# CS412: Machine Learning Homework 1

February 16, 2025

## Submission Guidelines

- **Jupyter Notebook:** Include all code cells and output. **(Ensure that all outputs remain in the notebook, as it will not be re-run during grading.)**
- **PDF Report:** Prepare a comprehensive PDF report summarizing your methodology, analysis, and results.
- **Notebook Link:** At the top of your PDF report, include the shareable link to your notebook (accessible via the Share button on the top right). **Place this link at the beginning of your report. The link should be set so that anyone with the link can access it.**
- **Submission Files:** Submit the following:
  - Your Jupyter Notebook as **CS412-HW1-YourName.ipynb** - *(e.g. CS412-HW1-MertPekey.ipynb)*
  - Your PDF report as **CS412-HW1-YourName.pdf** - *(e.g. CS412-HW1-MertPekey.pdf)*
- **Late Submissions:** Late submissions will be accepted with a penalty of **10 points per day**.

## 1 Overview

In this assignment, you will implement and evaluate two classifiers on the MNIST dataset:

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

This homework is designed to provide hands-on experience in data preprocessing, hyperparameter tuning, model evaluation, and reporting of results. All work must be documented in a Jupyter Notebook and the Report.

Figure 1: MNIST Dataset

# 2 Dataset and Preprocessing

The MNIST dataset contains $28 \times 28$ grayscale images of handwritten digits (0-9), where each pixel value ranges from 0 to 255.

## 2.1 Data Loading

1. Load the MNIST dataset using the Keras API (Keras MNIST Dataset).
2. Split the data as follows:
   - **Training Set:** Use 80% of the provided training data.
   - **Validation Set:** Use the remaining 20% of the training data.
   - **Test Set:** Use the given test set without modifications.
3. Print the shapes of your training, validation, and test sets to verify that the splits are correct.

## 2.2 Data Analysis

Before preprocessing, perform the following analysis:

1. **Class Distribution:** Compute and display the number of samples per digit to check for imbalances.
2. **Basic Statistics:** Calculate the mean and standard deviation of the pixel values.
3. **Visualization:** Create subplots showing at least one sample image for each digit.

## 2.3 Data Preprocessing

1. Normalize the images so that pixel values are scaled to the range [0, 1].
2. Ensure that all preprocessing steps are clearly explained.

# 3 k-NN Classifier

## 3.1 Model Initialization and Hyperparameter Tuning

1. Initialize a k-NN classifier.
2. Experiment with different numbers of neighbors: **1, 3, 5, 7, and 9**.
3. Use the validation set to determine the optimal value based on accuracy.
4. Plot the the number of neighbors and validation accuracy. Be sure to label your axes.

## 3.2 Final Model Training and Evaluation

1. Retrain the k-NN classifier using the combination of the training and validation sets with the best hyperparameter.
2. Evaluate the final model on the test set by reporting Accuracy, Precision, Recall, and F1-score
3. Generate and visualize a confusion matrix.
4. Discuss which digits are most frequently misclassified.
5. Display 5 random misclassified examples in a subplot.

# 4 Decision Tree Classifier

## 4.1 Model Training and Hyperparameter Tuning

1. Train a Decision Tree classifier on the MNIST dataset.
2. Tune the following hyperparameters:
   - **max_depth:** Try values `[2, 5, 10]`.
   - **min_samples_split:** Try values `[2, 5]`.
3. Document the results for each configuration and explain how you selected the best-performing model.

## 4.2 Evaluation

1. Evaluate the final model on the test set by reporting Accuracy, Precision, Recall, and F1-score
2. Generate a confusion matrix and provide an analysis of the results. In your discussion, highlight any patterns in misclassifications, such as which digits are most frequently confused with one another.
3. Plot the ROC curve for each digit on a single plot. Include the AUC score for each digit in the legend.

# 5 Final Report Guidelines

Your final PDF report should include:

- A clear overview of your methodology, including data analysis and preprocessing steps.
- Detailed explanations and justifications for your model choices and hyperparameter tuning.
- Comprehensive results supported by tables, plots, and visualizations.
- An analysis of misclassifications with potential explanations.
- The shareable link to your Jupyter Notebook at the top of the document.