CSCI 164 Project - AI Model Analysis for Handwritten Character Recognition

J. L. Bravo, Mattheau Casey, Miguel Ibarra

April 28, 2025

Abstract

This project analyzes three distinct datasets — MNIST, EMNIST Balanced, and IMDb Reviews — to compare the performance of classification models across different types of data. Specifically, we evaluate Logistic Regression and K-Nearest Neighbors algorithms, examining their accuracy, prediction scores, and overall effectiveness on each dataset. The goal is to assess both the differences between datasets and the comparative strengths of the models.

Contents

1 Dataset Selection

The objective of this project is to take 3 datasets and compare our models, analyzing their performance and how their hyperparameters change with respect to differing or more complicated data.

1.1 MNIST

The MNIST dataset is a widely used benchmark in machine learning and computer vision, consisting of 70,000 images of handwritten digits (0–9). Each image is a 28x28 pixel grayscale image, flattened into a 784-dimensional feature vector. In this project, MNIST serves as a standardized dataset for evaluating classification models. Its simplicity and well-understood structure allow us to effectively test and compare the performance of Logistic Regression and K-Nearest Neighbors algorithms on clean, labeled image data.

1.1.1 Reference Work: Keerthana G. and Usha Sree R.

In the paper "Handwritten Digit Recognition Using Logistic Regression" by Keerthana G. and Usha Sree R., the authors explore the application of logistic regression for classifying handwritten digits from the MNIST dataset. Their approach involves preprocessing the images by normalizing pixel values to a [0,1] range and flattening the 28×28 pixel images into 784-dimensional feature vectors. They implement logistic regression in a one-vs-rest framework to handle the multiclass nature of the problem and use gradient descent for optimization. The model's performance is evaluated using accuracy, precision, recall, and F1-score, achieving a final test accuracy of 92.4%. The authors also perform error analysis to understand misclassifications and highlight the model's potential applications in areas like optical character recognition (OCR). Their results demonstrate that logistic regression, despite being a relatively simple model, can be an effective baseline for handwritten digit recognition tasks.

1.2 EMNIST Balanced

The EMNIST Balanced dataset extends the MNIST format to include both handwritten letters and digits, offering a more diverse and challenging classification task. It contains 131,600 images across 47 balanced classes, with each image being a 28x28 pixel grayscale representation. The Balanced dataset classes include handwritten characters for digits (0 - 9), every capital letter of the english alphabet, and a small selection of under-case english letters. In this project, EMNIST Balanced is used to further evaluate the performance of Logistic Regression and K-Nearest Neighbors models, providing insight into how each algorithm handles a larger and more complex set of handwritten characters.

1.2.1 Reference Work: Buatoom, T. and Chaiyasoonthorn, W.

In their 2023 study, Buatoom and Chaiyasoonthorn introduce a novel approach to enhance image classification performance by integrating statistically weighted dimensions with principal component analysis (PCA). The method assigns weights based on three statistical behaviors: collection-class, inter-class, and intra-class variations. This weighting aims to emphasize features that are more discriminative across classes. The authors applied this technique to three datasets—MNIST, EMNIST, and Fashion-MNIST—using classifiers such as logistic regression, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM). Their findings indicate that combining statis-

tically weighted dimensions with PCA not only improves classification accuracy but also reduces feature dimensionality by over 50%. Notably, KNN showed an average accuracy improvement of 1.48%, while logistic regression improved by 0.48% when using this combined approach. The authors here demonstrate great methods of finding a balance of computational efficiency and accuracy, which is emphasized in this experiment and future work.

1.3 IMDb Reviews

The IMDb Reviews dataset is a collection of 50,000 movie reviews labeled as either positive or negative. Unlike MNIST and EMNIST, this dataset is based on natural language text rather than images, presenting a different type of classification challenge. In this project, we use the IMDb dataset to evaluate how Logistic Regression and K-Nearest Neighbors perform on text data after appropriate preprocessing, allowing us to compare model performance across different data modalities.

1.3.1 Reference Work: Rustam Talibzade

In the study "Sentiment Analysis of IMDb Movie Reviews Using Traditional Machine Learning Techniques and Transformers" by Rustam Talibzade (2023), the author investigates the effectiveness of different machine learning approaches for classifying sentiment in the IMDb review dataset. The dataset undergoes preprocessing steps including text cleaning and tokenization, followed by feature extraction using Bag-of-Words (BoW) and TF-IDF representations. Several models are evaluated, including Logistic Regression, Naïve Bayes, Support Vector Machines (SVM), and the transformer-based model BERT. The results show that while traditional machine learning models like Logistic Regression and SVM perform reasonably well, achieving competitive accuracy, BERT significantly outperforms them with an accuracy of 98%. However, this improvement comes at the cost of higher computational complexity and longer training times. The study highlights the trade-offs between traditional methods and modern deep learning approaches for sentiment analysis tasks.

2 Model Selection

The following classification models were implemented to meet the project requirements:

2.1 Logistic Regression

Logistic Regression is a widely used classification algorithm that models the probability of a data point belonging to a particular class. It operates by fitting a linear decision boundary and applying the logistic (sigmoid) function to map outputs to a probability between 0 and 1. In this project, Logistic Regression is used to classify both image and text data, allowing us to evaluate its effectiveness across structured pixel inputs and unstructured language inputs.

2.2 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm that classifies data points based on the majority class among their closest neighbors in the feature space. The algorithm is simple yet powerful, relying on distance metrics such as Euclidean distance to make predictions. In this project, KNN is applied to both the image datasets and the text-based IMDb dataset, providing a comparison to Logistic Regression in terms of performance and adaptability to different data types.

2.3 Naive Bayes

For this project, we use the Multinomial Naive Bayes variant, which is well-suited for classification tasks involving discrete features such as word counts in text data. Multinomial Naive Bayes applies Bayes' Theorem under the assumption that features are conditionally independent given the class label, and incorporates smoothing to handle unseen features. In this project, it is primarily applied to the IMDb Reviews dataset, allowing us to compare its performance against Logistic Regression and K-Nearest Neighbors on natural language text.

3 Hyperparameter Selection

For all three of the models used throughout this project, a Randomized Grid Search was used to find hyperparameters with a balance between the limited time to compute and a desirable improvement in accuracy. In many cases, the accuracy scores provided by the limited time to search felt unsatisfactory, however these results still showed valuable comparisons between the models of choice and how they reacted to different data modalities.

The amount of cv folds and iterations allowed for the grid search depended on the time complexity of each model/dataset combination. In the end the following hyperparameters were found:

- MNIST
  - Logistic Regression
    - Accuracy Score: 91.99%
    - C: 1.6176755124355395
    - Tolerance: 0.00127722364893473
  - K-Nearest Neighbors
    - Accuracy Score: 96.35%
    - n neighbors: 4
    - p: 1
    - weights: distance
  - Naive Bayes
    - Accuracy Score: 81.87%
    - alpha: 0.05808361216819946
- EMNIST Balanced
  - Logistic Regression
    - Accuracy Score: 69.73%
    - C: 8.774359226762291
    - Tolerance: 0.0029217098805921896
  - K-Nearest Neighbors
    - Accuracy Score: 70.91%
    - n neighbors: 4
    - p: 1
    - weights: distance
  - Naive Bayes
    - Accuracy Score: 53.84%
    - alpha: 0.05808361216819946
- IMDb Reviews
  - Logistic Regression

    - Accuracy Score: 81.5%
    - C: 50.475441833336404
    - Tolerance: 0.009100542683761407
  - K-Nearest Neighbors
    - Accuracy Score: 60.0%
    - n neighbors: 5
    - p: 2
    - weights: uniform
  - Naive Bayes
    - Accuracy Score: 79.5%
    - alpha: 0.15601864044243652

4 Performance Evaluation

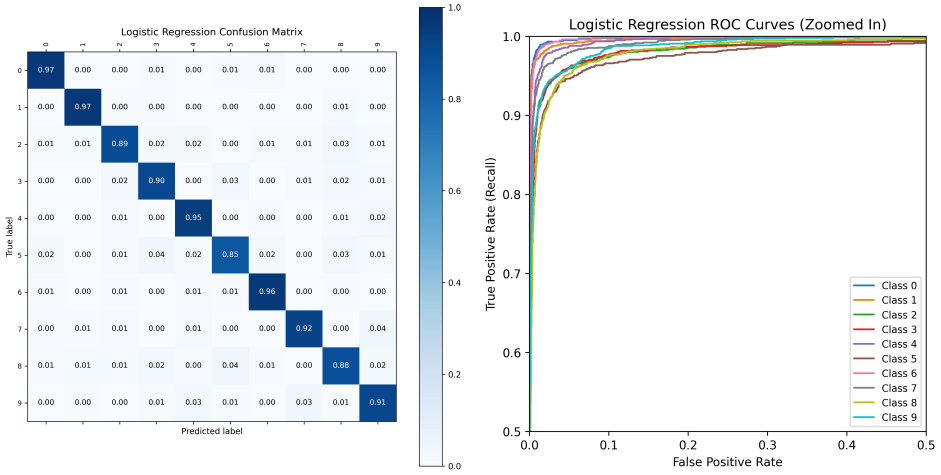
4.1 MNIST

For each model that was used to evaluate the MNIST dataset, the pixel dimensionality was reduced to a scale of [0,1] as part of our pre-processing. An attempt was made to reduce dimensionality further using PCA, as was shown to be helpful in some of our reference works, though for this project there were difficulties with resulting data that led to a necessity to forego using PCA entirely for our datasets. A train/test split of 20% was used.

4.1.1 Logistic Regression

The results for Logistic Regression on MNIST show the following weighted average scores for each class:

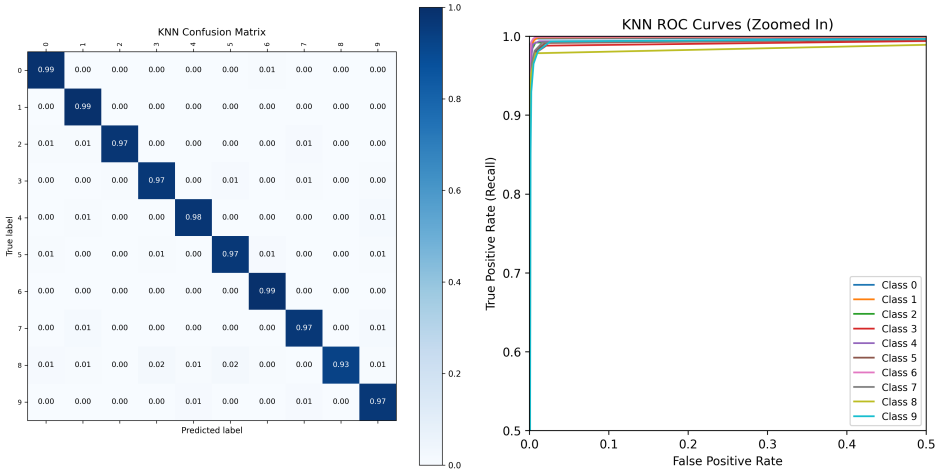
- Precision: 91.96%
- Recall: 91.98%
- F1 Score: 91.96%



4.1.2 K-Nearest Neighbors

The results for Logistic Regression on MNIST show the following weighted average scores for each class:

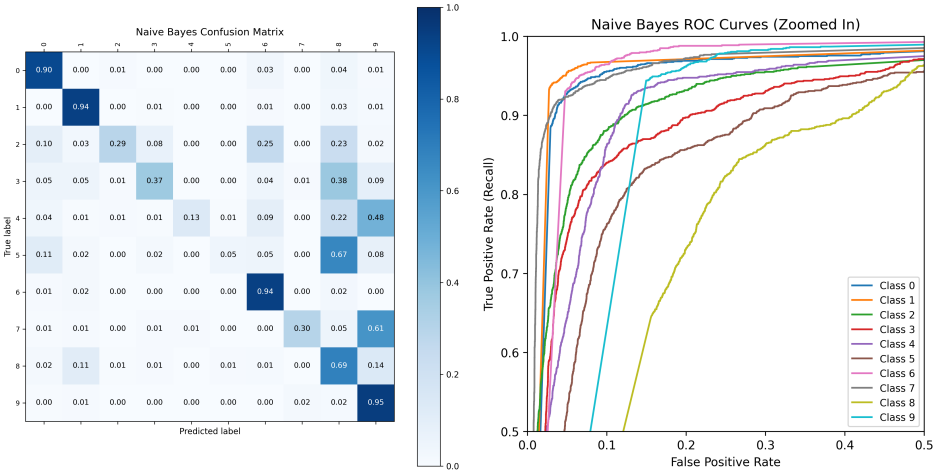
- Precision: 97.24%
- Recall: 97.22%
- F1 Score: 97.22%



4.1.3 Naive Bayes

The results for Logistic Regression on MNIST show the following weighted average scores for each class:

- Precision: 68.56%
- Recall: 56.24%
- F1 Score: 51.8%



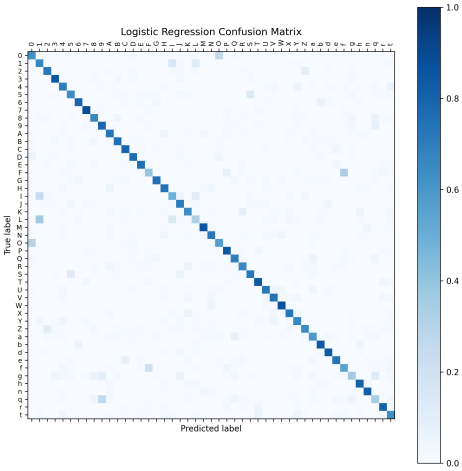
4.2 EMNIST Balanced

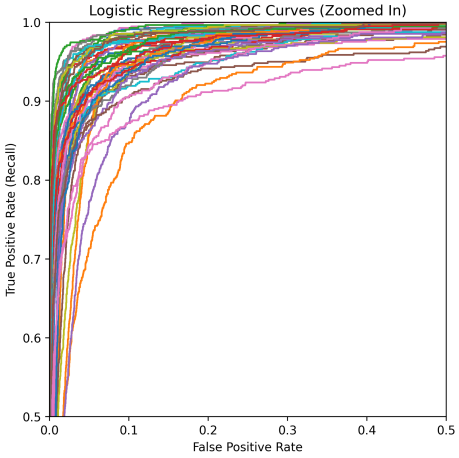
During our testing with the EMNIST dataset, the pixel dimensionality was also reduced to a scale of [0.1] as a part of pre-processing. Though due to difficulties involving the return of negative values we were required to foresake PCA on our chosen datasets. The dataset was divided into a train/test split of 20% for use in our examination.

4.2.1 Logistic Regression

The results for Logistic Regression on EMNIST show the following weighted average scores for each class:

- Precision: 69.5%
- Recall: 69.73%
- F1 Score: 69.44%

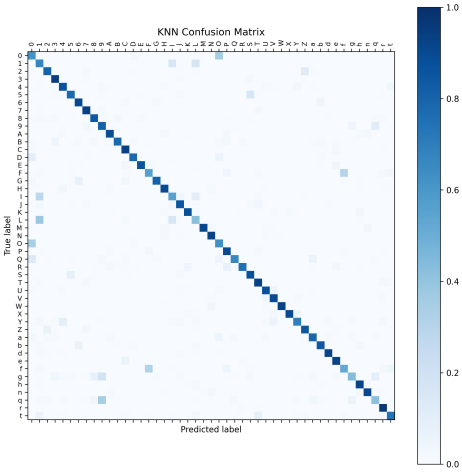


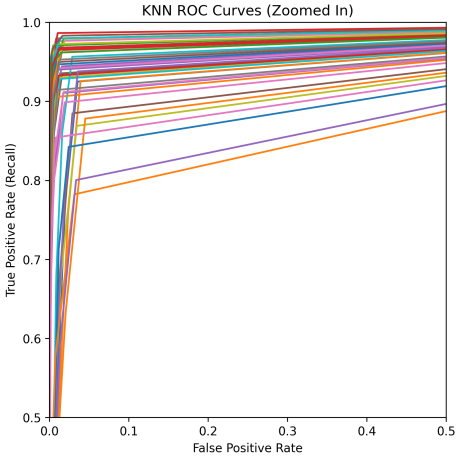


4.2.2 K-Nearest Neighbors

The results for K-Nearest Neighbors on EMNIST show the following weighted average scores for each class:

- Precision: 80.76%
- Recall: 79.76%
- F1 Score: 79.77%

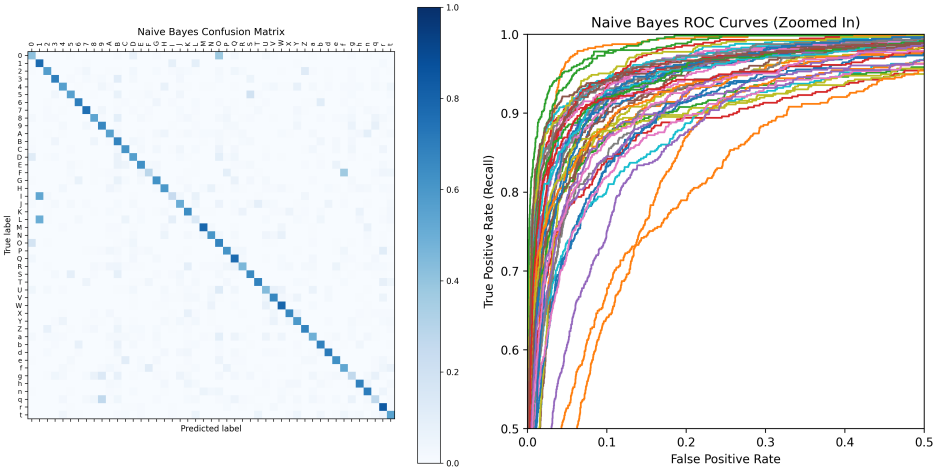




4.2.3 Naive Bayes

The results for Naive Bayes on EMNIST show the following weighted average scores for each class:

- Precision: 59.89%
- Recall: 59.11%
- F1 Score: 58.71%



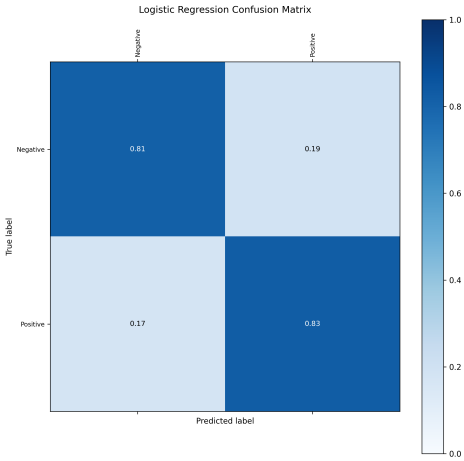
4.3 IMDB Reviews

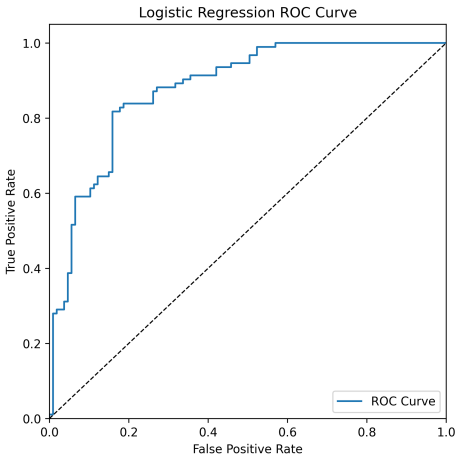
Out of the three datasets this was the simplest and thus produced the simplist graphs for examination, yet it is worth it as it provides examples that help describe the differences between the three model's preformances. In this data set, a train / test split of 20% was also used.

4.3.1 Logistic Regression

The results for Logistic Regression on IMDb Reviews show the following weighted average scores for each class:

- Precision: 82.10%
- Recall: 82.00%
- F1 Score: 82.01%

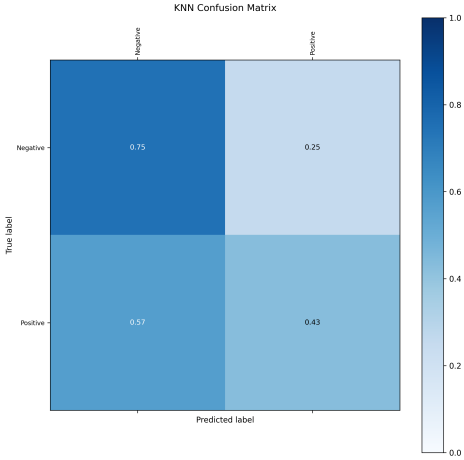


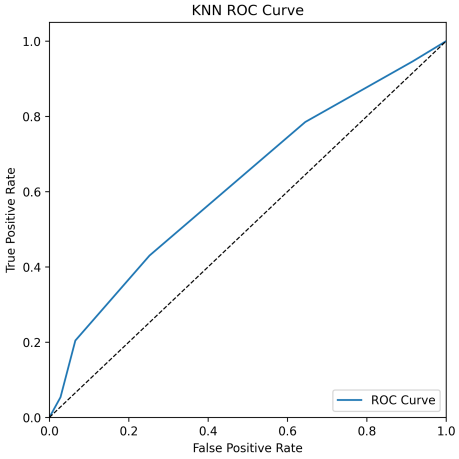


4.3.2 K-Nearest Neighbors

The results for K-Nearest Neighbors on IMDb Revciews show the following weighted average scores for each class:

- Precision: 59.94%
- Recall: 60.00%
- F1 Score: 58.91%



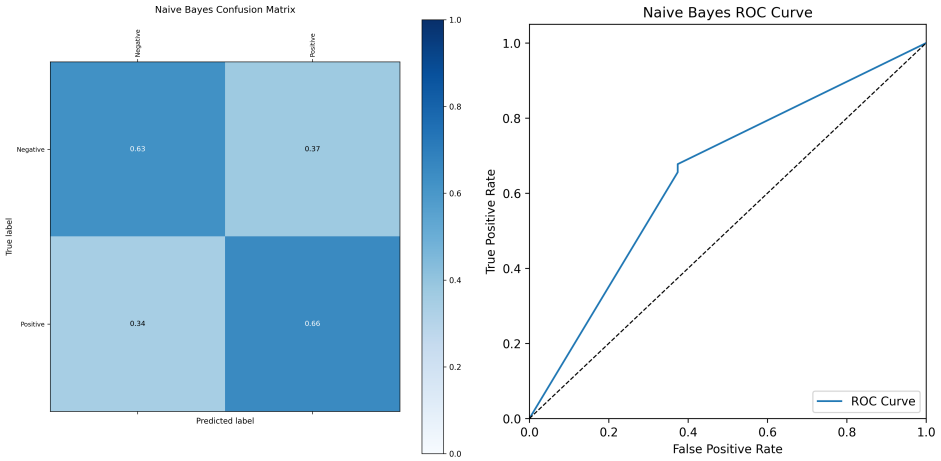


4.3.3 Naive Bayes

The results for Naive Bayes on IMDb Reviews show the following weighted average scores for each class:

- Precision: 64.29%

- Recall: 64.00%
- F1 Score: 64.04%



4.4 Result Analysis

In our experiments, we aimed to test models across datasets of varying complexity to observe how each model responded to different types of information. The simplest dataset was the IMDb Reviews, which involved binary classification (positive or negative sentiment) based on tokenized text data. This contrasts with the MNIST and EMNIST datasets, which are far more complex and involve image-based classification tasks.

4.4.1 MNIST vs EMNIST

A major focus of our experimentation was comparing model performance between MNIST and EMNIST, and observing which aspects of the expanded EMNIST dataset introduced new challenges. Since EMNIST contains all of MNIST plus additional handwritten letters, it provided a useful benchmark for how models adapted to increased complexity. As expected, performance on EMNIST was lower across all models compared to MNIST. Analysis of the confusion matrices showed that the largest errors arose from the similarity between certain characters, such as '0' and 'O', 'F' and 'f', and '1' and 'I'. Despite this, Naive Bayes achieved relatively strong precision on MNIST, though it struggled with recall and F1 scores, indicating it was better at avoiding false positives but worse at capturing all relevant instances.

4.4.2 Model Comparisons

Across the MNIST and EMNIST datasets, K-Nearest Neighbors consistently outperformed Logistic Regression and Naive Bayes. This outcome was anticipated, as KNN's instance-based learning approach is well-suited to image recognition tasks, where local pixel patterns play a significant role. In contrast, Logistic Regression, while strong for continuous input spaces, is less naturally aligned with the discrete pixel cluster structures needed for character recognition.

In the IMDb Reviews dataset, however, Logistic Regression achieved the best performance. This result aligns with expectations, given that while the IMDb dataset is binary in output, its input features (text embeddings) are more continuous and well-suited to a linear classifier.

It is also notable that Naive Bayes performed the worst overall, except on the IMDb dataset where it was competitive. This difference highlights a limitation of Naive Bayes: its assumption of feature independence, which clashes with image data where nearby pixels are strongly correlated. Without modeling the relationships between features, Naive Bayes is unable to capture the holistic patterns necessary for accurate image classification.

5 Future Work

It was during our experiments with the models that we found our limit early on, where it became clear to us that with our current machinery it would be infeasible to push the limits. Topics that we found interesting the journals pertaining to the dataset had to quickly be pushed aside due to said limits. It is our hope that with time we may acquire the hardware nessessary and conducive to further experimentation.

6 Conclusion

Throughout this project, we compared the performance of Logistic Regression, K-Nearest Neighbors, and Naive Bayes models across three distinct datasets: MNIST, EMNIST Balanced, and IMDb Reviews. Our experiments demonstrated that model performance is highly dependent on the structure and complexity of the data. K-Nearest Neighbors consistently achieved the highest scores on image-based datasets (MNIST and EMNIST) due to its non-parametric nature and ability to adapt to local patterns in pixel data. Logistic Regression performed well across all datasets, particularly excelling in the IMDb Reviews dataset, where the feature space is more continuous

and suited to a linear decision boundary. Naive Bayes, while competitive for text-based sentiment classification, struggled significantly on image data due to its strong independence assumptions between features. These results emphasize the importance of carefully selecting machine learning models based on data characteristics. Future work could explore deeper models, such as convolutional neural networks or transformer-based architectures, to further improve classification performance, especially on more complex datasets like EMNIST.

7 References

- Buatoom, T., & Chaiyasoonthorn, W. (2023). Improving Classification Performance with Statistically Weighted Dimensions and Dimensionality Reduction. Applied Sciences, 13(3), 2005. https://doi.org/10.3390/app13032005&8203;:
- Keerthana, G., & Usha Sree, R. (2024). Handwritten Digit Recognition Using Logistic Regression. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 4(6), 106–110. https://doi.org/10.48175/IJARSCT-22523
- Talibzade, R. (2023). Sentiment Analysis of IMDb Movie Reviews Using Traditional Machine Learning Techniques and Transformers. ResearchGate. https://doi.org/10.13140/RG.2.2.29464.16644