جامعة بيرزيت

BIRZEIT UNIVERSITY

Faculty of Engineering Technology

Electrical & Computer Engineering Department

ENCS3340, ARTIFICIAL INTELLIGENCE

Project 2 Report

Prepared by:

Faten Sultan	**ID Number:** 1202750
Mohammed Abu Shams	**ID Number:** 1200549

Instructor: Dr. Yazan Abu Farha

Date: 14.7.2023

Section: 2 && 4

Introduction

The persistent issue of spam classification within the realm of text processing and information retrieval has found potential resolution strategies in machine learning algorithms, notably K-Nearest Neighbors (KNN) and Multi-Layer Perceptron (MLP). The objective of this research is to juxtapose these two algorithms in the context of spam detection, applying a dataset derived from various features amassed from a multitude of emails.

Methodology

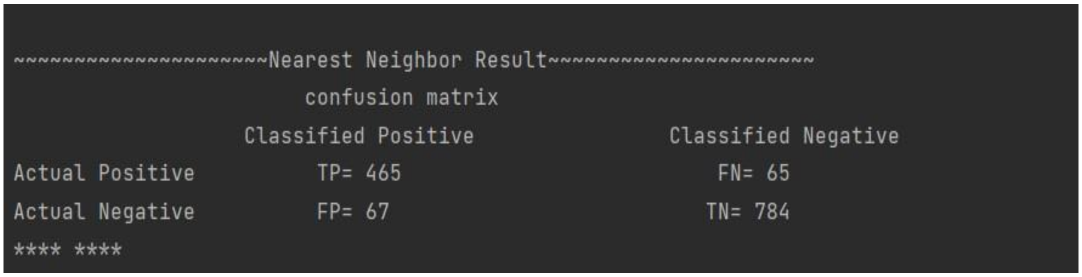
In this project we used KNN algorithm. This algorithm works on a proximity principle, allotting classifications to new instances derived from the majority class of their 'k' closest neighbors in the feature domain which is equal to 3. In contrast, MLP, as a form of neural network, leverages neuron layers, activation functions, and back propagation to discern intricate patterns in the data.

The initial step was to preprocess the dataset, normalizing each feature by deducting the mean and dividing by the standard deviation to ensure an equal contribution from all features towards the final verdict. This processed data was divided in a way that makes 70% for training while the rest 30% for testing sets this ratio is gauge for the models' efficacy.

Results and Discussion

For Nearest Neighbor

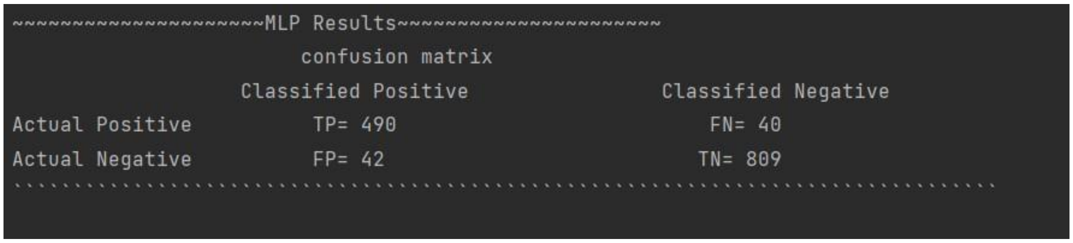
The confusion matrix is:



```
~~~~~~~~~~~~~~~~~~~~~Nearest Neighbor Result~~~~~~~~~~~~~~~~~~~~~~
                        confusion matrix
                   Classified Positive                Classified Negative
Actual Positive          TP= 465                          FN= 65
Actual Negative          FP= 67                          TN= 784
**** ****
```

For MLP

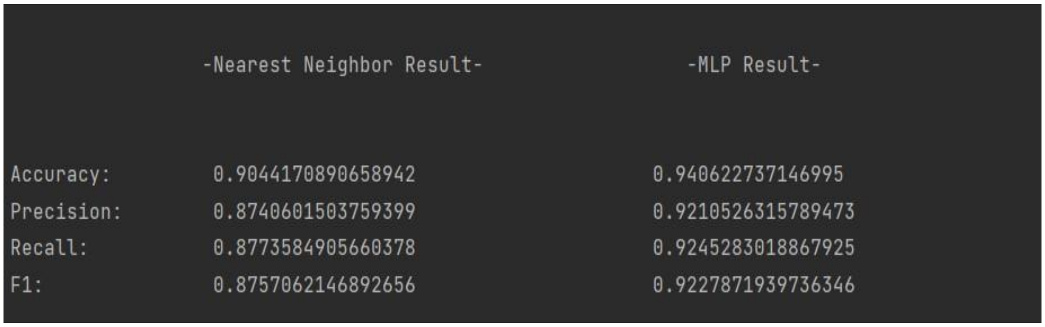
the confusion matrix is:



```
~~~~~~~~~~~~~~~~~~~~~MLP Results~~~~~~~~~~~~~~~~~~~~~~
                        confusion matrix
                   Classified Positive                Classified Negative
Actual Positive          TP= 490                          FN= 40
Actual Negative          FP= 42                          TN= 809
``````````````````````````````````````````````````````````````````````````````````
```
```

The result for Accuracy, precision, recall, f1

The following is the results for accuracy, precision, recall, f1 for both algorithms:



```
                 -Nearest Neighbor Result-                 -MLP Result-


Accuracy:         0.9044170890658942                0.940622737146995
Precision:        0.8740601503759399                0.9210526315789473
Recall:           0.8773584905660378                0.9245283018867925
F1:               0.8757062146892656                0.9227871939736346
```

We can check our results from the following rules:

Recall= $\frac{TP}{TP+FN}$ Precision= $\frac{TP}{TP+FP}$

Accuracy= $\frac{TP+TN}{TP+FN+FP+TN}$ F1= $\frac{2\ast Accuracy\ast Precision}{Accuracy+Precision}$

- From the obtained output we can concluded that the results of MLP is better than NN

Experiments

Multi experiments have applied to comprehend how various factors impacted the performance of the models. For the KNN algorithm, we tested it with different 'k' values, and for the MLP model, it was experimented with divergent architectures.

- KNN algorithm: is a supervised machine learning algorithm that can be used for both classification and prediction of the current output. This can be done by given value of K that determines the number of neighbors that are considered for prediction. A smaller K value may lead to more flexible decision boundaries, but it can also make the algorithm sensitive to noise. A larger K value may provide smoother decision.

- The MLP algorithm known as Multi-Layer Perceptron, it is one of artificial neural network (ANN) commonly used in machine learning for both classification and regression tasks. MLPs is used to model complex relationships between inputs and outputs. Once the MLP is trained, it can be used to make predictions on new, unseen data. The input data is passed through the network, and the output is obtained by forward propagation.

Improvements and Future Work

There is potential for enhancing both algorithms. For instance, the MLP model can be improved by adding more layers or neurons or by experimenting with diverse activation functions. While for the KNN model, we can explore different distance metrics or changing the number of neighbors. Moreover, feature selection methodologies could assist in reducing the dataset's dimensionality and possibly enhance the model's performance.

Conclusion

This project shows the different analysis of the KNN and MLP models for the context of spam detection. The results shows the importance of selecting a suitable model based on the data and problem context, and also underscore potential areas for further exploration and refinement.