Digit Recognizer using Logistic Regression and K-Nearest Neighbour Algorithms

Nethsarani D.W.D - EG/2020/4096 , Isanka J.R.A.U. - EG/2020/3974

Abstract — This paper gives a detailed analysis of a handwritten digit recognition system which is built using K-Nearest Neighbour (KNN) algorithms and logistic regression. The logistic regression model gives information on the likelihood that a given input will fall into a particular digit class.The KNN method effectively utilises the proximity of data points in feature space for categorization purposes. Throughout the project we compare these two models' performances on a dataset, showing the advantages and disadvantages of each model.

I. INTRODUCTION

The ability to correctly understand and read handwritten text is essential in an increasingly digitalized environment. In order to solve the fundamental problem of identifying and classifying handwritten numbers (0–9) derived from images, the Digit Recognizer project combines the fields of computer vision and image classification.

A. Background

Handwriting recognition has become essential in many domains, such as analyzing documents, computerized form processing, and character recognition systems, because to the explosion of digital information. This study explores the difficult task of using machine learning algorithms to comprehend handwritten numbers, advancing the field of character recognition technology.

B. Importance

Accurate digit recognition is important for purposes that go beyond computer vision and machine learning. Applications for it are numerous and include banking (processing checks), medical forms, postal automation, and other fields where accurate digitalization of handwritten records is required. This project intends to improve and optimize automated processes, boost efficiency and reliability by developing robust models that can recognize digits with high accuracy.

C. Project Goals

The main objective of this work is to create a machine learning model able to effectively identify and classify handwritten numbers in images. The purpose of this project and implementation of algorithms such as K-Nearest Neighbors (KNN) and Logistic Regression is to determine how well they perform this particular recognition task and to evaluate the advantages and disadvantages of each in relation to this case.

D. Algorithm Choice

The selection of KNN and logistic regression is based on their unique advantages for this particular project. For initial investigation and evaluation, logistic regression is a good choice because it gives a solid foundation for multi-class classification given its simplicity. However, KNN is a strong competitor for this image classification challenge due to its non-parametric character and its ability to identify complex correlations within data. But comparing it with Logistic Regression is necessary because of its computational requirements and scalability limitations.

II. METHODOLOGY

A. Dataset Overview

This project's main dataset source from the Kaggle Digit Recognizer competition. The handwritten numbers (0–9) in grayscale are included in this dataset. The 42,000 handwritten digit samples in the dataset are represented by pixel values, which become features for the machine learning model. Each image's matching digit (0–9) is assigned by the output variable.

Dataset Source: Kaggle

DatasetLink:https://www.kaggle.com/competitions/digit-recognizer/data

B. Data Pre-processing

Handling Missing Data: Fortunately, there are no missing values in the dataset. There was no need for a particular approach of missing data.

Feature Scaling: To ensure stability and numerical stability through model training, pixel values were normalized. The scale of the pixel values ranged from 0 to 1.

Training/Test Split: Using an 80/20 split, the dataset was split into training and test sets, with 80% of the data used for model training and 20% for evaluation of performance.

C. Algorithms

The project initially considered two algorithms: Logistic Regression and K-Nearest Neighbors (KNN).

Logistic Regression: Logistic Regression is a linear classification algorithm used for binary and multi-class classification problems.

Equation (Sigmoid Function): The logistic function $\frac{1}{1+e^{-z}}$ maps the linear combination z = ßo + ß1x1 + ß2x2 + ... + ßnxn. It models the probability of the occurrence of a class using a sigmoid function and optimizes coefficients to minimize error.

K-Nearest Neighbors (KNN): KNN is a non-parametric algorithm used for classification and regression.It uses the 'k' nearest neighbors' majority vote in the feature space to estimate the class of a sample.

D. Implementation

Code Source: The code for implementing Logistic Regression and KNN algorithms was obtained from the Scikit-Learn library for Python .

Modifications: No modifications were made to the core algorithm implementation.

Experiment Settings:

- Logistic Regression:
- Parameters: Regularization parameter , optimization algorithm, and multi-class handling.
- Settings: Experimented with different 'C' values and solvers to optimize performance.
- K-Nearest Neighbors:
- Parameters: 'K' (number of neighbors), distance metric, and weighting function.
- Settings: Evaluated performance with varying 'k' values and distance metrics.

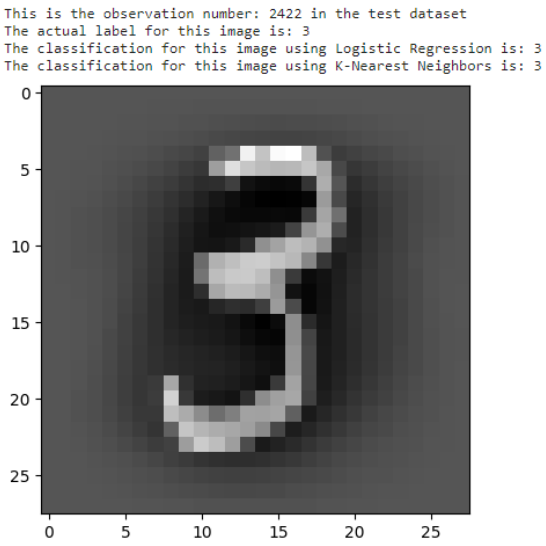
III. RESULTS

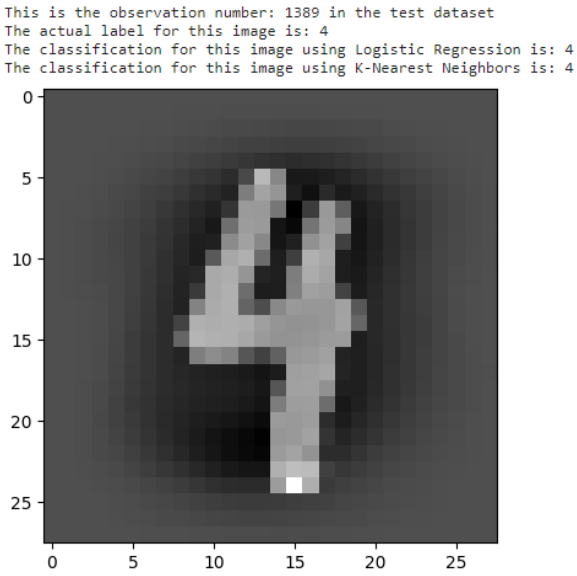
The logistic regression model was trained for 10 epochs on the MNIST dataset, consisting of 60,000 training images and 10,000 validation images. The training history shows a good, steady improvement in both training and validation accuracy over epochs. Starting with an initial accuracy of 91.89%, the model achieved an impressive 99.51% accuracy on the training set by the 10th epoch. The validation accuracy also shows a good improvement, reaching 96.86%.

A high accuracy and good precision on the validation dataset was shown by the logistic regression model. The overall accuracy on the validation set was 96.86%, showcasing the model's ability to generalize well to unseen data. Additionally, the precision score was 0.97 and it indicates the model's proficiency in correctly classifying digits.

The K-nearest neighbors model, trained using default parameters, achieved a commendable validation accuracy of 94.01%. Comparing these values of accuracy with the logistic regression model shows the distinct performance characteristics of each model.

Random observations from the validation set were selected, analyzed and the results showcase the model's capability to accurately classify digits. For example, observation 2422, with the actual label '3', was correctly classified by both the logistic regression and KNN models as '3'. Similarly, observation 1389, labeled as '4', received accurate classifications from both models.

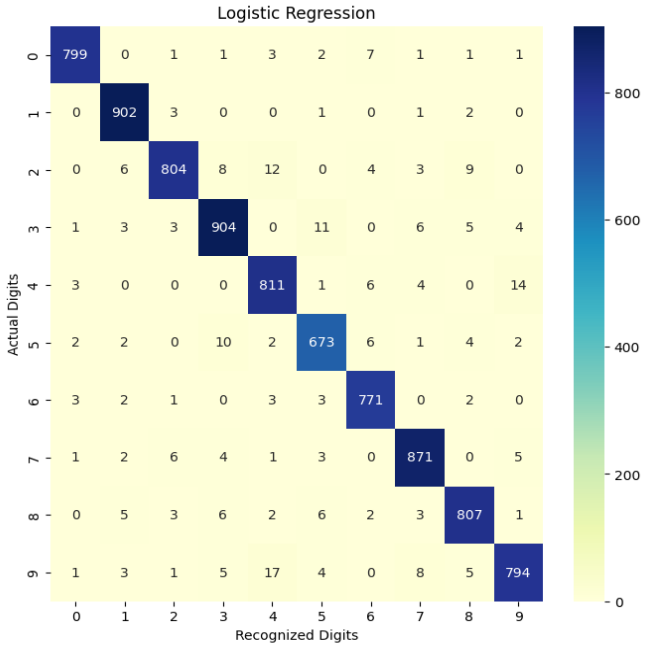




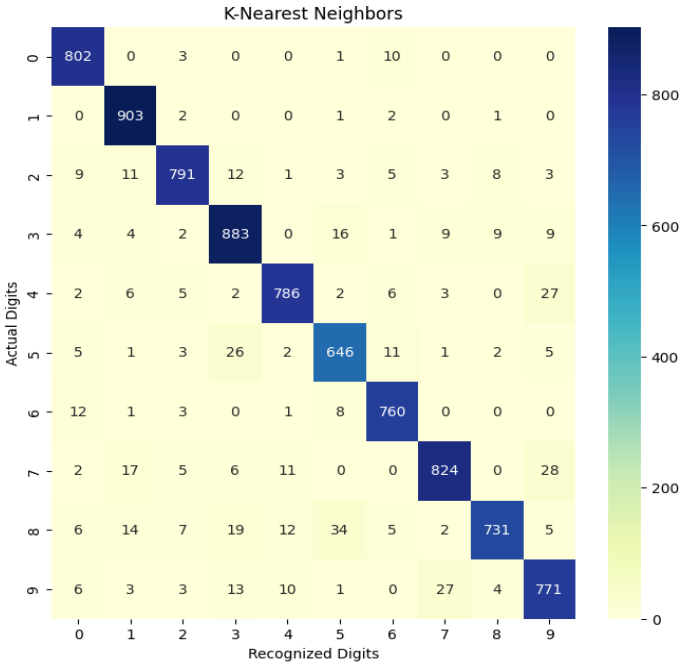
A. Confusion Matrices

Confusion matrices show the performances of both KNN model and the logistic regression model for recognizing handwritten digits. The confusion matrix for logistic regression and KNN are presented below, highlighting the number of correct and misclassified instances for each digit. These matrices provide a clear idea of the model's strengths and potential areas that need to be improved.

Logistic Regression Confusion Matrix:



K-Nearest Neighbors Confusion Matrix:



The confusion matrices provide both models' experiences at recognizing digits while showing the potential areas that need to be improved.

Both logistic regression and KNN show an excellent performance in handwritten digit recognition.

The results show the efficiency of both logistic regression and K-nearest neighbors models.It provides valuable information for us to do improvements and optimizations for our project in the future.

.

IV. DISCUSSION

1. Model Performance and Generalization:

We first describe the noteworthy results of the K-nearest neighbors (KNN) and logistic regression models for digit identification. In both the training and validation sets, the logistic regression model showed a strong capacity to identify complex patterns within the training set and achieved high accuracy. On the other hand, the KNN model performed marginally better than logistic regression, despite its excellent accuracy.

2. Ethical Considerations:

a. Bias and Fairness:

A crucial ethical consideration concerns the possibility of bias present in the models. When it comes to underrepresented digits or demographic groups, it is especially important to examine if the models show biases in their projections. To guarantee the fair treatment of all digits and user groups, bias mitigation measures such as varied dataset curation and algorithmic fairness procedures must be investigated..

b. Transparency and Explainability:

Attention must be paid to these models' explainability and openness. We need to improve the models since they become more combined into decision-making. Building trust and encouraging ethical AI techniques are facilitated by making sure that stakeholders understand the decision-making processes of the models.

3. Model Interpretability:

It is important to understand how machine learning models made choices for applications. Knowing how a model classifies a digit builds user confidence. To understand the reasoning behind the models' decisions, one can investigate methods like layer-wise relevance propagation and SHAP (SHapley Additive exPlanations) values.

4. Scalability and Deployment:

The models' implementation in real-life situations and scalability are taken into consideration. Wide accessibility and usefulness depend on the digit recognizer models' ability to be simple to integrate into a variety of applications, from web platforms to mobile devices.

V. CONCLUSION

This study concludes with a look into digit recognition using the K-nearest neighbors and logistic regression techniques. The models that are shown work rather well; the logistic regression model is quite accurate, and the KNN model is a competitive option. When using such models, it is important to concern about in removing biases, guaranteeing

fairness, and fostering openness as the examination of ethical issues makes clear.

The ethical issues raised highlight how important it is to put interpretability and fairness first when developing machine learning algorithms. There are major ethical issues with the use of digit recognition systems in banking and healthcare, among other industries

To make these models better, resolve moral issues, create an environment of cooperation that values responsible AI activities, more effort will be need in the future. These results show new opportunities for improve ethical stability of machine learning systems.

REFERENCES

[1] https://medium.com/@nikhilanandikam/handwritten-digit-recognition-hdr-using-k-nearest-neighbors-knn-f4c794a0282a

[2] U. R. Babu, Y. Venkateswarlu and A. K. Chintha, "Handwritten Digit Recognition Using K-Nearest Neighbour Classifier," *2014 World Congress on Computing and Communication Technologies*, Trichirappalli, India, 2014, pp. 60-65

[3] https://github.com/Ronny-22-Code/Handwritten-Digit-Classification-using-KNN

[4] https://medium.com/analytics-vidhya/handwritten-digit-recognition-using-logistic-regression-8d3b3f7e31c0.

[5] https://www.geeksforgeeks.org/identifying-handwritten-digits-using-logistic-regression-pytorch/

[6] https://atmamani.github.io/projects/ml/mnist-digits-classification-using-logistic-regression-scikit-learn/

[7] https://github.com/Yuvrajchopra25/Project-5-MNIST-Handwritten-Digit-Recognition-using-Sklearn-and-LogisticRegression