**Open Ended Lab Report**

**For**

**Machine Learning**

**2022-2026**

**Submit To: Engr. Awais Rathore**

**Submit By: 2022-SE-06**



**The University of Azad Jammu and Kashmir Muzaffarabad**

**Department of Software Engineering**

**Fifth Semester**

**Classification of MNIST Handwritten Digits Using Machine Learning**

**1. Introduction**

The MNIST dataset consists of handwritten digits (0-9) represented as 28x28 pixel grayscale images. The objective of this study is to classify these digits using various machine learning models and compare their performance. The dataset has been preprocessed by flattening the images into a 1D vector of 784 features and splitting them into training and testing sets, stored in CSV files.

**2. Methodology**

**2.1 Data Preprocessing**

* Loaded the MNIST dataset using pandas.
* Checked for missing values and filled them using the median.
* Separated features and labels.
* Scaled the features using Min-Max scaling to bring values between 0 and 1.
* Split the dataset into training (80%) and testing (20%) sets.

**2.2 Models Used and Justification The following machine learning models were trained and evaluated:**

* K-Nearest Neighbors (KNN):
  + Chosen due to its simplicity and effectiveness for image classification.
  + Works well with small to medium-sized datasets by comparing new samples to existing ones.
* Decision Tree:
  + Selected for its ability to handle complex decision boundaries.
  + Provides interpretable classification results.
* Logistic Regression:
  + Used as a baseline model for comparison.
  + Works well for linear separable problems and provides probabilistic outputs.
* Naive Bayes (GaussianNB):
  + Chosen for its fast training time and ability to handle small datasets efficiently.
  + Assumes feature independence, which can limit its performance on complex datasets like MNIST.

**2.3 Model Evaluation Each model was evaluated using the following metrics:**

* Accuracy
* Classification report (precision, recall, f1-score)
* Confusion matrix visualization using seaborn.

(Insert classification report here)

**2.4 Hyperparameter Tuning RandomizedSearchCV was used for hyperparameter tuning:**

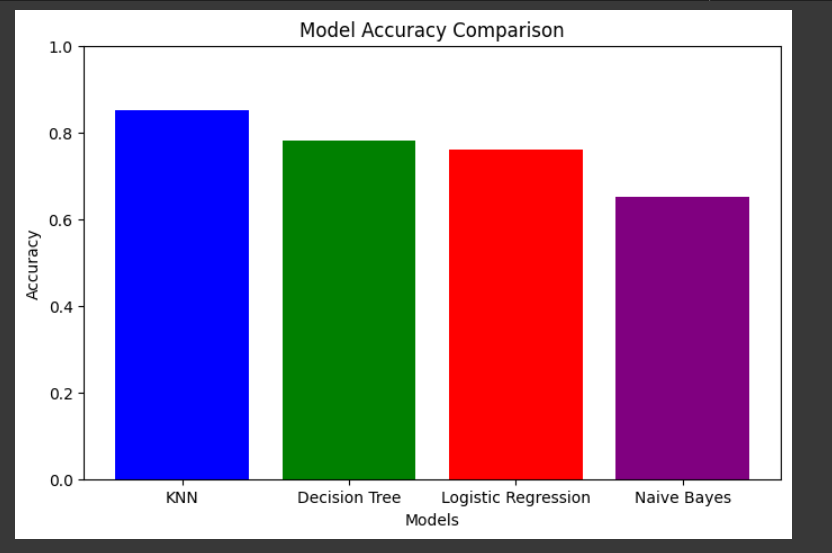
* KNN: Optimized n\_neighbors (range: 3 to 10).
* Decision Tree: Tuned max\_depth and min\_samples\_split.
* Logistic Regression: Optimized C values.
* Naive Bayes: Adjusted var\_smoothing.

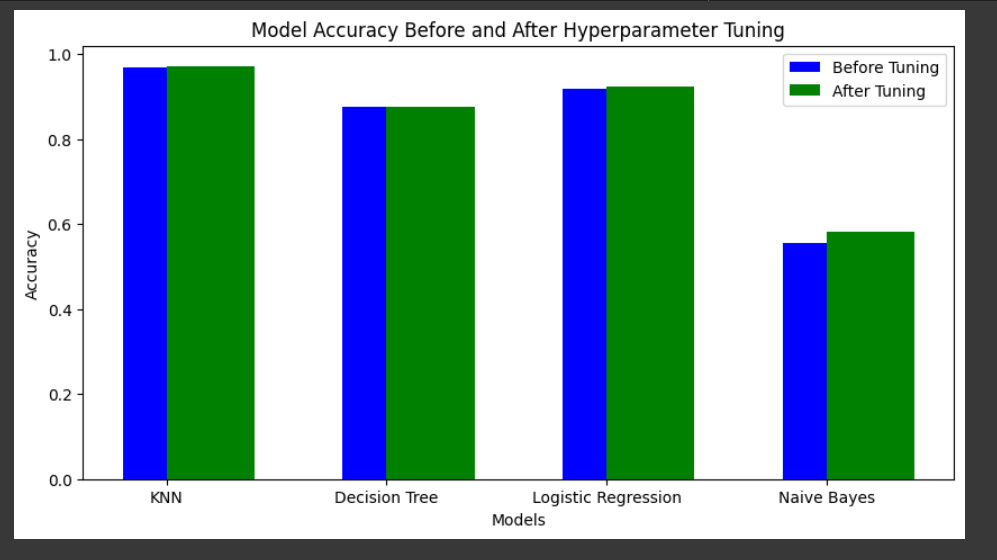
**3. Results**

| **Model** | **Accuracy (Before Tuning)** | **Accuracy (After Tuning)** |
| --- | --- | --- |
| **K-Nearest Neighbors** | **0.9695** | **0.9714** |
| **Decision Tree** | **0.8770** | **0.8763** |
| **Logistic Regression** | **0.9192** | **0.9232** |
| **Naive Bayes** | **0.5552** | **0.5828** |

**4. Graphical Representation**

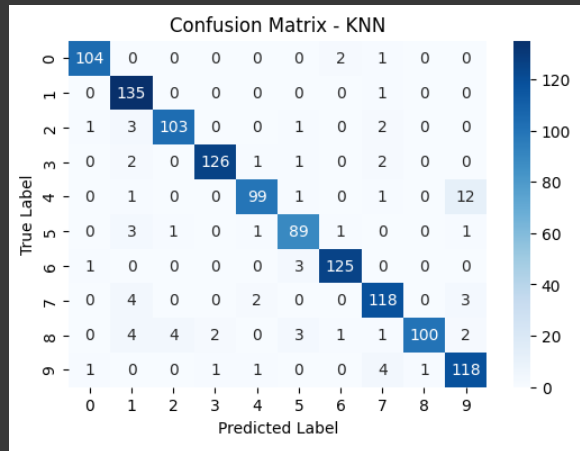
**4.1 Accuracy Comparison**



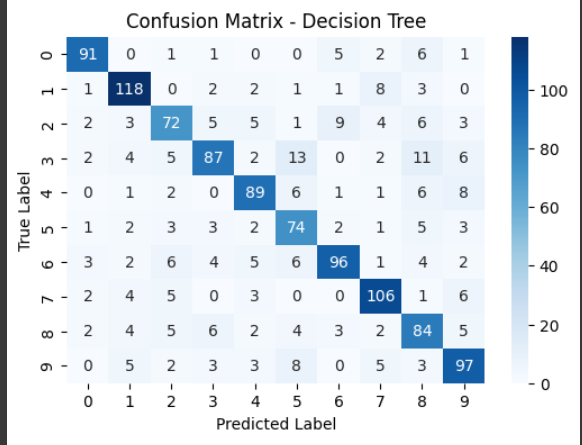


**4.2 Confusion Matrices:**

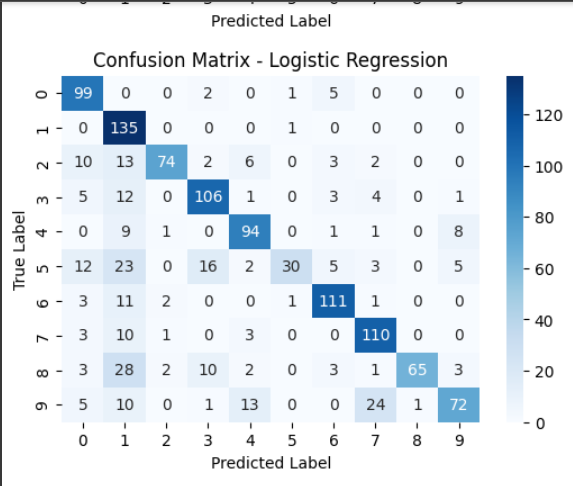
**a.Knn**



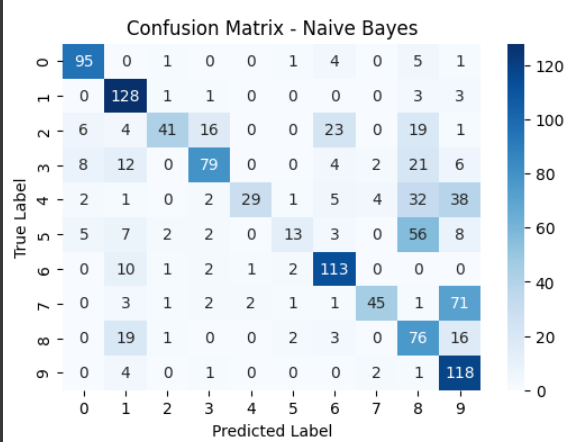
**b. Decision Tree:**



**c.Logistic Regression:**



**d. Naive Bayes:**



**5. Discussion**

* Best Performing Model: KNN achieved the highest accuracy (97.14%) after tuning.
* Worst Performing Model: Naive Bayes performed poorly due to its assumption of feature independence, which does not hold for image data.
* Effect of Hyperparameter Tuning:
  + KNN and Logistic Regression saw improvements in accuracy.
  + Decision Tree remained stable, indicating that further feature engineering may be required.
  + Naive Bayes showed slight improvement but remained the weakest.

**6. Conclusion** Among the models tested, the K-Nearest Neighbors (KNN) algorithm demonstrated the highest classification accuracy on the MNIST dataset, closely followed by Logistic Regression. Decision Trees performed moderately well, offering interpretability but slightly lower accuracy. In contrast, the Naive Bayes model struggled, likely due to its assumption of feature independence, which does not align well with the nature of image data. To further enhance classification accuracy, future work could explore advanced techniques such as Convolutional Neural Networks (CNNs), which are specifically designed for image recognition tasks and have been proven to yield superior performance on similar datasets.