**IMAGE COMPRESSION USING K-MEANS CLUSTERING**

**ON MNIST DATASET**

**1. Introduction** This project applies K-Means clustering to the MNIST dataset to perform image compression. The goal is to reduce the amount of data needed to represent each image while maintaining the essential features necessary for recognition.

**2. Dataset Description** The dataset used is the MNIST dataset, a collection of 28x28 grayscale images of handwritten digits (0-9). It consists of 60,000 training images and 10,000 test images.

In this experiment:

* The dataset was normalized to scale pixel values between 0 and 1.
* Each image was flattened from a 28x28 matrix into a 784-dimensional vector.
* A reduced sample of 2,000 training images and 500 test images was used to optimize computational efficiency.

**3. Methodology**

* **Clustering with K-Means:** The dataset was clustered using K-Means with 16 clusters, which helps in representing the image pixels with fewer values.
* **Compression Approach:** After clustering, each pixel value in an image was replaced with the centroid of its respective cluster.
* **Reconstruction:** The compressed images were reconstructed by mapping each pixel value to its corresponding cluster center, reshaped back to the original 28x28 format.

**4. Results & Observations**

* The compressed images show slight blurring and loss of fine details due to the limited number of clusters.
* The original images retained their recognizable structures after compression, demonstrating the effectiveness of K-Means in preserving key features.
* Increasing the number of clusters would improve reconstruction quality but at the cost of higher storage requirements.

**5. Advantages of the K-Means Model for Image Compression**

* **Data Reduction:** The model significantly reduces the number of unique pixel values by clustering similar pixel intensities together.
* **Faster Processing:** With fewer unique pixel values, storing and transmitting images becomes more efficient.
* **Feature Preservation:** Despite compression, the essential shape of digits remains intact, allowing potential applications in handwritten digit recognition.
* **Computational Efficiency:** K-Means clustering is relatively simple and computationally efficient, making it feasible for real-time applications.

**6. Limitations and Future Improvements**

* **Blurring Effect:** The limited number of clusters introduces blurring, which could be mitigated by increasing clusters.
* **Alternative Approaches:** More advanced techniques like PCA, Autoencoders, or Deep Learning-based compression could provide better results.
* **Application to Other Datasets:** The same approach could be tested on more complex datasets to evaluate its effectiveness beyond MNIST.

**7. Conclusion** K-Means clustering provides a simple yet effective method for image compression. While it introduces some loss in image quality, it successfully reduces storage requirements while maintaining recognizable digit structures. Future work can involve optimizing the number of clusters or integrating this approach with deep learning techniques for improved performance.