Classification and Clustering of Handwritten Digits using Machine Learning Techniques

Garrett Fincke*

Abstract. This project uses machine learning algorithms for classifying and clustering handwritten digits using the MNIST dataset. In this project, I implemented K-Nearest Neighbors (KNN), Logistic Regression (LR), and Support Vector Machines (SVM) to classify digits and evaluated the models performance through various metrics such as accuracy, precision, recall, and F1-score. SVM achieved the highest accuracy at 95.3. Additionally, K-Means clustering was applied to group digit images, with Principal Component Analysis (PCA) for visualization.

1. Introduction.

Handwritten digit classification has significant applications in computer vision. This project involves classifying digits using methods such as KNN, LR, and SVM, and grouping them with K-Means clustering to explore unsupervised learning on MNIST data. The experiments focus on optimizing model performance through hyperparameter tuning and evaluating clustering efficacy through visualization.

2. Related Work.

Handwritten digit recognition has been extensively studied, with numerous algorithms proposed over the years. The MNIST dataset was first created by LeCun et al. [1], and they tested it via the use of Convolutional Neural Networks (CNNs) for digit recognition. Later, Support Vector Machines have been applied successfully, as demonstrated by Cortes and Vapnik [2]. While deep learning approaches now dominate the field, algorithms like KNN, LR, and SVM are relevant because of lower computational requirements. Thus, my approach differs by focusing on these algorithms' performance after hyperparameter tuning and by analyzing the effectiveness of K-Means Clustering on the MNIST dataset.

3. Data.

The MNIST dataset contains 70,000 images of digits (28x28 pixels), with 60,000 for training and 10,000 for testing.

3.1. Data Preprocessing.

- Random Sampling: Randomly selected 10,000 images from the original training set to create the working dataset (computer couldn't handle much more)
- Normalization: Pixel values were scaled to [0, 1].
- Flattening: Images were flattened into 784-dimensional vectors.
- Train-Test Split: An 80/20 split as required

4. Methods.

4.1. Classification Algorithms.

Three classification algorithms were implemented:

- **KNN**: K=3 yielded an accuracy of 94.4.
- **Logistic Regression**: Achieved an accuracy of 91.15.
- **SVM**: Produced the highest accuracy at 95.3 with the RBF kernel and regularization parameter tuning.

*Penn State University, ggf5038@psu.edu

4.2. Hyperparameter Tuning.

- **KNN**: k values 1-10 were tested, with k=3 providing optimal results.
- **SVM**: Cross-validated with RBF and linear kernels. RBF outperformed with a regularization parameter of C=1.

4.3. Clustering Algorithm.

- **K-Means Clustering**: K-Means was applied to identify clusters within the MNIST dataset. Although the Elbow Method and Silhouette Scores did not reveal a definitive optimal k, $k = 10$ was selected based on the known number of digit classes. PCA was used to reduce data dimensionality for visualization, illustrating general clustering patterns with some overlap in handwritten digits.

5. Experiments.

Model	Accuracy	Precision	Recall	F1 Score
KNN	0.9440	0.9453	0.9440	0.9439
Logistic Regression	0.9115	0.9113	0.9115	0.9111
SVM	0.9530	0.9529	0.9530	0.9528

Table 1
Performance metrics for classification models

5.1. Classification Performance Evaluation.

5.2. Clustering Analysis.

- **Optimal k Determination**: The Elbow Method and Silhouette Score provided no sharp indication of an optimal k, though $k = 10$ was chosen based on prior knowledge of MNIST's 10 digit classes. The gradual decrease in inertia and fluctuating silhouette scores reflect the complexity and overlapping nature of MNIST data.
- **Visualization**: PCA-reduced clusters and visualizations of cluster centers revealed general grouping, with some digits, such as '0' and '1', forming relatively distinct clusters. However, overlap between certain clusters indicates variability in handwritten digits, limiting perfect separation.

Confusion matrices are provided below to illustrate model strengths and weaknesses.

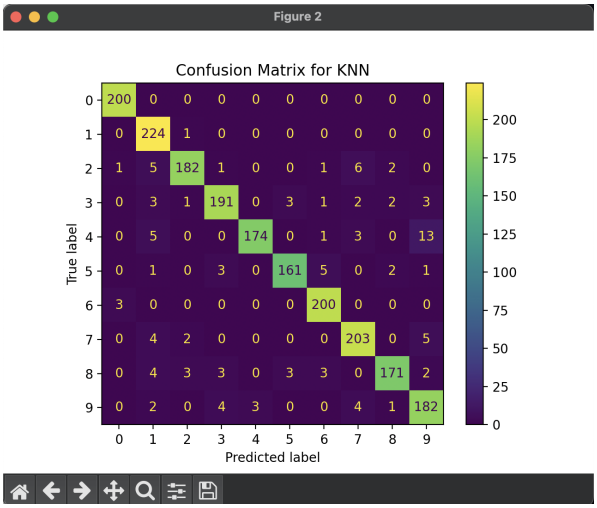

Figure 1. *Confusion Matrix for KNN*

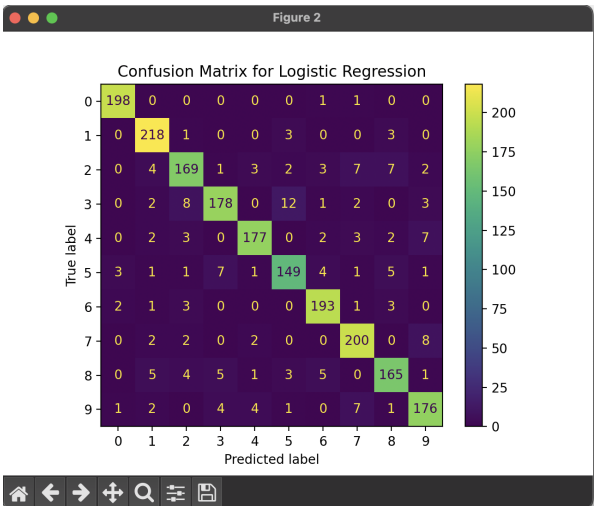

Figure 2. *Confusion Matrix for Logistic Regression*

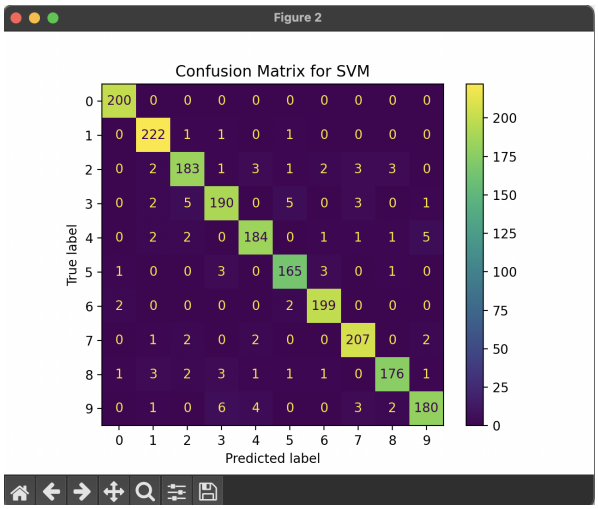

Figure 3. *Confusion Matrix for SVM*

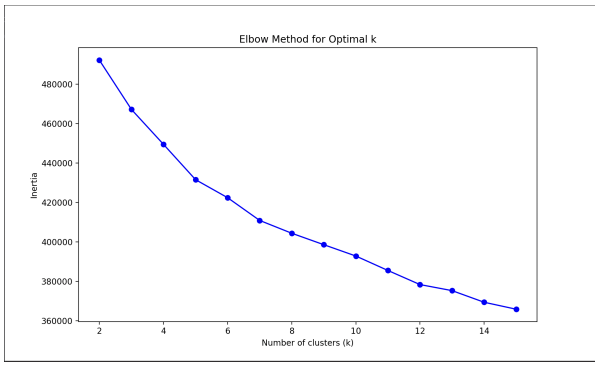

Figure 4. *Elbow Method for Optimal k*

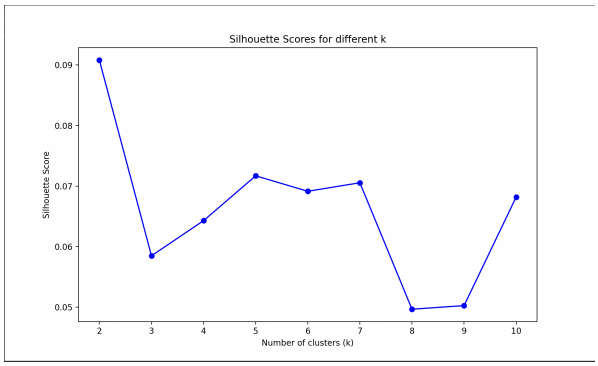

Figure 5. *Silhouette Scores for different k values*

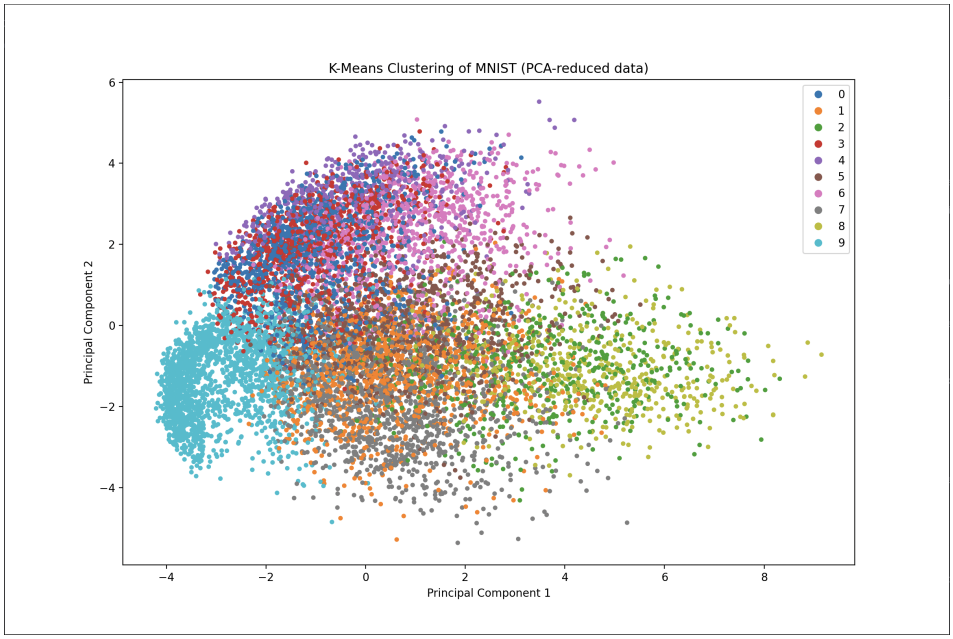

Figure 6. *K-Means Clustering of MNIST (PCA-reduced data)*

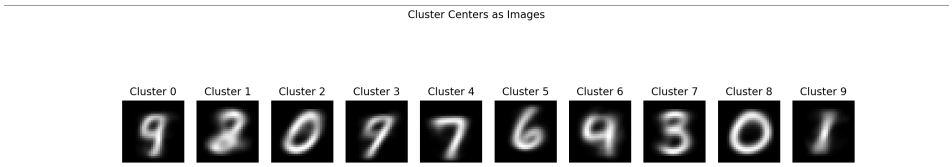

Figure 7. *Cluster Centers as Images*

6. Conclusion. In this study, I implemented and compared KNN, Logistic Regression, and SVM classifiers on the MNIST dataset, achieving maximum accuracies of 94.4, 91.15, and 95.3, respectively. The SVM classifier, utilizing an RBF kernel with optimized hyperparameters, outperformed the other models, which showed its effectiveness for this task. My clustering analysis using K-Means and PCA revealed inherent challenges due to the overlapping nature of handwritten digits, limiting perfect separation.

However, I found limitations from the computational intensity of training SVMs on high-dimensional data, and the difficulty in selecting an optimal number of clusters for K-Means. Future work could involve implementing deep learning approaches like Convolutional Neural Networks (CNNs). LeCun et al. [1] achieved 99 accuracy using deep neural networks. Incorporating such techniques would likely offer improved accuracy and better handling of the complex patterns inherent in handwritten digit data.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.