**Methods and tools for data analysis**

Report

Assignment 3

Student : Akzhan Mataev

Group : CSE-2404M

Instructor : Dr.Adai Shomanov

# Comparative Analysis of Supervised Learning Algorithms on the MNIST Dataset

## 1. Introduction

The goal of this study is to compare three supervised learning algorithms—Logistic Regression, Decision Tree, and Random Forest—on the MNIST dataset. MNIST consists of grayscale images of handwritten digits (0-9) and is widely used as a benchmark dataset for classification tasks. The objective is to evaluate the performance of these models based on classification metrics and identify the most effective approach for digit recognition.

## 2. Data Exploration and Preprocessing

The MNIST dataset consists of 60,000 training samples and 10,000 test samples, with each image having dimensions of 28x28 pixels. The preprocessing steps included:

* **Reshaping**: Each image was converted into a one-dimensional vector of 784 features.
* **Normalization**: Pixel values were scaled to the range [0,1] for better convergence of models.
* **Exploratory Data Analysis**:
  + A histogram of digit distributions was plotted to ensure balanced classes.
  + A correlation matrix was generated to analyze feature redundancy.
* **Data Splitting**: The dataset was split into training (80%) and validation (20%) subsets using a stratified approach to ensure balanced class representation.

## 3. Model Development

Three classification models were implemented:

* **Logistic Regression**: A simple linear classifier that models the probability of each class using a softmax function.
* **Decision Tree**: A tree-based model that recursively splits the data based on the most informative features.
* **Random Forest**: An ensemble of decision trees that aggregates multiple predictions to improve accuracy and reduce overfitting.

Hyperparameters used:

* **Logistic Regression**: Solver = ‘lbfgs’, Max iterations = 500
* **Decision Tree**: Max depth = 15, Min samples split = 5
* **Random Forest**: Number of estimators = 100, Max depth = 20, Min samples split = 5, Random state = 42

## 4. Results and Discussion

The models were evaluated using the following classification metrics:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Logistic Regression | 92.0% | 92.0% | 92.0% | 92.0% |
| Decision Tree | 88.0% | 88.0% | 88.0% | 88.0% |
| Random Forest | 97.0% | 97.0% | 97.0% | 97.0% |

* **Confusion Matrix Analysis**:
  + Logistic Regression performed well but struggled with certain digits (e.g., 3 and 8).
  + Decision Tree showed lower performance due to overfitting, even with max depth restriction.
  + Random Forest significantly outperformed both models, demonstrating strong generalization capabilities.
* **ROC and AUC Analysis**:
  + The ROC curve was plotted for Logistic Regression and Random Forest.
  + The AUC score for Logistic Regression was **0.92**, while for Random Forest it was **0.98**, indicating superior performance.
* **Error Analysis**:
  + The most common misclassifications occurred between digits with similar shapes (e.g., 4 and 9, 3 and 8).
  + Logistic Regression misclassified 8% of digit 8 as digit 3.
  + Decision Trees had difficulty distinguishing digits with high pixel variance.

## 5. Conclusion

Among the three models tested, Random Forest achieved the highest accuracy (97%) and overall performance. This suggests that ensemble learning is highly effective for digit classification. Potential improvements could include:

* Using deep learning approaches such as Convolutional Neural Networks (CNNs) for better feature extraction.
* Fine-tuning hyperparameters through Grid Search or Random Search.
* Experimenting with additional preprocessing techniques like feature extraction using PCA.
* Implementing ensemble stacking methods to combine multiple classifiers.

## 6. References

* Scikit-learn Documentation: https://scikit-learn.org/stable/
* TensorFlow Keras Documentation: https://www.tensorflow.org/api\_docs

This report provides a structured analysis of supervised learning algorithms for MNIST classification and highlights key performance differences among them.

## 7. Visualizations

Below are the key visualizations used in the study:

1. **Histogram of Digit Distribution**
2. **Confusion Matrices for all models**
3. **ROC Curves for Logistic Regression and Random Forest**
4. **Examples of Misclassified Digits**

These visualizations provide insights into the model performances and error patterns, further justifying the results.