A Comparative Study on the Accuracy of Various ML Techniques in Detecting DDoS Attacks

By

Benjamin Dempsey

Student No.: 230150566

Under the supervision of

Dr. Saha

Department of Computer Science

University of Northern British Columbia

3333 University Way, Prince George BC, Canada

Contents

CHAPTER 1

ABSTRACT

While distributed denial of service (DDoS) attacks are on the rise, methods of detecting and preventing attacks are also making strides. A common method of DDoS detection involves using various machine learning (ML) techniques to identify commonalities between packets that are malicious in nature to detect and prevent them from reaching the target. This paper compares and contrasts the effectiveness of using Multi-layer Perceptron (MLP) Classifiers, Decision Trees, and Logistic Regression on the CICDDoS2019 dataset. The conclusion of the experiments showed that Logistic Regression performed the worst with 98.0% success rate. The MLP Classifier had improved results with 99.0% accuracy. Finally, the decision tree outperformed all methods with a 100.0% success rate in identifying malicious packets.

CHAPTER 2

METHODOLOGY

The experiment was split into five distinct steps as follows:

1. Dataset selection:

2. Data preprocessing

3. Feature selection

4. Model training

5. Result comparision

Each step will be explained in detail in this section.

Dataset Selection

A literature review of similar works (Golduzian, 2023) showed that the most complete and thorough dataset was CICDDoS2019 from University of New Brunswick. The dataset has 88 total features including source IP, destination IP, packet length, etc with each row labeled. This proved a good choice and made the preprocessing step simple. This experiment was repeated on the DNS data and UDP data.

Data Preprocessing

All preprocessing was done in Python using Google Collab and a number of data science libraries. The first step was reducing the 2GB csv file to a more manageable 32MB by reducing the number of rows to 80000. Since most ML techniques exclusively work with numbers, the labels were converted so that the 'BENIGN' label became zero and 'DDoS' became 1. All other non-numeric attributes were converted to equivalent numeric types. All NaN became zeros and all infinities became MAX_INT. For the MLP the data's range was reduced to fall between 0 and 1 by taking the original value x and applying the equation $x' = \frac{x}{x_{max}}$ where x_{max} is the maximum value in the column.

Feature Selection

Columns with inconsistent data types were dropped entirely. A Variance Threshold method was used to drop attributes that varied too greatly. This reduced the number of attributes from 88 to 66.

Model Training

For each of the three ML techniques used (MLP, Logistic Regression, Decision tree), the dataset was split into two subsets. The first contained everything but the label, and the second contained just the corresponding label for each row. These sets were split again into training and testing data with 75% being used for training and the remaining 25% used for testing. Models were trained and then tested on the dataset.

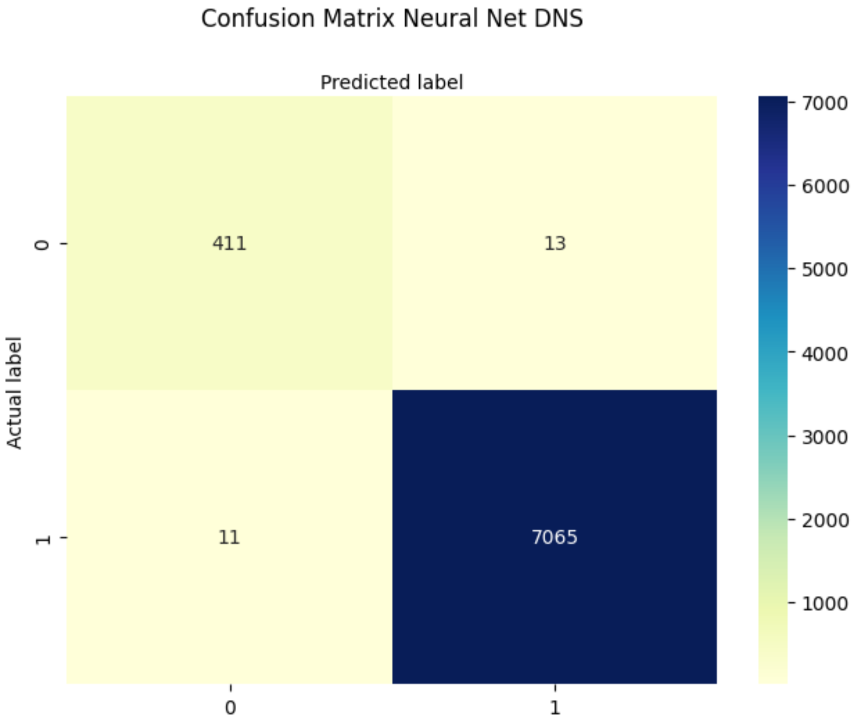
Result Comparison

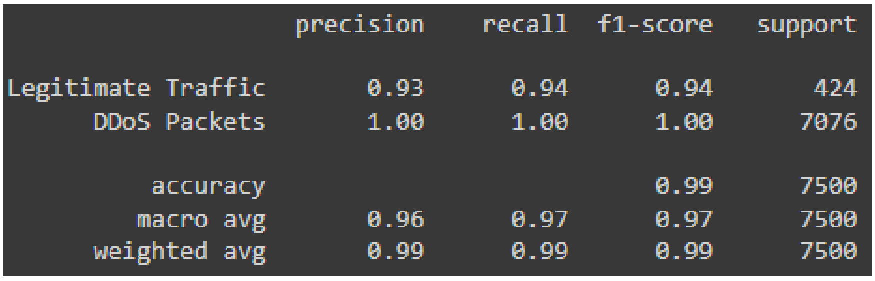
The results from the three methods were displayed as a confusion matrix, and their precision, f1-score, recall, and accuracy were displayed. These results can be found in the next section.

CHAPTER 3

RESULTS AND CONCLUSION

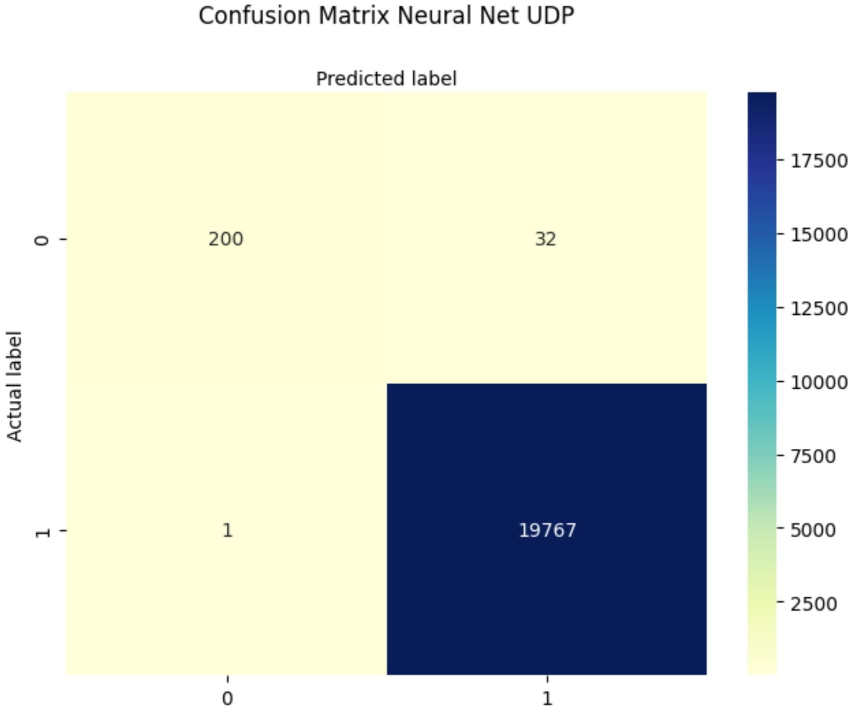
Multi-layer Perceptron Classifier:

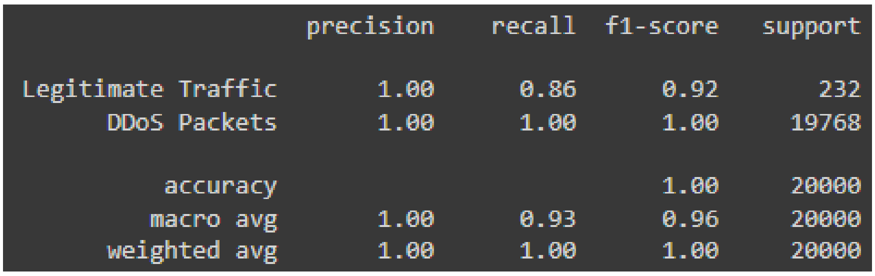




```
                   precision    recall  f1-score   support

Legitimate Traffic      0.93      0.94      0.94       424
      DDoS Packets      1.00      1.00      1.00      7076

          accuracy                          0.99      7500
         macro avg      0.96      0.97      0.97      7500
      weighted avg      0.99      0.99      0.99      7500
```

Figure 3.1: Results from the MLP DNS trials

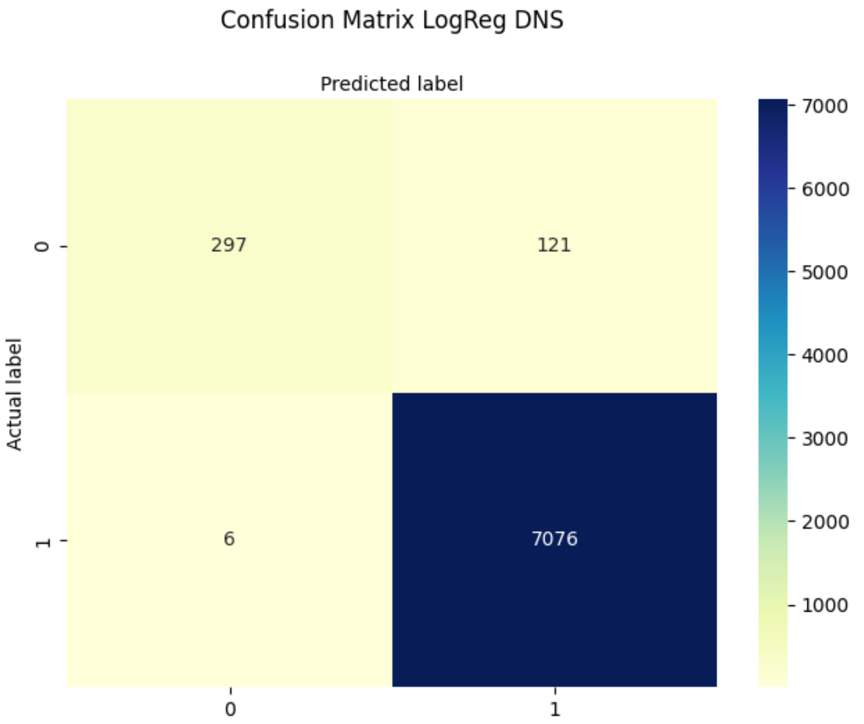


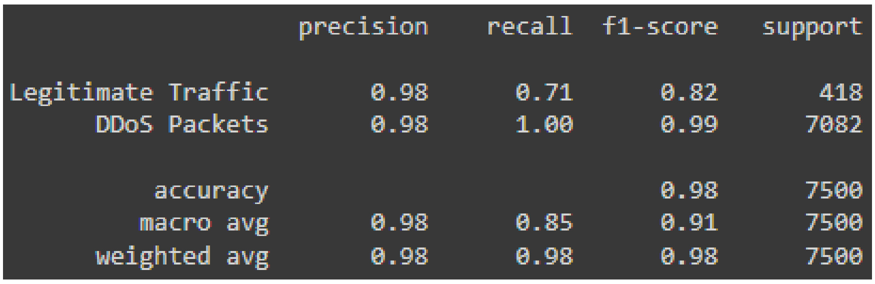


```
                   precision    recall  f1-score   support

Legitimate Traffic      1.00      0.86      0.92       232
      DDoS Packets      1.00      1.00      1.00     19768

          accuracy                          1.00     20000
         macro avg      1.00      0.93      0.96     20000
      weighted avg      1.00      1.00      1.00     20000
```

Figure 3.2: Results from the MLP UDP trials

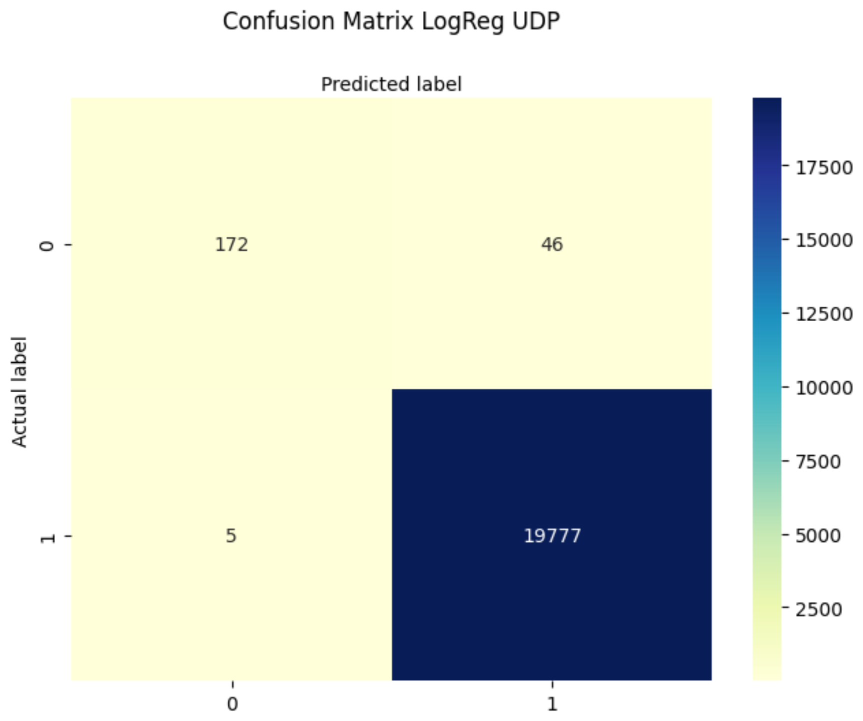
Logistic Regression Classifier:

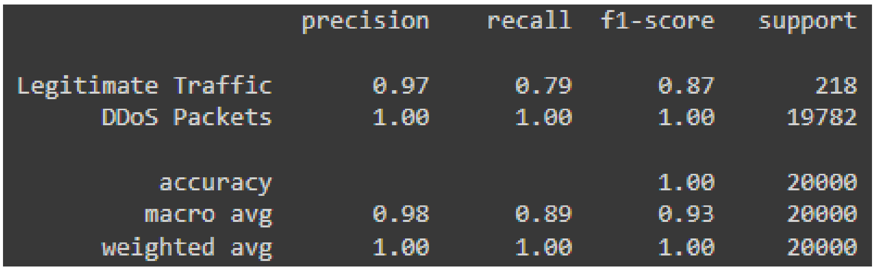




```
                   precision    recall  f1-score   support

Legitimate Traffic       0.98      0.71      0.82       418
      DDoS Packets       0.98      1.00      0.99      7082

          accuracy                           0.98      7500
         macro avg       0.98      0.85      0.91      7500
      weighted avg       0.98      0.98      0.98      7500
```

Figure 3.3: Results from the Logistic Regression DNS trials

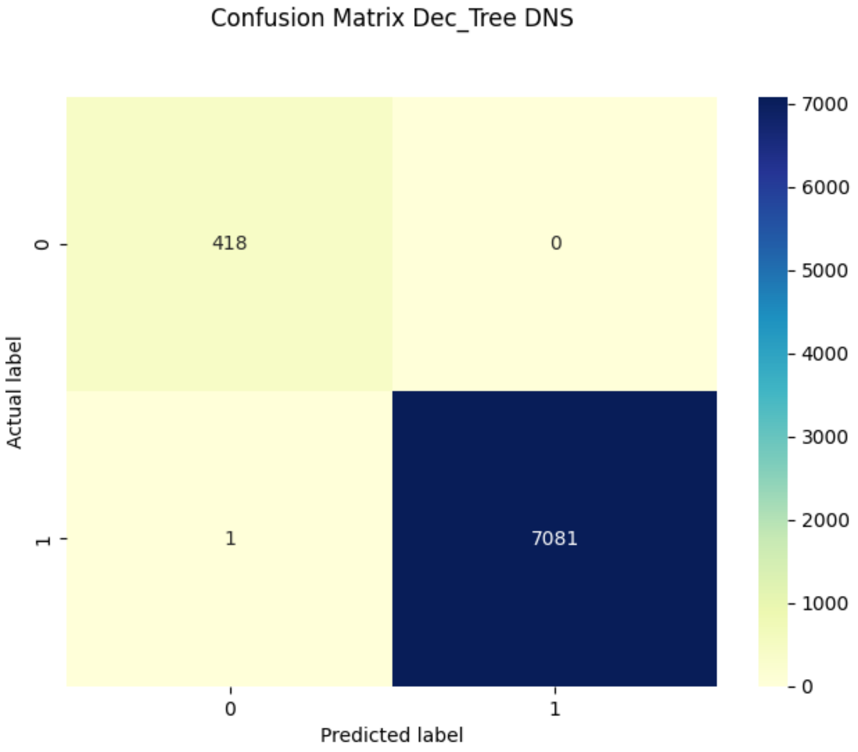


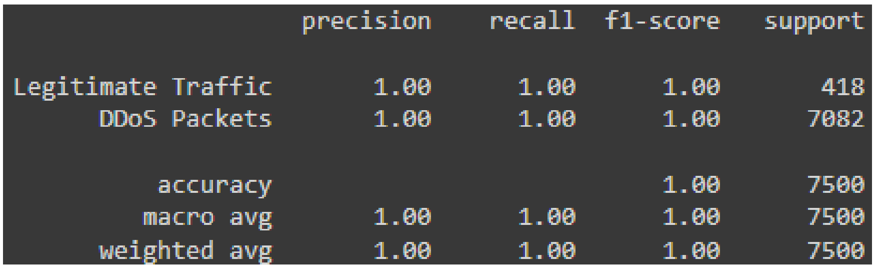


```
                   precision    recall  f1-score   support

Legitimate Traffic      0.97      0.79      0.87       218
      DDoS Packets      1.00      1.00      1.00     19782

          accuracy                          1.00     20000
         macro avg      0.98      0.89      0.93     20000
      weighted avg      1.00      1.00      1.00     20000
```

Figure 3.4: Results from the Logistic Regression UDP trials

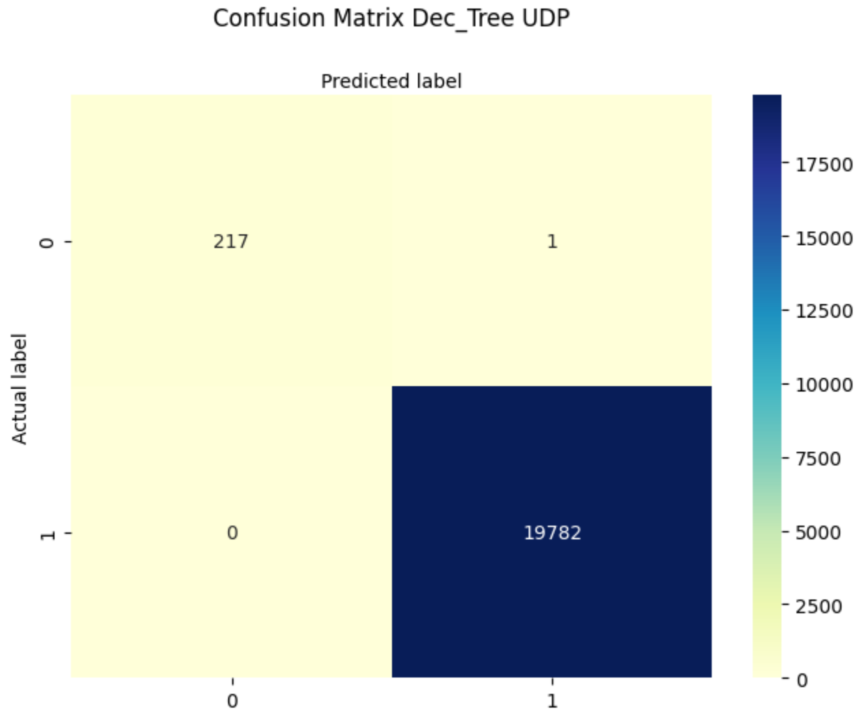
Decision Tree:

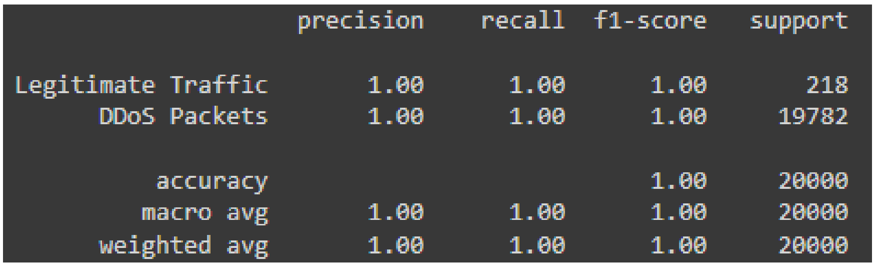




```
                    precision    recall  f1-score   support

Legitimate Traffic       1.00      1.00      1.00       418
      DDoS Packets       1.00      1.00      1.00      7082

          accuracy                           1.00      7500
         macro avg       1.00      1.00      1.00      7500
      weighted avg       1.00      1.00      1.00      7500
```

Figure 3.5: Results from the Decision Tree DNS trials





```
                   precision    recall  f1-score   support

Legitimate Traffic       1.00      1.00      1.00       218
      DDoS Packets       1.00      1.00      1.00     19782

          accuracy                           1.00     20000
         macro avg       1.00      1.00      1.00     20000
      weighted avg       1.00      1.00      1.00     20000
```

Figure 3.6: Results from the Decision Tree UDP trials

Conclusion:

On the DNS data the MLP incorrectly identified 24 out of 7500 of the training packets with 11 being false negatives and 13 being false positives. This was much better than the logistic regression classifier model which incorrectly identified 127 out of 7500 packets, 6 of which were false negatives and 121 were false positives. The best performer was the decision tree which had only incorrectly identified 1 DNS packet out of 7500 with 1 false negative and 0 false positives.

On the UDP data the MLP incorrectly identified 33 out of 20000 of the training packets with 1 being false negatives and 32 being false positives. This was much better than the logistic regression classifier model which incorrectly identified 51 out of 20000 packets, 5 of which were false negatives and 46 were false positives. The best performer was the decision tree which, again, had only incorrectly identified 1 UDP packet out of 20000 with 0 false negatives and 1 false positive.

All models performed better on the larger UDP dataset however I believe this is largely due to the poor distribution of DDoS packets and benign packets. Since the vast majority (98.01%) of UDP packets were malicious DDoS packets the model's would have had a 98.01% accuracy if they had guessed the DDoS label every time. Future research should examine datasets with more even distributions of labels or datasets with mostly legitimate benign packets to more accurately reflect real world networks.

The decision tree model consistently gave the best results on this dataset with 99.99986% and 99.99995% accuracy. Out of the machine learning methods chosen I believe a decision tree was the best fit for this problem. Decision trees are great for classification and performed well with this thorough dataset.

CHAPTER 4 WORK CITED

“Search UNB.” University of New Brunswick est.1785. Accessed July 23, 2024. https://www.unb.ca/cic/datasets/ddos-2019.html.

Predict and prevent DDOS attacks using machine ... Accessed July 24, 2024. https://arxiv.org/pdf/2308.

Singh, Sumit. “ML Beginner’s Guide to Ddos Attack Detection Model.” Labellerr, April 14, 2024. https://www.labellerr.com/blog/ddos-attack-detection/#:̃:text=Machine %20learning%20algorithms%20play%20a,and%20mitigation%20of%20DDoS%20attacks.