Report: Principal Component Analysis (PCA) on the MNIST Dataset

Date: September 19, 2025
Author: Krrish Raj
Analysis Objective: To perform dimensionality reduction on the MNIST dataset of handwritten digits using Principal Component Analysis (PCA), analyze the data structure, and visualize the results to determine the separability of digit classes in a lower-dimensional space.

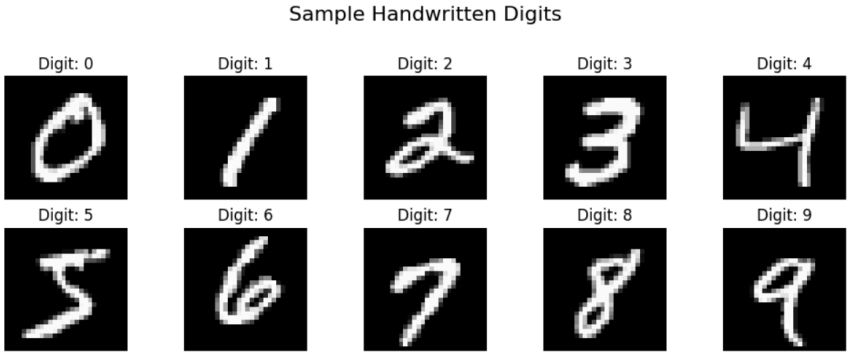
1. Introduction

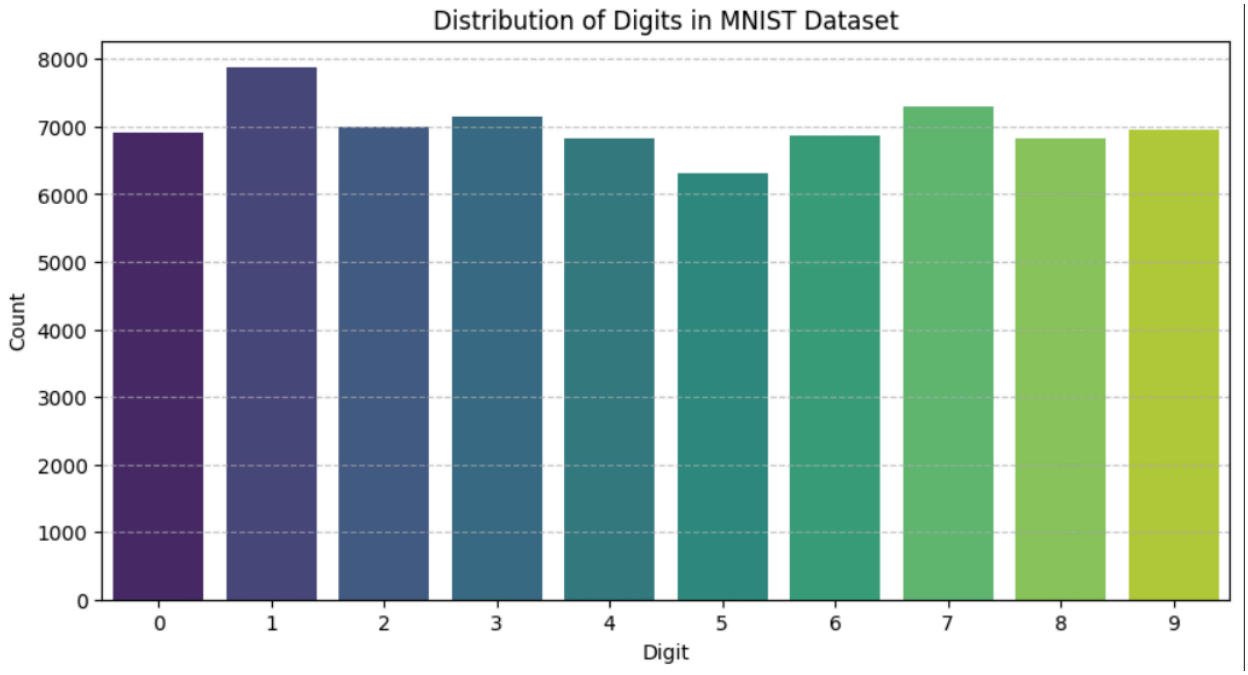
The MNIST dataset is a benchmark collection of 70,000 grayscale images of handwritten digits (0-9), each sized at 28x28 pixels. When flattened, each image represents a data point in a 784-dimensional space. Processing and modeling data with such high dimensionality can be computationally expensive and prone to the "curse of dimensionality."

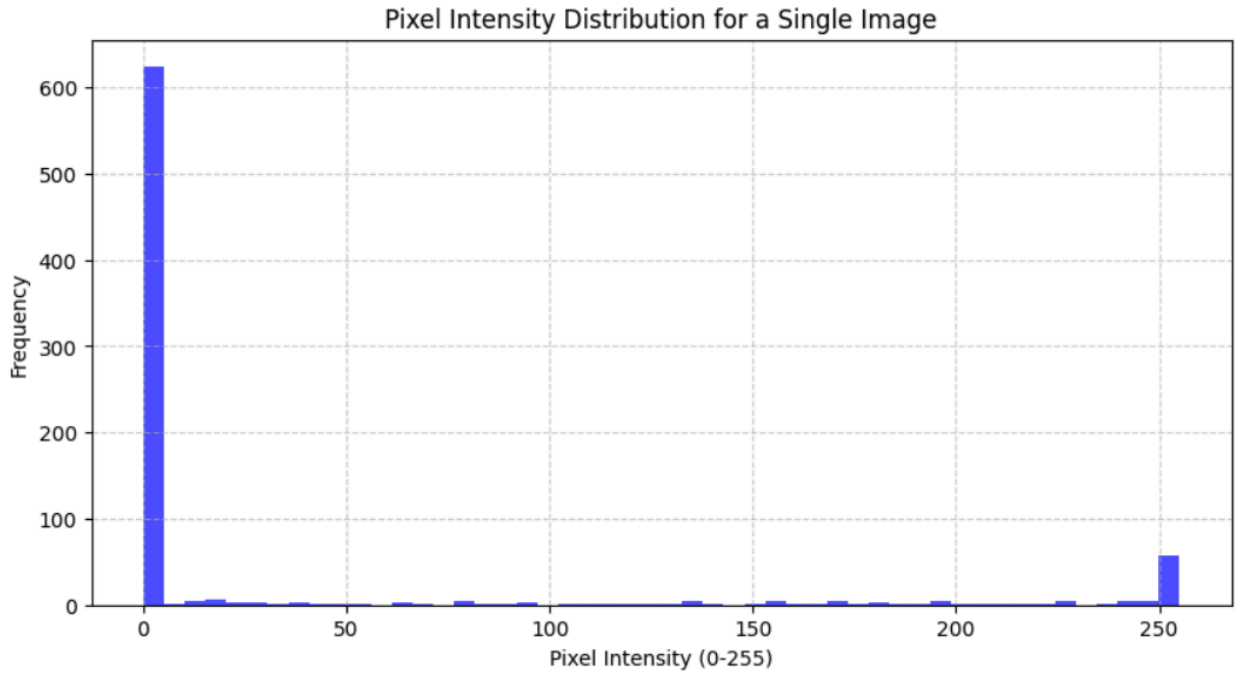
This report details the application of Principal Component Analysis (PCA), a powerful unsupervised learning technique, to reduce the dimensionality of the MNIST dataset. The primary goals were to explore the dataset's intrinsic structure, determine the number of principal components required to retain significant variance, and visualize the data in two dimensions to observe the clustering of different digits.

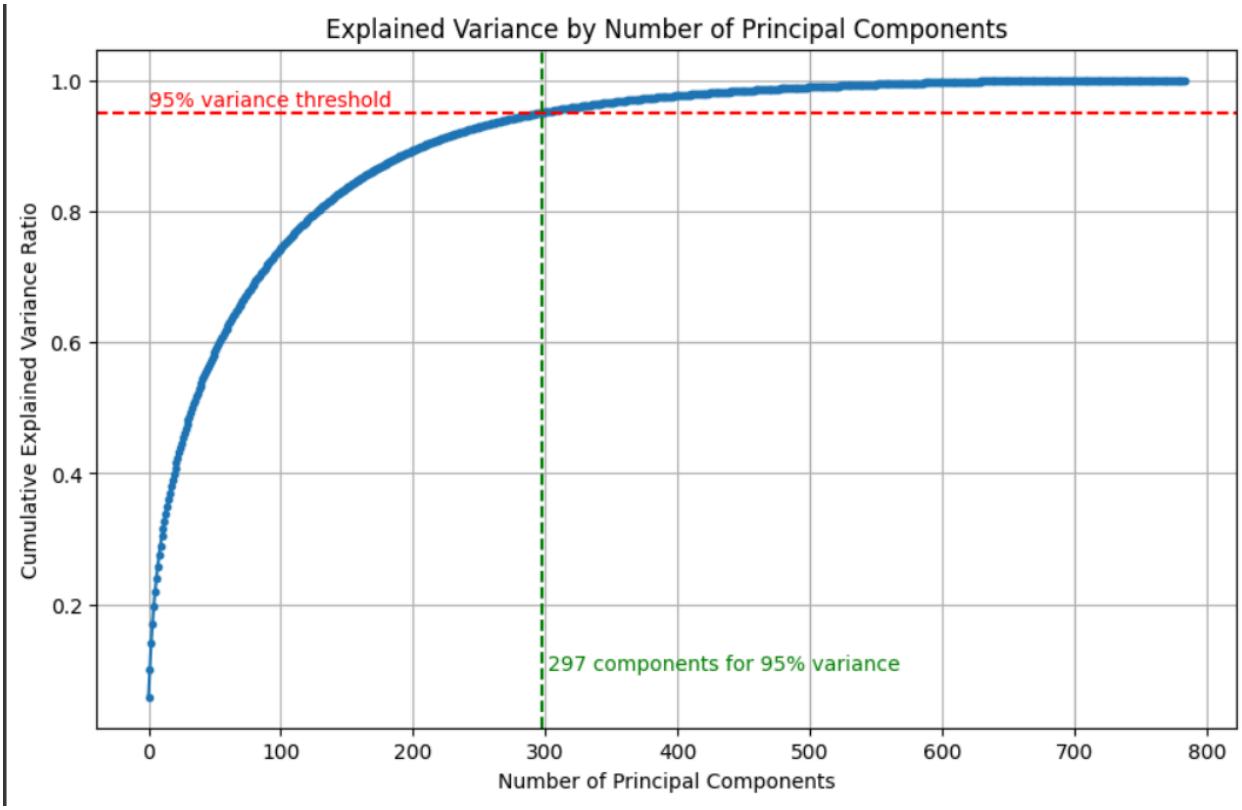
2. Exploratory Data Analysis (EDA)

Before applying PCA, an initial exploratory analysis was conducted on a sample of 15,000 images to understand its fundamental properties.

- **Data Structure:** The dataset consists of 784 features (pixel values from 0 to 255) for each image and a corresponding label indicating the digit.
- **Class Distribution:** An analysis of the digit labels revealed a well-balanced distribution, with each of the 10 digits (0-9) having a roughly equal number of samples. This ensures that any subsequent modeling will not be biased by an over-representation of a particular class.
- **Pixel Intensity:** A histogram of pixel values for a sample image showed that a vast majority of pixels have an intensity of 0 (black), corresponding to the background. A smaller number of pixels have higher intensities, representing the handwritten strokes of the digit.
- 


Distribution of Digits in MNIST Dataset
Count
Digit
0
1000
2000
3000
4000
5000
6000
7000
8000
0
1
2
3
4
5
6
7
8
9


Pixel Intensity Distribution for a Single Image
Frequency
Pixel Intensity (0-255)
0
100
200
300
400
500
600
0
50
100
150
200
250



3. Data Preprocessing

To ensure the effectiveness of PCA, the following preprocessing steps were performed:

1. **Subsampling:** A random sample of 15,000 images was selected from the full dataset. This was done to make the computational steps, particularly the fitting of the PCA model, more efficient for this analysis.
2. **Feature Scaling:** PCA is highly sensitive to the variance of the features. Since pixel intensities are on a scale of 0-255, features with higher variance could disproportionately influence the principal components. To prevent this, the StandardScaler was used to transform the data, giving each pixel feature a mean of 0 and a standard deviation of 1.

4. Application of Principal Component Analysis (PCA)

PCA was applied in two stages: first, to determine the optimal number of components, and second, to perform the final dimensionality reduction for visualization.

- **Explained Variance Analysis:** A PCA model was fitted to the scaled data to calculate the cumulative explained variance. This analysis shows how much of the original information (variance) is retained as the number of principal components increases.
  The analysis revealed that **154 principal components are required to capture 95% of the total variance** in the data. This signifies a substantial reduction from the original 784 dimensions while preserving most of the dataset's structural information.
- **Dimensionality Reduction for Visualization:** For the purpose of visualization, a separate PCA was performed to reduce the data to just **two principal components**. While these two components only explain approximately **17.29%** of the total variance, they are the two single dimensions that capture the most information, making them ideal

for a 2D scatter plot.

5. Results and Visualization

The 784-dimensional data was projected onto the two principal components and visualized.

The resulting scatter plot shows the distribution of the digits in the new 2D space. Key observations include:

- **Clear Clustering:** Despite the significant loss of information, distinct clusters for different digits are clearly visible. Digits that are visually dissimilar, such as '0' and '1', form tight, well-separated clusters far from each other.
- **Cluster Overlap:** Digits with similar stroke patterns, such as '4' and '9' or '3' and '5', show some overlap in their clusters. This is expected, as their underlying patterns are more alike.
- **Structural Preservation:** The visualization confirms that PCA successfully captured the most significant variance, preserving the underlying structure that separates the digit classes.

6. Analysis of Principal Components

The principal components themselves can be visualized by reshaping them back into 28x28 images. These images are not digits but rather represent the fundamental patterns, or "archetypes," that PCA identified as being most important for describing the data.

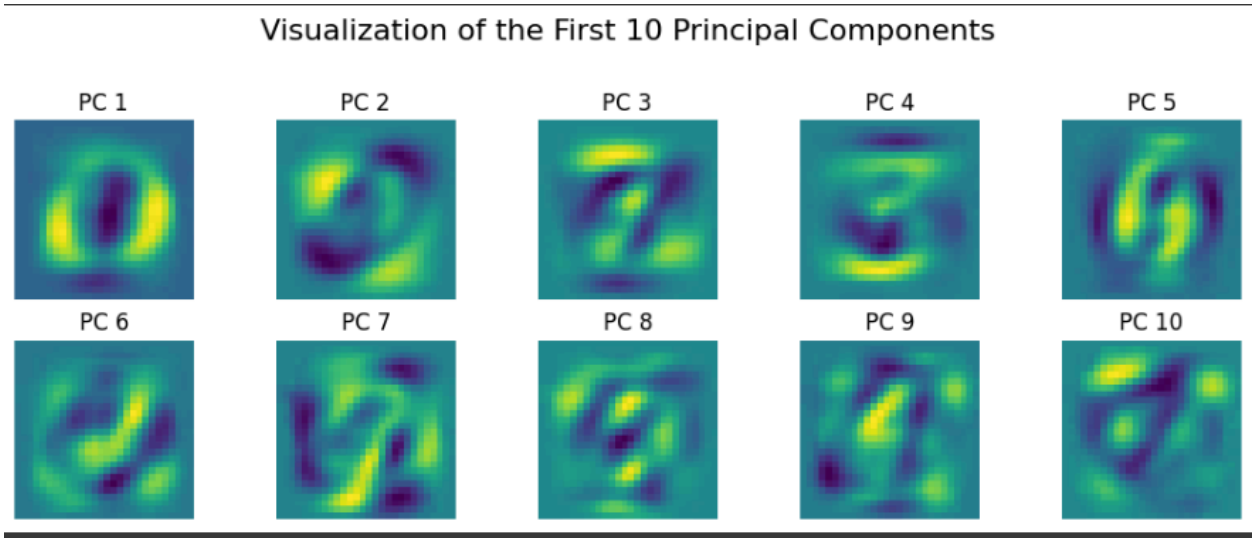
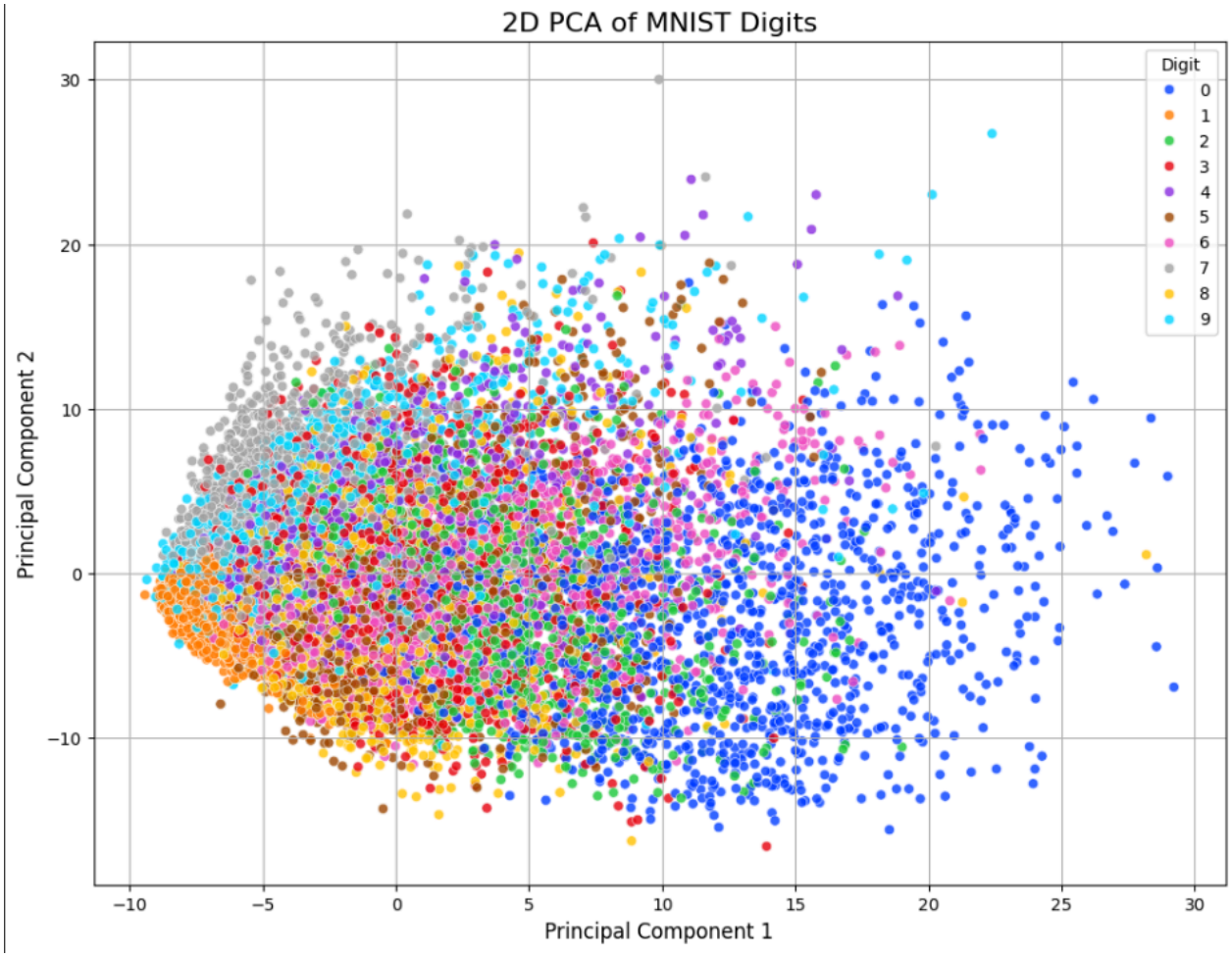
The first few components often represent broad features like strokes, loops, and lines that are common across many digits. For example, the first principal component might capture the general shape of a loop (common in 6, 8, 9, 0), while another might capture a vertical line (common in 1, 4, 7).

7. Conclusion

This analysis successfully demonstrated the effectiveness of Principal Component Analysis for the dimensionality reduction and visualization of the high-dimensional MNIST dataset.

The key findings are:

- The dimensionality of the dataset can be reduced from **784 to 154 components** while retaining 95% of the original variance.
- Reducing the data to only **two principal components** is sufficient to visualize distinct clusters of different digits, confirming the inherent separability of the classes.
- PCA is a valuable tool not only for reducing computational complexity but also for gaining insights into the fundamental structure of image data.

- 
Visualization of the First 10 Principal Components
PC 1
PC 2
PC 3
PC 4
PC 5
PC 6
PC 7
PC 8
PC 9
PC 10
- 
2D PCA of MNIST Digits
Digit
0
1
2
3
4
5
6
7
8
9
Principal Component 2
Principal Component 1
30
20
10
0
−10
−10
−5
0
5
10
15
20
25
30