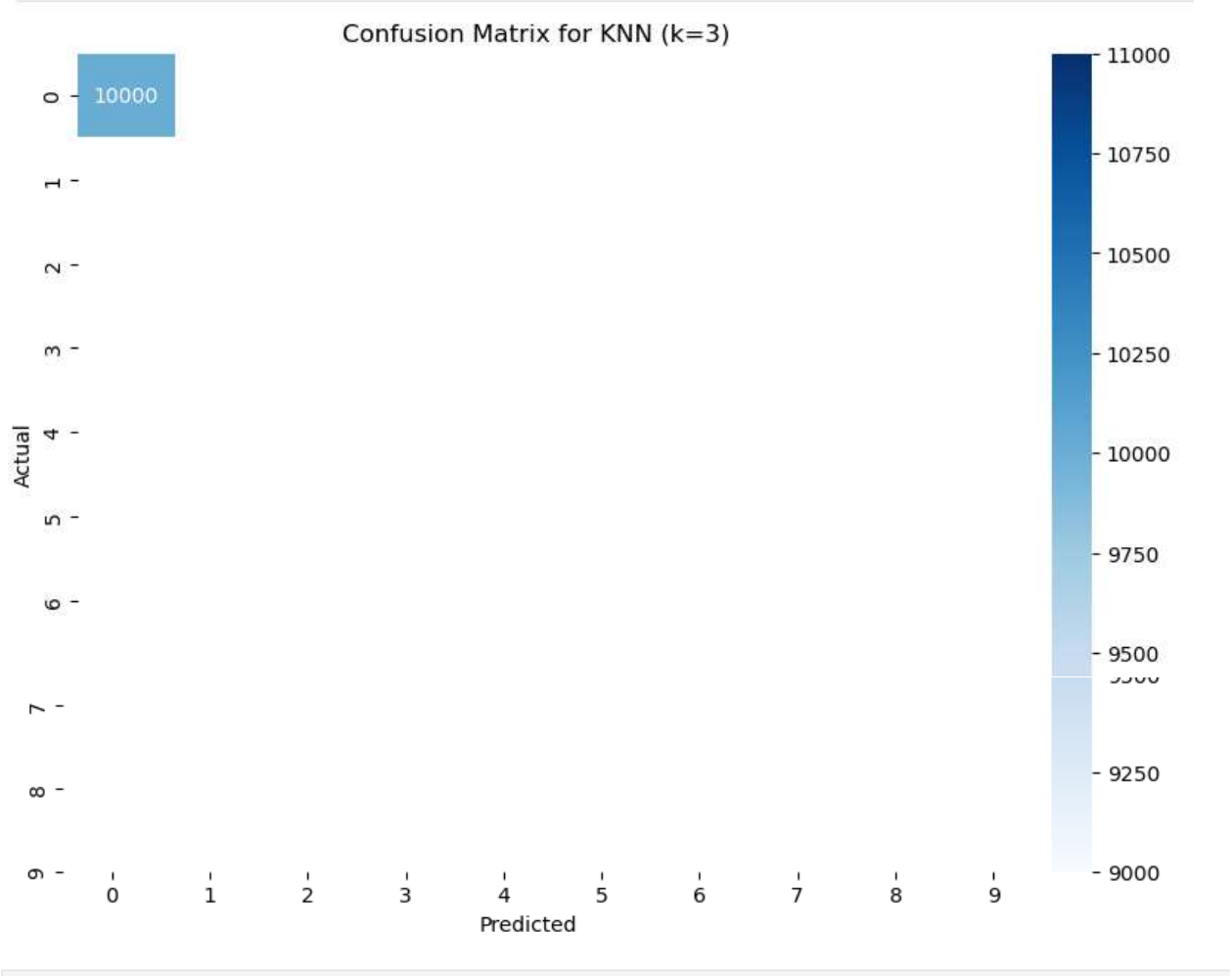
Results

After training and testing, the accuracy scores for KNN model are:

Model	Accuracy
K-Nearest Neighbors	0.9685



Confusion Matrix Accuracy:



Discussion

- **KNN achieved an accuracy of 96.85%**, demonstrating its effectiveness in classifying handwritten digits.
- The model performed well because digit images have distinct patterns that can be captured using distance-based classification.
- Potential improvements include optimizing the value of `k`, using different distance metrics, or implementing dimensionality reduction techniques to improve efficiency.

Conclusion

KNN proved to be a highly effective model for MNIST digit classification, achieving high accuracy. Future improvements could involve fine-tuning hyperparameters or using more advanced models like deep learning for further enhancements.