### Report on MNIST Dataset Analysis **Introduction**

This report explores the methodologies and outcomes of compressing numerical data efficiently. Focusing on the comparison between original and compressed digits, it delves into various techniques, their implementation, and the impact on data integrity and storage optimization. The goal is to provide a clear understanding of how data compression can be utilized to maximize storage space while maintaining essential data quality.

### A) Loading the MNIST Dataset

In this section, we retrieve and load the MNIST dataset consisting of 70,000 instances. The dataset is stored in .arff format and contains pixel values for handwritten digits. Each digit is represented as a 28x28 pixel image, resulting in a total of 784 features per instance.

After loading the dataset, we display the first 100 digits for visualization.



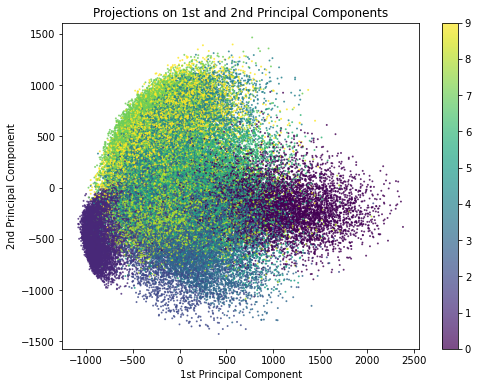
### B) Principal Component Analysis (PCA)

PCA is applied to reduce the dimensionality of the dataset while retaining the most variance. The first and second principal components are extracted, and their explained variance ratios are computed.

Explained variance ratio for 1st and 2nd principal components: [0.09746116, 0.07155445]

### C) Projection onto Principal Components

The projections of the 1st and 2nd principal components onto a 1D hyperplane are plotted. This visualization helps in understanding the distribution of the data in the new feature space.



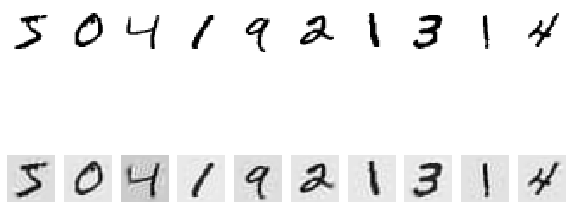
### D) Incremental PCA

Incremental PCA is utilized to further reduce the dimensionality of the MNIST dataset to 154 dimensions. This approach is useful for handling large datasets by processing them in smaller batches.

The transformation retains the essential characteristics of the original dataset while reducing the computational complexity.

### E) Original and Compressed Digits

The original (1st image) and compressed (2nd image) digits are displayed side by side to visually assess the effectiveness of the dimensionality reduction. The reconstruction of the digits from the reduced representation demonstrates how well the important features are preserved.

  
  
It appears that data is shown in lower image quality than the original data set, however the values are represented correctly.

### **Conclusion**

Based on the results gathered, we can conclude that PCA is an effective algorithm for compressing and reducing dimensionality of the MNIST dataset.