**Mete Kerem Berk**

**CS412 Machine Learning Homework 1**

**February 22, 2025**

Notebook Link: <https://colab.research.google.com/drive/1rFPa8QgWvX14T9wMAs9KTQ1evAB0Qe7P?usp=sharing>

1. **Overview**

In this assignment, k-Nearest Neighbors (k-NN) and Decision Tree classifiers are implemented on the MNIST dataset. The MNIST dataset contains 28×28 grayscale images of handwritten digits (0-9), where each pixel value ranges from 0 to 255.

1. **Dataset and Processing**
   1. **Data Loading**

First, the MNIST dataset is loaded using the Keras API. The dataset is loaded as training and test sets. The initial training set is split between training and validation set 80/20 respectively and the test set is left as is.



Figure 1 - Data Loading Code

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure 2 - Data Loading Output

The initial training set is split between the training set and the validation set to help in hyperparameter training and prevent overfitting. The validation set acts as a checkpoint to detect problems before testing.

* 1. **Data Analysis**

After loading the dataset, data analysis is done to identify dataset characteristics and necessary preprocessing steps.

First, class distribution is controlled. The number of samples per digit is computed and displayed to check for imbalances.



Figure 3 - Class Distribution Code

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure 4 - Class Distribution

From Figure 4, it is seen that the samples are balanced in the dataset. This shows that preprocessing techniques like resampling or class weighting are not necessary.

Then, the mean and the standard deviation of the pixel values are computed. The mean helps understand the general brightness level of images and the standard deviation shows whether the images have high contrast or are mostly uniform in brightness.



Figure 5 - Mean & Standard Deviation Code



Figure 6 - Mean & Standard Deviation

The mean and standard deviation values are moderate on a 0-255 scale. These values will need to be normalized (0-1 scale) in the preprocessing steps.

Finally, random samples from each digit are displayed to confirm data quality and structure. It also helps detect noise, mislabeling and unexpected artifacts.

