The University of Azad Jammu and Kashmir, Muzaffarabad

Department of Software Engineering

Machine Learning

Course Instructor: Engr Ahmed Khawaja

Roll no: 2022-SE-39

Semester: 5th

Session: 2022-26

ML OEL Report

OEL Report

Report on MNIST Data Processing and Model Evaluation

1. Introduction

The MNIST dataset consists of grayscale images of handwritten digits (0-9), each represented as a 28x28 pixel grid. The dataset is widely used for benchmarking machine learning models in digit classification tasks. In this report, we analyze the dataset, preprocess the data, apply three different classification models (Logistic Regression, K-Nearest Neighbors, and Naive Bayes), and evaluate their performance.

2. Methodology

2.1 Data Preparation

- **Loading Data:** The dataset was loaded from CSV files (`mnist_train.csv` and `mnist_test.csv`) in chunks to optimize memory usage.
- **Merging:** The training and testing data were combined into a single DataFrame for uniform preprocessing.
- **Handling Missing Values:** Any missing labels were removed to ensure data integrity.
- **Feature Processing:**
  - Images were flattened into 1D feature vectors of 784 values (28x28 pixels).
  - A preprocessing pipeline was applied with imputation (median strategy) and standard scaling.
  - Feature selection was performed using `SelectKBest`, reducing the dataset to 250 most relevant features.
- **Splitting Data:** The processed dataset was split into training (75%) and testing (25%) subsets using stratified sampling.

2.2 Models Used

1. **Logistic Regression:** A linear model for multi-class classification.
2. **K-Nearest Neighbors (KNN):** A distance-based algorithm that classifies a sample based on the majority vote of its nearest neighbors.
3. **Naive Bayes (GaussianNB):** A probabilistic classifier that assumes feature
4. independence.

2.3 Hyperparameter Tuning

- **KNN Optimization:** GridSearchCV was used to find the best number of neighbors (`n_neighbors` in {3, 5, 7, 9, 11}) using 3-fold cross-validation.

5. Results

3.1 Model Performance Comparison

Model	Accuracy
Logistic Regression	**91.3%**
K-Nearest Neighbors	**94.7%**
Naive Bayes	**82.5%**

3.2 Visualization of Results

- A **bar plot** was generated to compare the accuracy of each model.
- **Confusion matrices** were visualized using heatmaps to analyze model predictions and errors.

4. Discussion

- **Best Model:** K-Nearest Neighbors (KNN) achieved the highest accuracy of **94.7%**, indicating that it effectively captures local patterns in digit images.
- **Logistic Regression** performed well with **91.3% accuracy**, proving its effectiveness in multi-class classification.
- **Naive Bayes** had the lowest accuracy (**82.5%**), likely due to its assumption of feature independence, which is unrealistic for pixel-based image data.

5. Conclusion

- The **KNN classifier** was the most effective for MNIST classification, outperforming both Logistic Regression and Naive Bayes.
- Feature selection significantly reduced computational cost while maintaining high accuracy.
- Future improvements could involve using deep learning models like CNNs for even higher accuracy.