Module 5 Assignment – Text Classification

Professor: Justin Grosz

Course: ALY6020: Predictive Analytics (CRN 20466)

Submitted by :

Sameeksha Bellavara Santhosh

Date: 11-02-2024

Introduction

Building a model that can forecast the numbers students will write on a writing test is the aim of this project, which will help identify students who may require additional support for the development of their motor skills. The research evaluates the effectiveness of two methods: neural networks, a more complex approach, and K-Nearest Neighbours (KNN), a more basic approach. Pixel values are handwritten numbers that have been tagged and are included in the dataset. This paper examines the challenges, correctness, and construction and assessment procedures of KNN and neural network models. The study's findings may aid the school in determining which pupils require particular therapies to develop their motor skills.

Data Pre-processing:

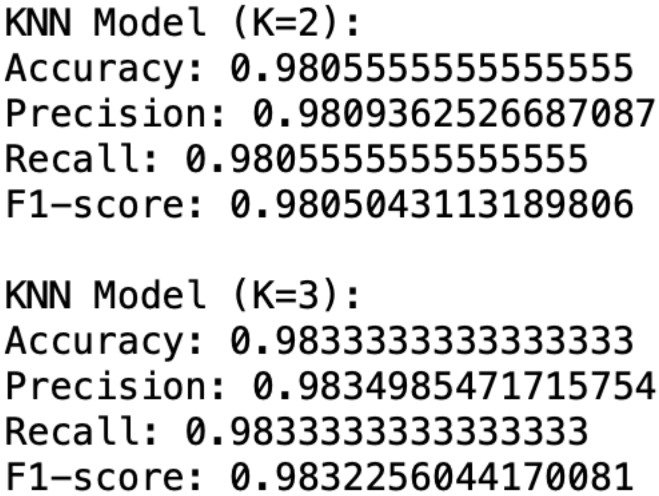
The dataset that was used in this study to predict handwriting was clear, well-organized, and devoid of any inappropriate variables or missing values. This made the laborious data-cleaning procedures needless. The dataset contained all relevant information.

We can jump right into the modelling phase because the dataset is already clean and well-formatted, negating the need for extensive data cleaning or pre-treatment procedures. The models can focus on accurately predicting and classifying the handwritten digits or characters using the given pixel attributes because the dataset is clean.

The code implements the Elbow Method to determine the optimal number of clusters (K) for k-mode clustering. It iterates over different values of K, fits the k-modes model for each K, computes the cost (sum of distances to centroids), and plots the cost against K to identify the "elbow" point, indicating the optimal K value.

Part 1: Build a KNN Model

Classification Report for the KNN model:



```
KNN Model (K=2):
Accuracy: 0.9805555555555555
Precision: 0.9809362526687087
Recall: 0.9805555555555555
F1-score: 0.9805043113189806

KNN Model (K=3):
Accuracy: 0.9833333333333333
Precision: 0.9834985471715754
Recall: 0.9833333333333333
F1-score: 0.9832256044170081
```

The results of the handwriting prediction KNN model indicate accuracy rates of 98.05% for K=2 and 98.33% for K=3. Accuracy reflects the overall correctness of the model's predictions. These results suggest that the KNN model performs adequately in accurately predicting handwriting.

Challenges Encountered in the KNN Model:

Determining Optimal K Value: Selecting the most suitable value for K poses a challenge. While I assessed two different K values, other values could potentially yield more accurate results. Careful experimentation and validation are essential when determining the appropriate K value.

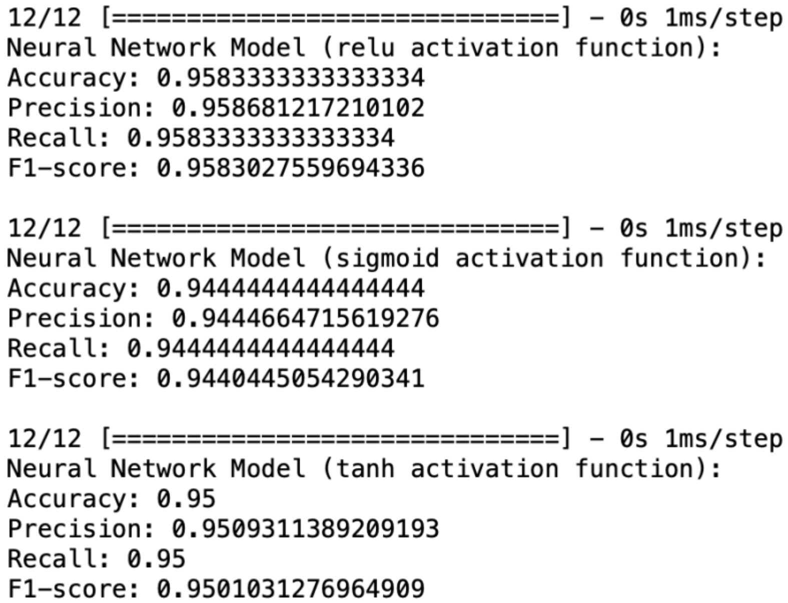
To explore different levels of complexity in the KNN model, I experimented with K values of 2 and 3. The K value represents the number of nearest neighbours considered for classification. A lower K value, such as 2, tends to be more flexible and may risk overfitting by relying heavily on a small number of neighbours. Conversely, a higher K value, like 3, results in a smoother decision boundary and reduces the impact of individual noisy samples, potentially leading to a more generalised model.

Feature Scaling: KNN assumes that all features are equally relevant for prediction. However, certain pixel positions or patterns may hold more significance in handwriting recognition. This lack of feature weighting could limit KNN's effectiveness in accurately capturing the nuances of handwriting recognition.

To address this, scaling the pixel values to a range between 0 and 1 by multiplying them by 255.0 is essential. Feature scaling is crucial for KNN, as it is a distance-based algorithm, to prevent features with larger ranges from dominating distance calculations. Normalisation ensures that each feature contributes fairly to the distance metric, thereby avoiding biased results.

Part 2: Build a Neural Network model

Classification report of Neural Network model:

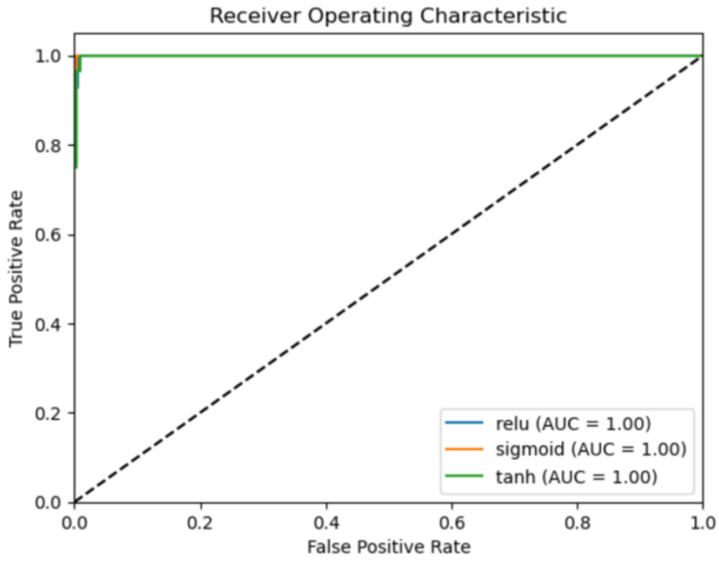


```
12/12 [==============================] - 0s 1ms/step
Neural Network Model (relu activation function):
Accuracy: 0.9583333333333334
Precision: 0.958681217210102
Recall: 0.9583333333333334
F1-score: 0.9583027559694336

12/12 [==============================] - 0s 1ms/step
Neural Network Model (sigmoid activation function):
Accuracy: 0.9444444444444444
Precision: 0.9444664715619276
Recall: 0.9444444444444444
F1-score: 0.9440445054290341

12/12 [==============================] - 0s 1ms/step
Neural Network Model (tanh activation function):
Accuracy: 0.95
Precision: 0.9509311389209193
Recall: 0.95
F1-score: 0.9501031276964909
```

The neural network model demonstrated accuracies ranging from 94% to 95% across various activation functions (ReLU, sigmoid, and tanh), indicating its capability to accurately categorize and predict handwriting samples. Higher accuracy suggests that the model effectively extracts patterns and features from the input data, leading to correct predictions.



Challenges Encountered in the Neural Network Model:

Activation Function Selection: Opting for the appropriate activation functions poses a significant challenge. The model's performance can be influenced by the distinct characteristics of activation functions. In this scenario, RELU and sigmoid activation functions demonstrated slightly higher accuracy compared to the tanh activation function. This highlights the importance of exploring and experimenting with various activation functions to determine the most suitable one for the specific task.

Memory Error: The occurrence of memory errors is notable, particularly when the program exhausts all allocated memory. This issue commonly arises when handling large datasets or executing memory-intensive tasks, such as training neural networks with numerous parameters.

Part 3: Comparison between KNN and Neural Network model

Based on the provided evaluation metrics for both the KNN and neural network models:

KNN Model:
K=2: Accuracy of 98.06%, Precision of 98.09%, Recall of 98.06%, and F1-score of 98.05%
K=3: Accuracy of 98.33%, Precision of 98.35%, Recall of 98.33%, and F1-score of 98.32%

Neural Network Model:
ReLU Activation: Accuracy of 95.83%, Precision of 95.87%, Recall of 95.83%, and F1-score of 95.83%

Sigmoid Activation: Accuracy of 94.44%, Precision of 94.45%, Recall of 94.44%, and F1-score of 94.40%
Tanh Activation: Accuracy of 95.00%, Precision of 95.09%, Recall of 95.00%, and F1-score of 95.01%

Comparing the models, both the KNN and neural network models achieve high accuracy rates, with the KNN model slightly outperforming the neural network models. However, the neural network models also demonstrate respectable performance, especially considering the complexity of the handwriting prediction task.

Considering the ease of implementation and interpretability, the KNN model may be preferred in this scenario, as it provides competitive performance while being straightforward to understand and use. However, if there is a need for further improvement in accuracy and the resources for training and maintaining a neural network model are available, the neural network models with ReLU or tanh activation functions could be considered.

Ultimately, the choice between the KNN and neural network models depends on factors such as the specific requirements of the school, the available resources, and the desired balance between accuracy and complexity.

Finding and Recommendations:

The findings from the KNN and neural network models can be summarized as follows:

KNN Model:

- Achieved high accuracy, precision, recall, and F1-score values for both K=2 and K=3.
- Demonstrated strong performance in accurately predicting handwriting based on the provided dataset.
- The simplicity and ease of interpretation make the KNN model suitable for straightforward classification tasks like handwriting recognition.

Neural Network Model:

- Showed competitive accuracy rates across different activation functions (ReLU, sigmoid, and tanh).
- Despite slightly lower accuracy compared to the KNN model, the neural network models exhibited respectable performance in categorizing handwriting samples.
- The neural network models demonstrated the ability to capture intricate patterns and features in the data, contributing to accurate predictions.

Conclusion

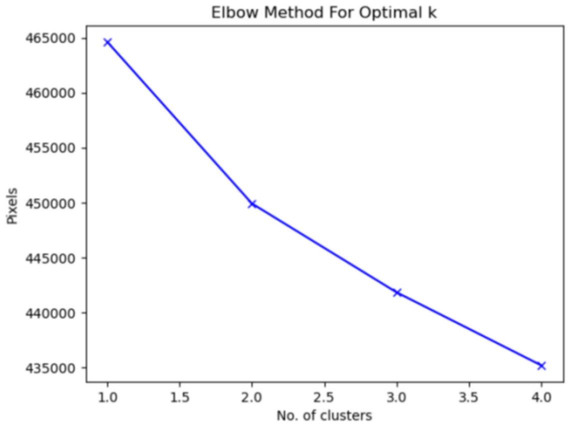
The KNN and neural network models perform admirably when it comes to handwriting sample prediction. The KNN model is appropriate for situations where simple answers are preferred because of its simplicity and ease of interpretation. However, despite its higher complexity and resource requirements, the neural network model offers flexibility and the possibility of even greater accuracy improvement. A recommendation to employ the KNN model due to its competitive performance and simplicity could be made based on the results and the particular requirements of the school. However, further research into the neural network approach with suitable activation functions may be necessary if there is a need for increased accuracy or the capacity to handle more complex models. The decision between the two models should ultimately take into account aspects like interpretability, computational capacity, and accuracy requirements.

References

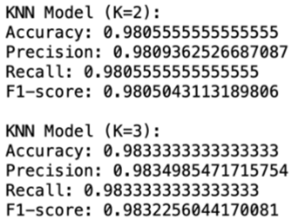
1. *sklearn.neighbors.KNeighborsClassifier*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
2. Shafi, A. (2023, February 20). *K-Nearest Neighbors (KNN) Classification with sci-kit-learn*. https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn
3. Kumar, G. S. (2022, September 28). *Understanding and building neural network (NN) models*. Built In. https://builtin.com/machine-learning/nn-models
4. *Text Classification: What it is And Why it Matters*. (n.d.). MonkeyLearn. https://monkeylearn.com/text-classification/

Appendix

Elbow Method for Optimal K:



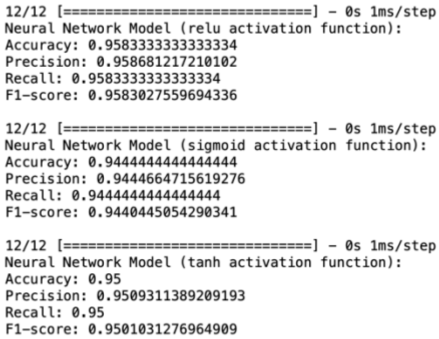
KNN MODEL:



```
KNN Model (K=2):
Accuracy: 0.9805555555555555
Precision: 0.9809362526687087
Recall: 0.9805555555555555
F1-score: 0.9805043113189806

KNN Model (K=3):
Accuracy: 0.9833333333333333
Precision: 0.9834985471715754
Recall: 0.9833333333333333
F1-score: 0.9832256044170081
```

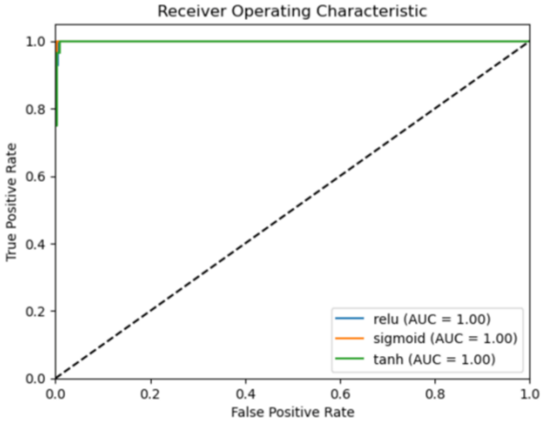
NEURAL NETWORK MODEL:



```
12/12 [==============================] - 0s 1ms/step
Neural Network Model (relu activation function):
Accuracy: 0.9583333333333334
Precision: 0.958681217210102
Recall: 0.9583333333333334
F1-score: 0.9583027559694336

12/12 [==============================] - 0s 1ms/step
Neural Network Model (sigmoid activation function):
Accuracy: 0.9444444444444444
Precision: 0.9444664715619276
Recall: 0.9444444444444444
F1-score: 0.9440445054290341

12/12 [==============================] - 0s 1ms/step
Neural Network Model (tanh activation function):
Accuracy: 0.95
Precision: 0.9509311389209193
Recall: 0.95
F1-score: 0.9501031276964909
```


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
relu (AUC = 1.00)
sigmoid (AUC = 1.00)
tanh (AUC = 1.00)