MNIST Digit Classification with k-NN and Decision Trees

Ada Boran Yılmaz

1 Overview

This report presents a comprehensive approach to classifying handwritten digits (0–9) from the MNIST dataset using two different machine learning algorithms:

- k-Nearest Neighbors (k-NN)
- Decision Tree

We begin with an overview of the dataset and our preprocessing steps. Next, we detail the hyperparameter tuning for each model and present the final evaluation results on a held-out test set. We conclude with an analysis of misclassifications and potential explanations for errors.

2 Dataset and Preprocessing

2.1 Data Loading

- Loads the MNIST data using Keras API
- Splits the training data as 80% training set and 20% validation set
- Verifies the shapes

Training set shape: (48000, 28, 28)

Validation set shape: (12000, 28, 28)

Test set shape: (10000, 28, 28)

2 Dataset and Preprocessing

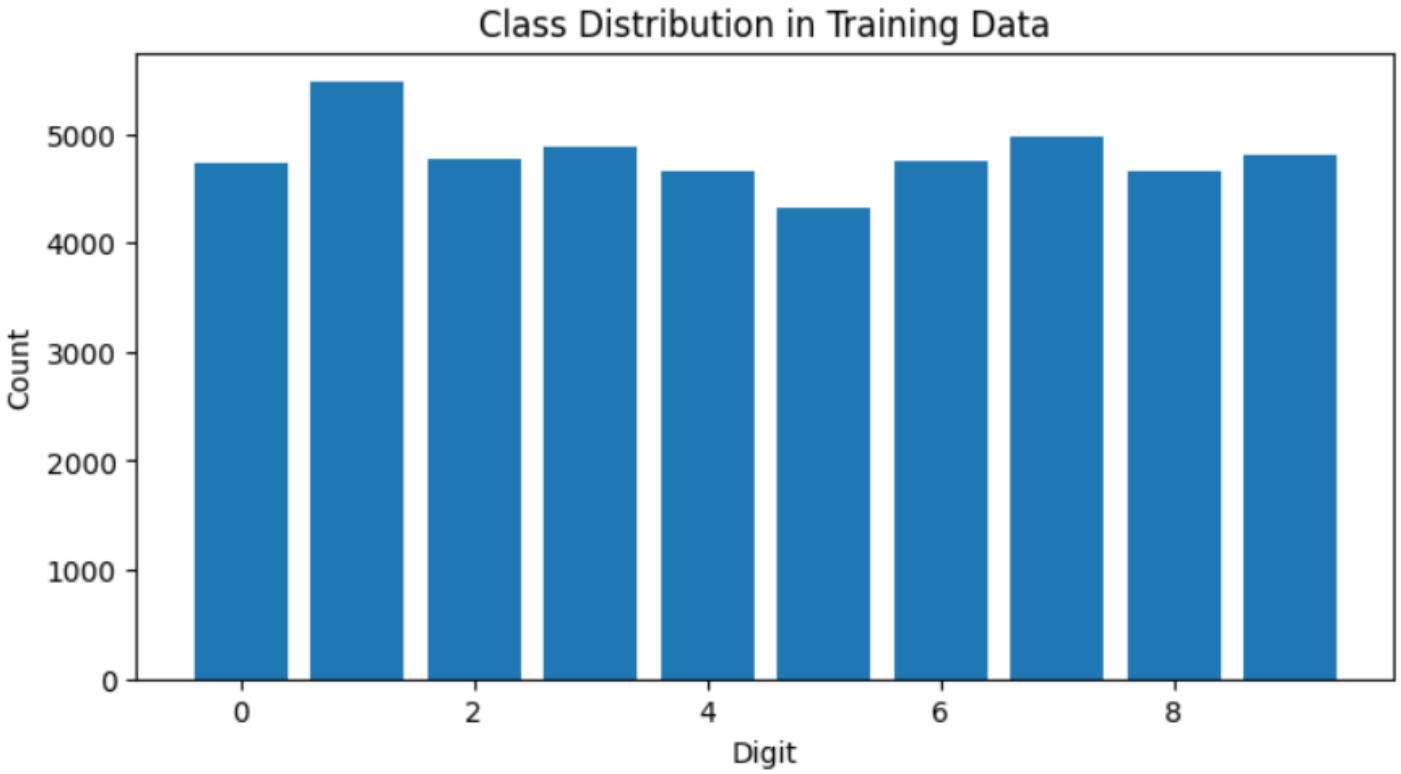
2.2 Data Analysis

- Computes class distribution for training data
- Plots class distribution as a bar chart
- Calculates the basic statistics such as mean and standard deviation of pixel values.
- Visualizes one sample per digit

Class Distribution: {0: 4729, 1: 5470, 2: 4762, 3: 4889, 4: 4655, 5: 4324, 6: 4748, 7: 4968, 8: 4653, 9: 4802}

2 Dataset and Preprocessing

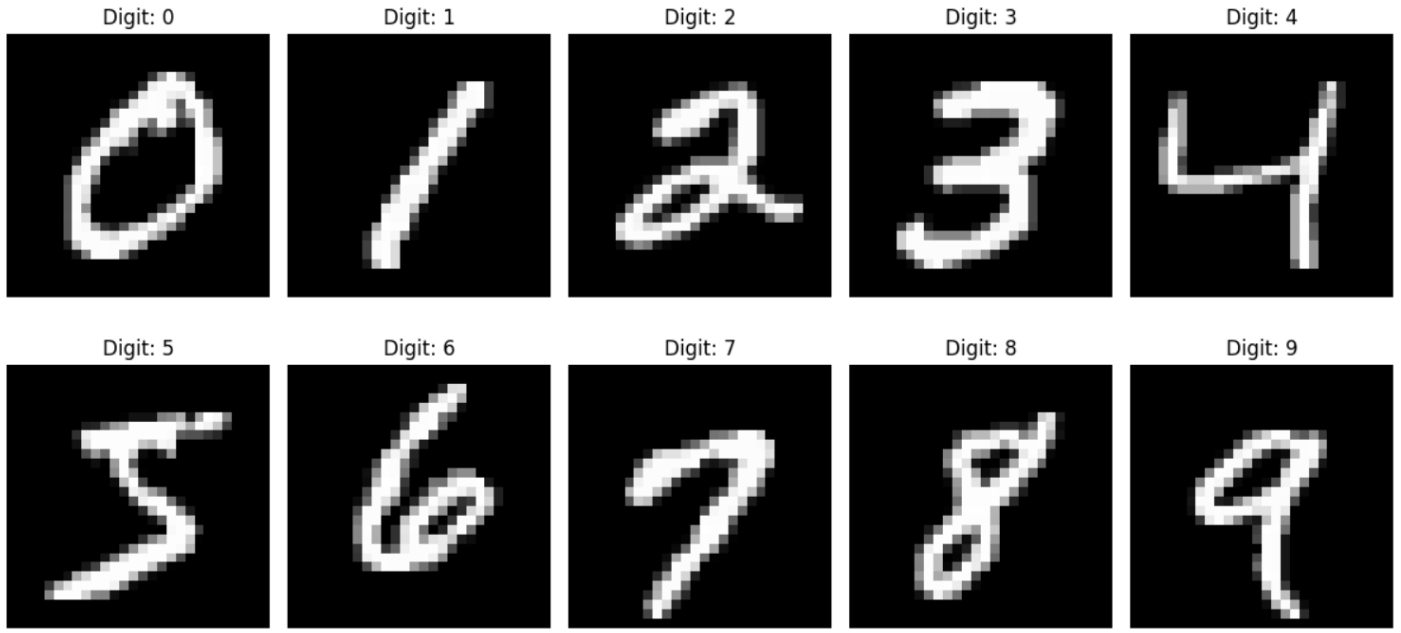
2.2 Data Analysis



2 Dataset and Preprocessing

2.2 Data Analysis

Mean pixel value: 33.37, Standard Deviation of pixel values: 78.64



2 Dataset and Preprocessing

2.3 Data Preprocessing

- Normalizes the images by scaling pixel values to [0, 1]
- Verifies that the shapes are unchanged

Training set shape (normalized): (48000, 28, 28)

Validation set shape (normalized): (12000, 28, 28)

Test set shape (normalized): (10000, 28, 28)

3 k-NN Classifier

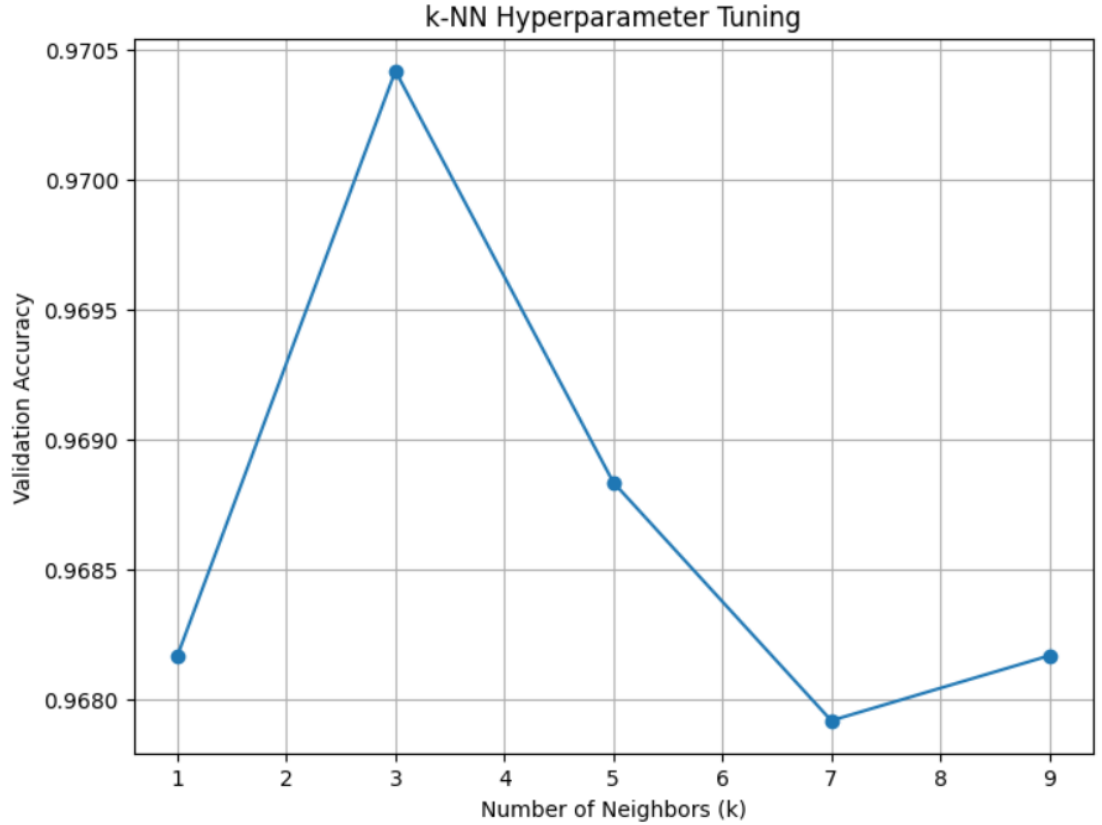
3.1 Model Initialization and Hyperparameter Tuning

- Flattens the 28x28 images for training and validation
- Defines the neighbor values to try
- Loops over each neighbor value and evaluate on the validation set
- Plots the validation accuracy versus the number of neighbors

k = 1, Validation Accuracy: 0.9682; k = 3, Validation Accuracy: 0.9704;
k = 5, Validation Accuracy: 0.9688; k = 7, Validation Accuracy: 0.9679;
k = 9, Validation Accuracy: 0.9682

3 k-NN Classifier

3.1 Model Initialization and Hyperparameter Tuning



3 k-NN Classifier

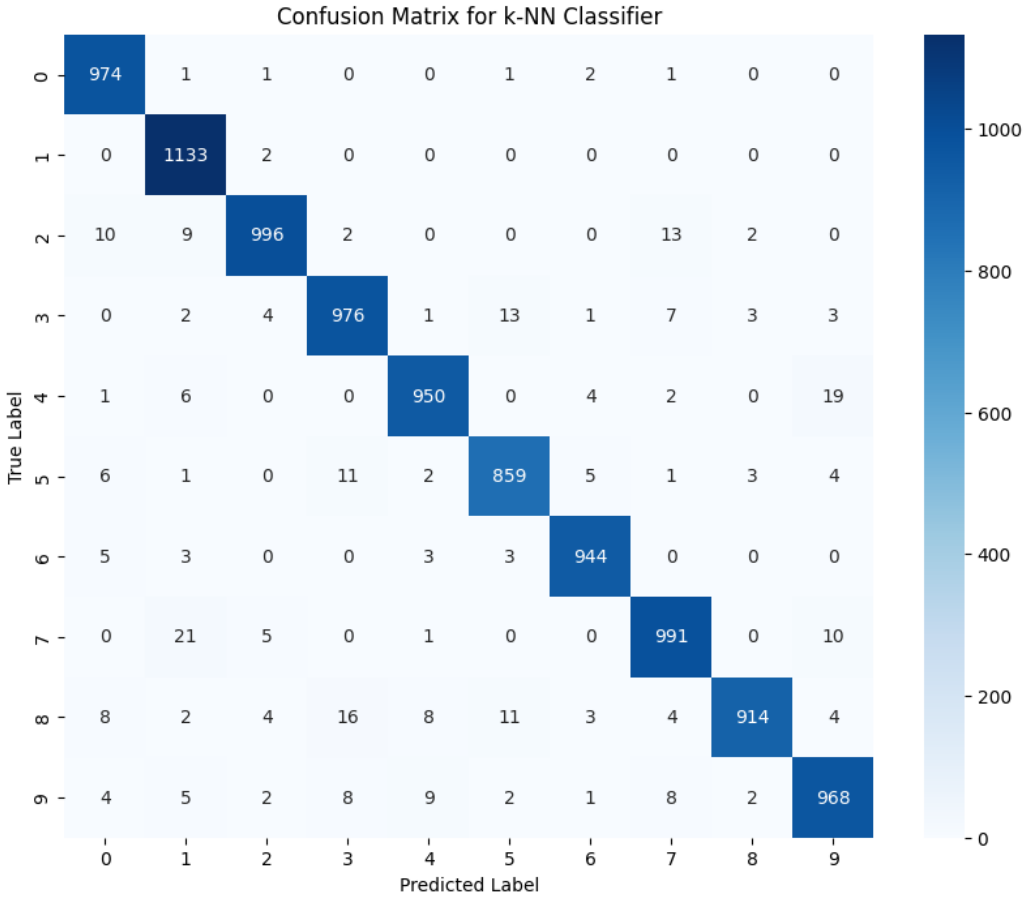
3.2 Final Model Training and Evaluation

- Concatenates training and validation sets, then reshapes images into 2D arrays.
- Uses the best k from validation to fit the final k-NN classifier.
- Predicts labels on the test set and computes accuracy, precision, recall, and F1-score.
- Displays a confusion matrix and calculates per-digit misclassifications.
- Identifies and prints the top three digits with the highest misclassification rates.
- Randomly selects and visualizes five misclassified examples with true and predicted labels.

3 k-NN Classifier

3.2 Final Model Training and Evaluation

```
Optimal number of neighbors (k): 3
Test Accuracy: 0.9705
Test Precision: 0.9707
Test Recall: 0.9705
Test F1 Score: 0.9705

Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       980
           1       0.96      1.00      0.98      1135
           2       0.98      0.97      0.97      1032
           3       0.96      0.97      0.96      1010
           4       0.98      0.97      0.97       982
           5       0.97      0.96      0.96       892
           6       0.98      0.99      0.98       958
           7       0.96      0.96      0.96      1028
           8       0.99      0.94      0.96       974
           9       0.96      0.96      0.96      1009

    accuracy                           0.97     10000
   macro avg       0.97      0.97      0.97     10000
weighted avg       0.97      0.97      0.97     10000
```



3 k-NN Classifier

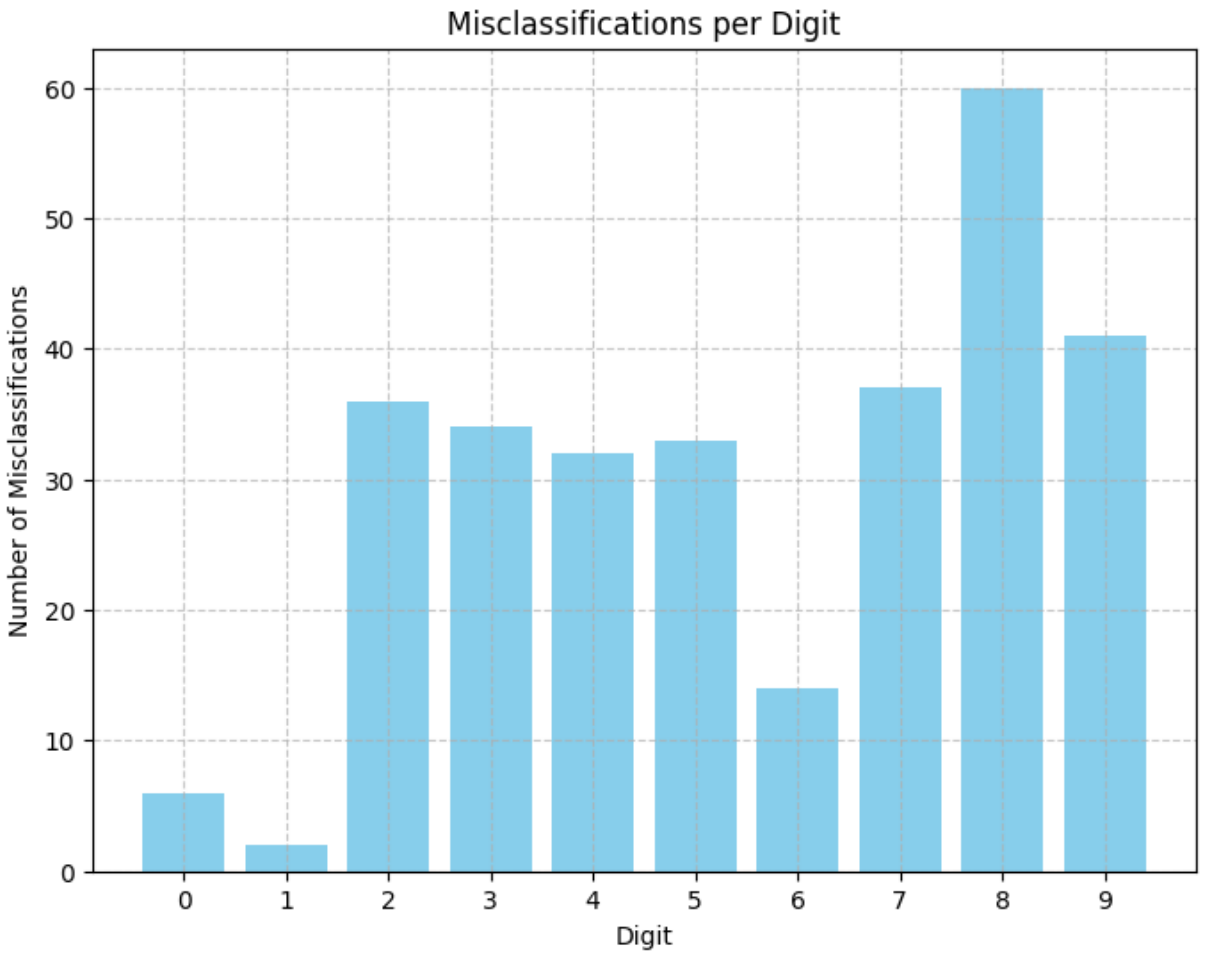
3.2 Final Model Training and Evaluatior

Total misclassified examples: 295

- Digit 0: 6 misclassifications, 0.61% misclassification rate
- Digit 1: 2 misclassifications, 0.18% misclassification rate
- Digit 2: 36 misclassifications, 3.49% misclassification rate
- Digit 3: 34 misclassifications, 3.37% misclassification rate
- Digit 4: 32 misclassifications, 3.26% misclassification rate
- Digit 5: 33 misclassifications, 3.70% misclassification rate
- Digit 6: 14 misclassifications, 1.46% misclassification rate
- Digit 7: 37 misclassifications, 3.60% misclassification rate
- Digit 8: 60 misclassifications, 6.16% misclassification rate
- Digit 9: 41 misclassifications, 4.06% misclassification rate

Top 3 Most Frequently Misclassified Digits:

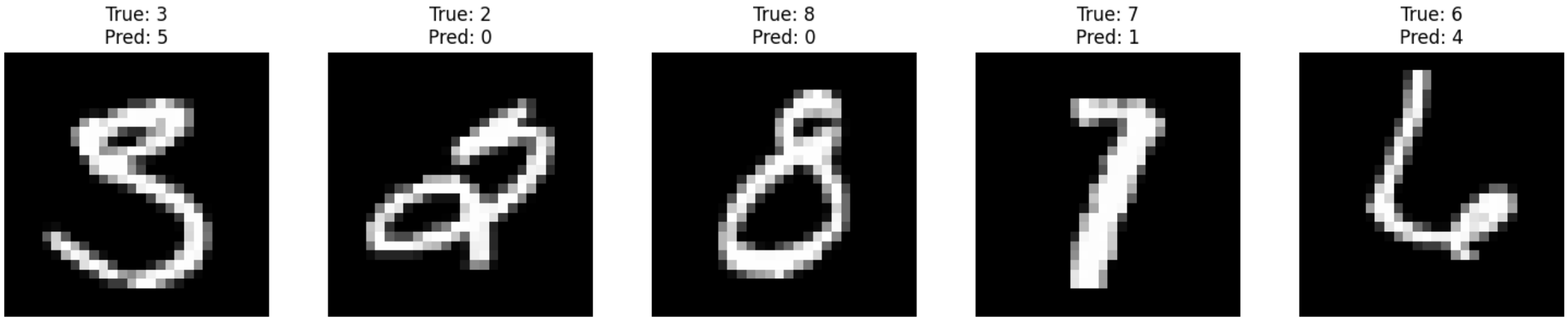
- Digit 8: 60 misclassifications, Rate: 6.16%
- Digit 9: 41 misclassifications, Rate: 4.06%
- Digit 7: 37 misclassifications, Rate: 3.60%



3 k-NN Classifier

3.2 Final Model Training and Evaluation

5 Random Misclassified Examples:



4 Decision Tree Classifier

4.1 Model Training and Hyperparameter Tuning

- Reshapes the training and validation sets into 2D arrays so that the Decision Tree classifier can process the data. Defines the grid of hyperparameter values for max_depth and min_samples_split to explore during tuning.
- Creates a dictionary to store the validation accuracy for each hyperparameter combination and initializes variables to track the best configuration.
- Iterates over every combination of max_depth and min_samples_split, training a Decision Tree for each setting. For each configuration, initializes the classifier with the current hyperparameters, fits it on the flattened training data, and predicts on the validation set.
- Computes the validation accuracy for the current configuration and stores the result, printing the configuration and its accuracy.
- Updates the best hyperparameters if the current configuration yields a higher validation accuracy. Prints out the best-performing hyperparameter combination along with its corresponding accuracy.

4 Decision Tree Classifier

4.1 Model Training and Hyperparameter Tuning

Max Depth: 2, Min Samples Split: 2 => Validation Accuracy: 0.1872
Max Depth: 2, Min Samples Split: 5 => Validation Accuracy: 0.1872
Max Depth: 5, Min Samples Split: 2 => Validation Accuracy: 0.2855
Max Depth: 5, Min Samples Split: 5 => Validation Accuracy: 0.2855
Max Depth: 10, Min Samples Split: 2 => Validation Accuracy: 0.4176
Max Depth: 10, Min Samples Split: 5 => Validation Accuracy: 0.4177

Best Hyperparameters:
Max Depth: 10, Min Samples Split: 5, Accuracy: 0.4177

4 Decision Tree Classifier

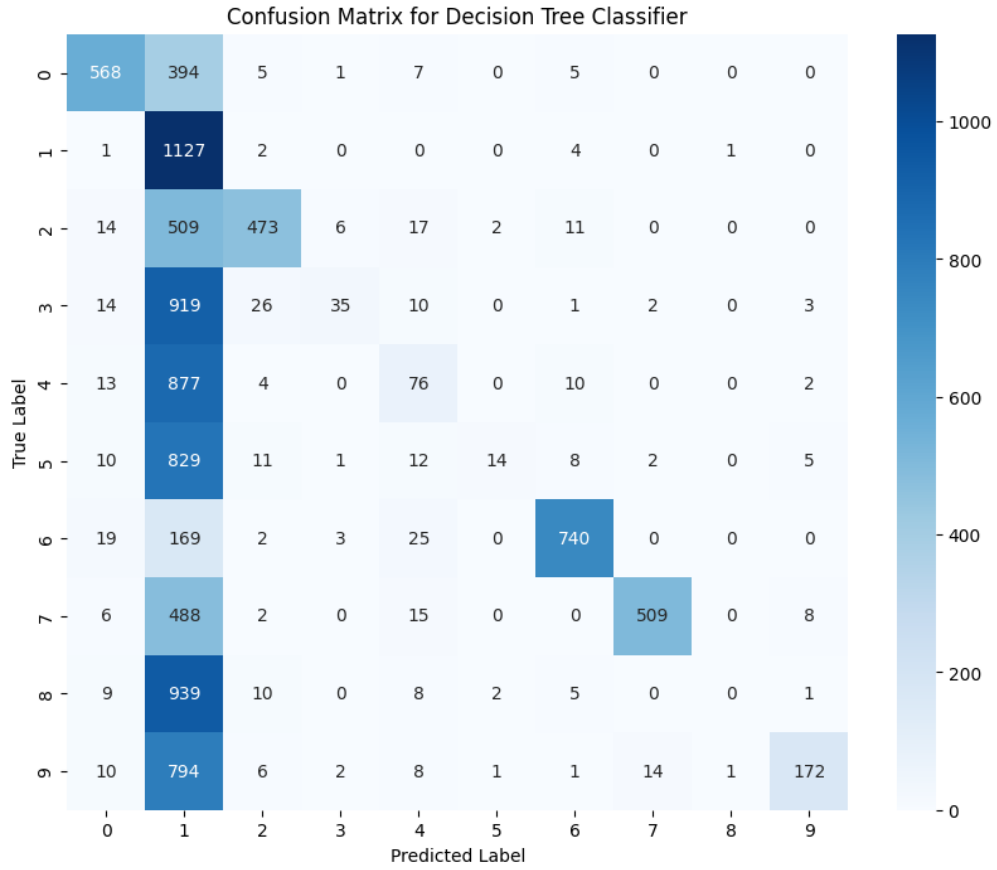
4.2 Evaluation

- Combines the training and validation sets and reshapes the images into 2D arrays so the classifier can process them.
- Initializes the final Decision Tree classifier with the best hyperparameters (max_depth=10, min_samples_split=5) and trains it on the combined dataset.
- Predicts the labels on the test set and computes evaluation metrics including accuracy, precision, recall, and F1-score, then prints a detailed classification report.
- Generates and visualizes a confusion matrix using a heatmap to analyze which classes are most often misclassified.
- Provides a discussion prompt to examine the confusion matrix for patterns in misclassification (e.g., identifying digits that are frequently confused).
- Computes predicted probabilities for each class, binarizes the test labels for a multi-class ROC analysis, calculates ROC curves and AUC scores for each digit, and plots all ROC curves on a single graph with AUC values included in the legend.

4 Decision Tree Classifier

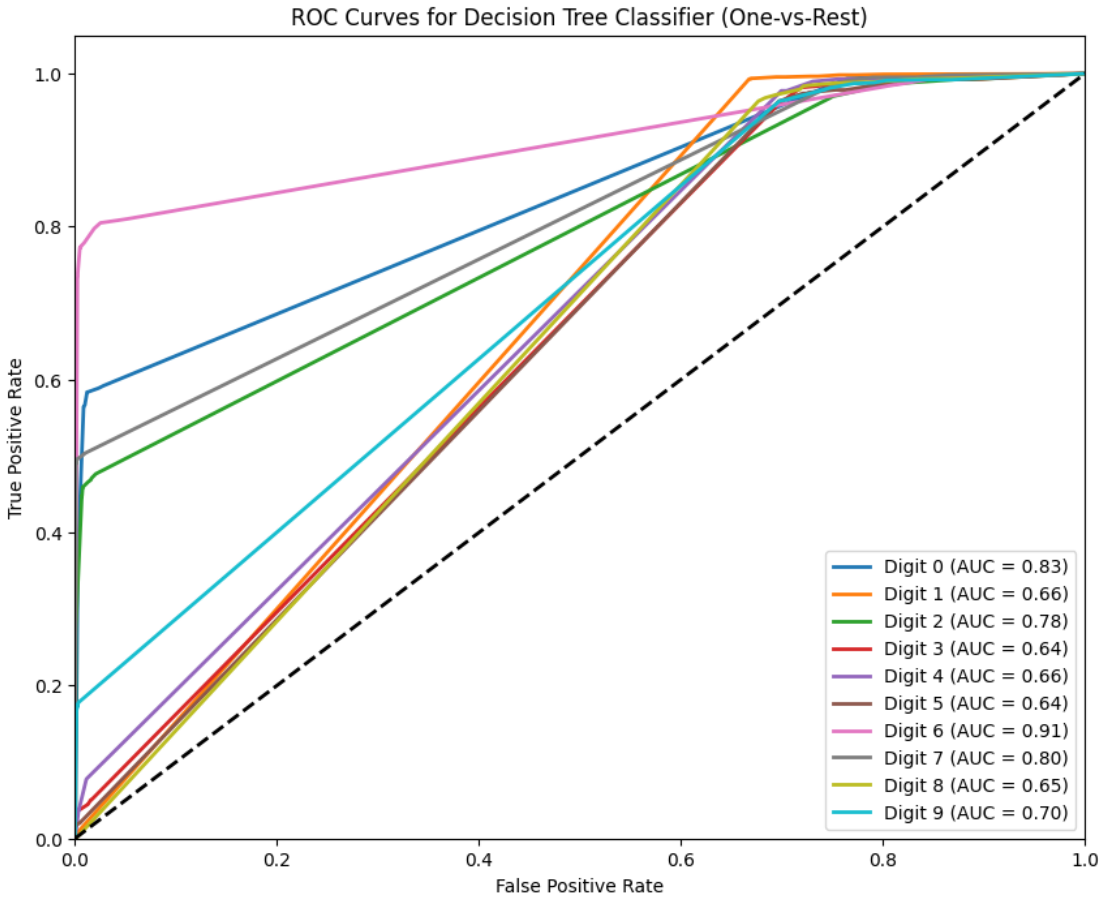
4.2 Evaluation

```
Test Accuracy:  0.3714
Test Precision: 0.6540
Test Recall:    0.3714
Test F1 Score:  0.3609

Detailed Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.58      0.69       980
           1       0.16      0.99      0.28      1135
           2       0.87      0.46      0.60      1032
           3       0.73      0.03      0.07      1010
           4       0.43      0.08      0.13       982
           5       0.74      0.02      0.03       892
           6       0.94      0.77      0.85       958
           7       0.97      0.50      0.65      1028
           8       0.00      0.00      0.00       974
           9       0.90      0.17      0.29      1009

    accuracy                           0.37     10000
   macro avg       0.66      0.36      0.36     10000
weighted avg       0.65      0.37      0.36     10000
```



4 Decision Tree Classifier

4.2 Evaluation



4 Decision Tree Classifier

4.2 Evaluation

Decision Tree Confusions:

- Almost all numbers other than 6 are confused for number 1.
- 0, 1, 6 and 7 are the only numbers that has the top predictions in their own category.
- Number 3, 4, 5, 8 and 9 all were predicted as mostly 1.