Optimized SVM for Digit Classification - Report

Objective

The notebook’s goal is to classify handwritten digit images using machine learning, focusing on building and optimizing a Support Vector Machine (SVM) classifier. It compares results with Decision Tree and Random Forest classifiers.

Dataset and Preprocessing

Dataset: Pen-Based Recognition of Handwritten Digits (UCI). Contains 10 digit classes (0–9) with 16 integer-valued features per sample and ~11,000 samples total. Preprocessing: Verified no missing values or duplicates, split into features (X) and labels (y), then divided into 90% training (9892 samples) and 10% testing (1100 samples). Scaling applied within pipelines.

Approach and Optimization

Primary classifier: Support Vector Machine (SVC) in a pipeline with MinMaxScaler. Hyperparameters (C, gamma, kernel) tuned via GridSearchCV with 10-fold cross-validation. Comparative models: Decision Tree and Random Forest, also tuned with GridSearchCV.

Key Implementation Steps

1. Load and inspect dataset. 2. Split into train/test sets. 3. Build pipelines with MinMaxScaler + classifier. 4. Tune hyperparameters with GridSearchCV. 5. Evaluate models on test set using accuracy, F1-score, and confusion matrix.

Results

Support Vector Machine (SVM): 97.8% accuracy, macro-F1 ~0.978. Decision Tree: 95.5% accuracy, macro-F1 ~0.955 (overfits training data). Random Forest: 87.5% accuracy, macro-F1 ~0.871 (underfits). Comparison: SVM > Decision Tree > Random Forest in performance.

Conclusions

Optimized SVM provided the best results with strong generalization and high accuracy. Decision Tree showed overfitting, while Random Forest underperformed. The study highlights the effectiveness of SVM with tuned hyperparameters in digit recognition.