

**TEXAS A&M INTERNATIONAL UNIVERSITY**

**College of Arts and Science**

Carlos E. Gonzalez

Mario Juarez

Mentor: Dr. Mustafa Al-Lail

**Abstract**: In the Internet era, the current society relies on internet technologies to carry out their daily activities and businesses. Any disturbance on the internet can have a significant impact on the well-being of society. Attackers can exploit the new vulnerabilities of these technologies. It is crucial to maintain computer networks safe and keep the integrity, confidentiality, and availability of services to the users. Security researchers have proposed many countermeasures to tackle the cyber-attacks with the surge of new networking paradigms. One promising technique is the use of Machine Learning (ML) to detect intrusion attacks. ML can analyze patterns in network traffic and classify them as benign or attacks. This gives ML models the ability to predict future attacks at a suitable time. We applied the ML techniques using the CIC-IDS 2017 dataset that reflects the most up-to-date attacks. This dataset contains over 2.2 million samples represented by 79 features/columns and the last column being the label. Our first step was to first explore the dataset using a method known as Exploratory Data Analysis (EDA), this method uses data visualization to draw meaningful patterns and insights. We then implemented a Multi-layer Perceptron regressor (MLP). Further, we applied two feature selection techniques: Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) to eliminate noise in the dataset and improve the performance and running time of the models.

**Intro**

The world of Information Technology has brought new paradigms such as Cloud Computing and the Internet of Things (IoT) [1]. These new services bring new requirements such as shared data, real-time response, and data security.

Data security is one of the most important requirements because new challenges have emerged demanding new security measures for the users [3]. There have been many attacks that have proven how much of a threat cyber-attack is nowadays. Attacks that have occurred in recent times are the LinkedIn attack that occurred in July of 2021, in which 700 million users were affected [2]. In August of 2019 Facebook had a data breach, in which users had their phone numbers, account names, and Facebook ID compromised [2].

The old methods are not useful with the new types of attacks, giving high false alarms and lacking a more dynamic adaptation. Machine Learning is one of the most researched methods for the process of intrusion detection in computer systems. This method uses computer algorithms to automatically improve its accuracy by using data as input. It adapts to new attacks making it a more robust Intrusion Detection System (IDE).

Our research is based on first analyzing the dataset using visualizations to summarize their main characteristics using EDA. Once we got a better understanding of the data we concluded that using Multi-layer Perceptron was one of the best options because of its effectiveness in gathering accurate information from an imbalanced dataset. Additionally, we used Principal Component Analysis and Recursive Feature Elimination as feature selection techniques. Our approach was to import the CIC-IDS 2017 dataset containing eight CSV files. We then concatenated all eight CSV files into one file. Once the file gets concatenated the data was cleaned from empty and infinite cells, leaving only cells containing data. The target values were independently encoded as enumerated data types. Data was split into train and test data using the train test split method. We then applied the scalar method MinMaxScaler from sklearn to normalize the range of features in the dataset. Lastly, we tested the models with unseen data and scored them to get the accuracy, precision, recall, and f-score. Our findings reflected an average a high recall of 90% and above.

This research paper is organized the following way. Related Work investigates other similar approaches to the same problem. Methodology explains the implementation of our supervised models. Results show a completely accurate solution to the model. The conclusion presents the findings and final details of our research paper.

**Related Work**

Research in 2020 done by Thaseen et. al. [4] on the use of machine learning algorithms such as Naïve Bayes, Random Forest, K-Nearest-Neighbors, and Support Vector Machine demonstrated the performance of each algorithm in detecting difference intrusions in a network. This research used accuracy, precision, recall, and f1-score as evaluation metrics. Random Forest was proven to be the superior algorithm in this research by predicting the attacks with 99.81 percent accuracy, precision, and recall on the dataset generated using the Wireshark tool. Other research has been done using different datasets with the same goal in mind similar to the research done by Çavuşoğlu [5]. The research was done using the NSL-KDD dataset, and the machine learning algorithms used were Naïve Bayes, Random Forest, Decision Tree, and K-Nearest-Neighbors. Çavuşoğlu’s research uses a combination of feature selection and machine learning algorithms to improve the detection of intrusion attacks. Alqahtani et. al. [6] research used an older data set but uses a different machine learning algorithm Artificial Neural Networks. This correlates to our research since we are using feature selection algorithms. This research shows that Artificial Neural Networks are a good algorithm but there are better ones for this type of problem that have higher accuracy. Taher et. al. [7] used an older dataset than the one used in our research but used the same machine learning algorithms, as well as using feature selection. This research used a Support Vector Machine and Artificial Neural Networks. Taher concluded that the best algorithm out of the two was Artificial Neural Networks with a wrapper feature selection that is comparable to our work. Master et. al. [8] use the CICIDS-2017 dataset which at this point is the most up-to-date dataset that includes the most modern type of intrusion attacks in a network. This research concluded that the best algorithm used was Decision Tree.

**Methodology**

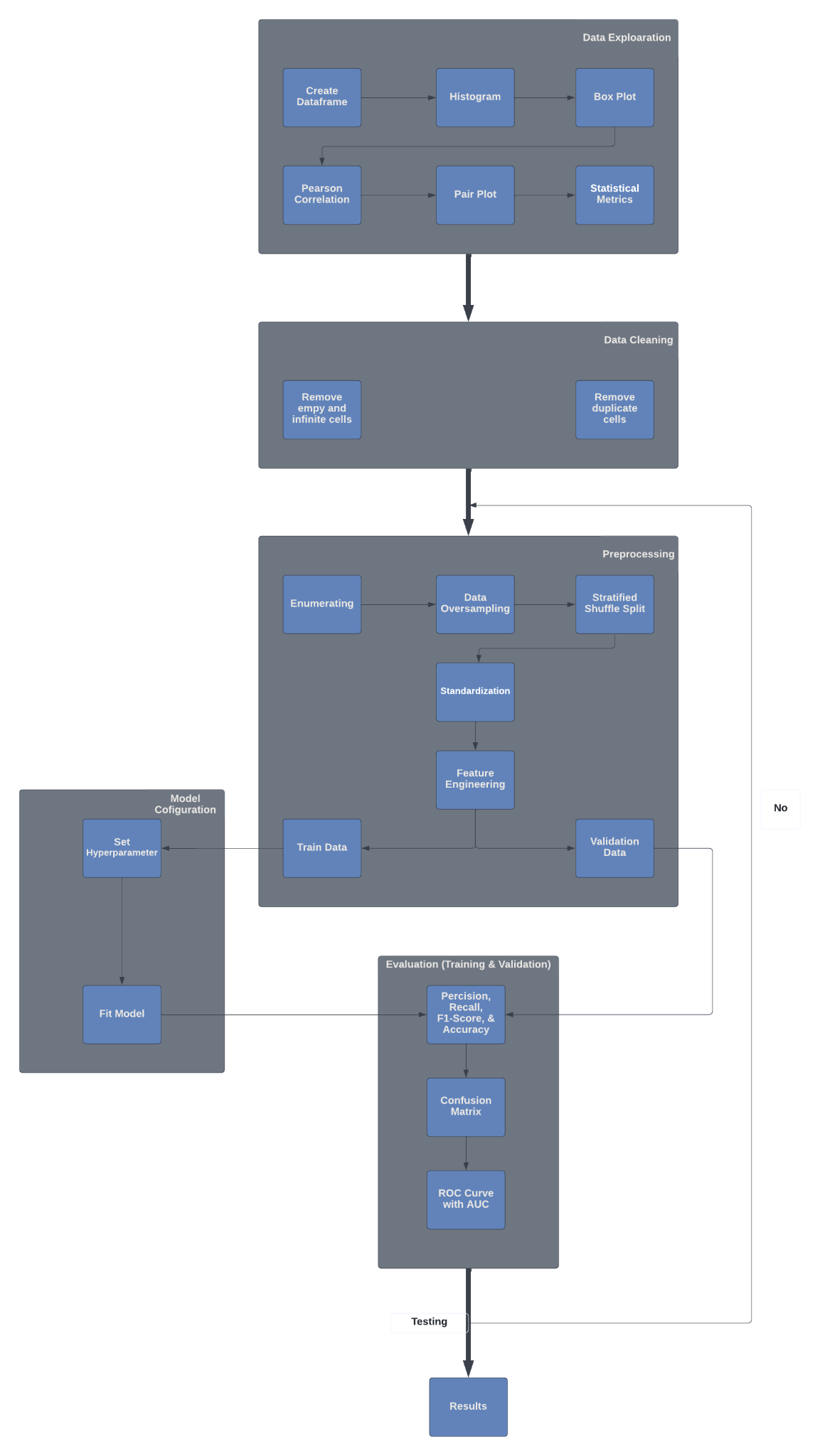
Our first approach was to clean the data by removing duplicates, empty, and inf cells to make our models error-free. Our next step was to create a dataframe and analyze the dataset using EDA method by gathering statistical metrics of each feature/column. Histograms for every feature were also collected to get a visual description of every feature. A pearson correlation was to identify the variables to identify what variables had the strongest relationship. To get a sense of how dispersed our attack and benign labels were, we used a pair plot and identified what algorithm would be best suited for capturing attack labels. This was achieved by using a combination of features and giving a visual representation to choose an algorithm that will perform well.

The next step was to split the data into 80% for validation and 20% for testing. We then preprocessed the dataset for our validation models, starting by encoding our data using the Label Encoder function to convert to machine readable form. This labeled all our label columns independently enumerating each label with its unique number. A stratified shuffle split merges StratifiedKFold and ShuffleSplit, which returns stratified randomized folds. This method preserves the percentage of samples for each class.

A normalization technique using MinMaxScaler was used to get the data points closer to each other for better model generalization.

Our model was now set and evaluated to get our classification report to show the precision, recall, F1 Score, and support. A confusion matrix was also used to measure recall, f1-score, and most importantly AUC-ROC curves.

Lastly a K-fold cross validation was performed to get more realistic results, giving us a more accurate data.

**** *Fig. 1 Methodology*

**Results**

Our results have shown that when trying to detect intrusion attacks using MLP/RFE, resulted in the best performing ROC curve out of our two classifications as shown in Figure 4.1.

For our first model when using the ROC curve to test our model. We can see that MLP with PCA was our worst model tested as shown in (Figure 1.1). MLP with PCA had the highest False Positive rate out of all the models tested. We also used the confusion matrix to see how our models were tested. For the MLP with PCA (Figure 1.2) we can see that the model did not predict correctly between classes. The model predicted that for class 0, 240 were class 3. It was only able to predict 180 correct for class 0 and for the rest of the classes the correct predictions were either too low or just zero. This shows that MLP with PCA is our worst performing model thus far. Our classification report shows us that MLP and PCA (Table 1) is still the worst-performing algorithm we have tested. All classes had below-average precision, recall, and F1-Score. Class 3 had the worst results scoring zero for precision, recall, and F1 score.

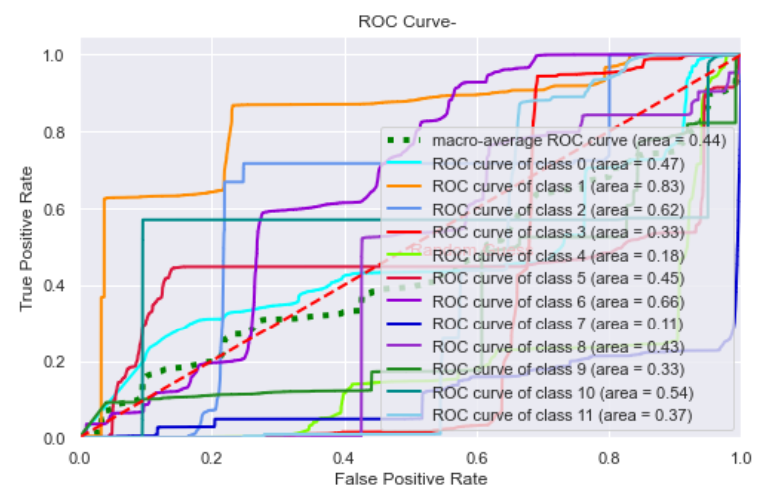


Figure 1.1 Multi-Layer Perceptron with Principal Component Analysis (ROC Curve)

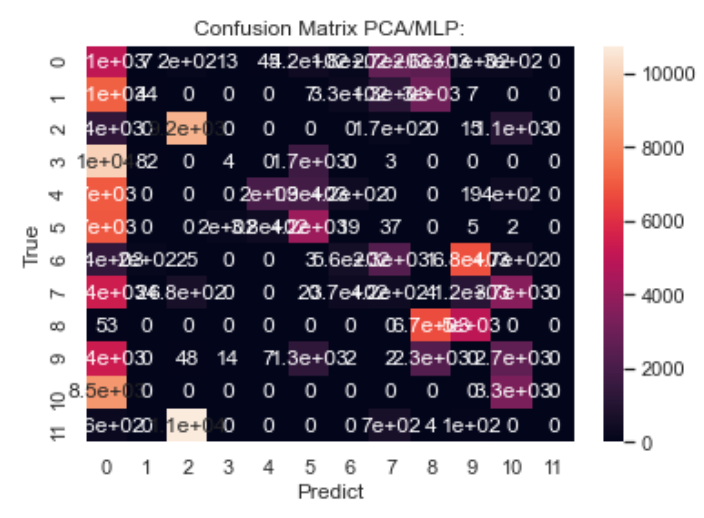


Figure 1.2 Multi-Layer Perceptron with Principal Component Analysis (Confusion Matrix)

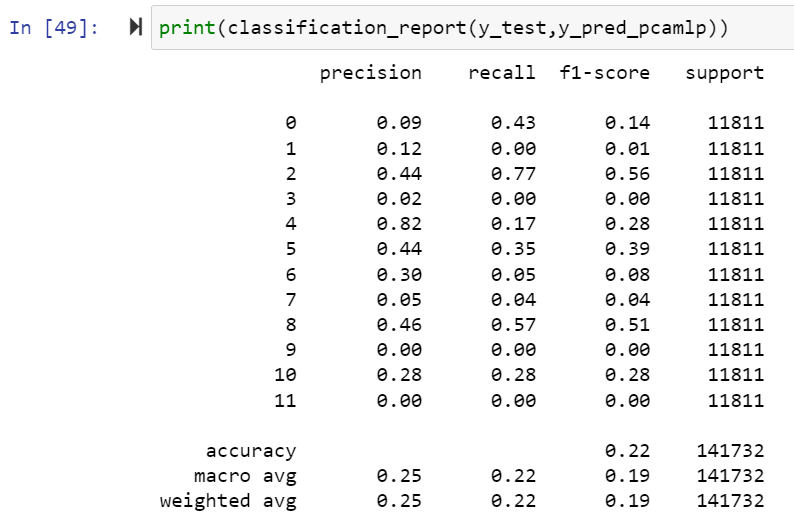


Table 1 Multi-Layer Perceptron with Principal Component Analysis (Classification Report)

The second model we tested was MLP with RFE (shown in Figure 2.1). This model was our best performing model using the ROC curve. It had the least number of false positives detected. When comparing this model (Figure 2.1) to our previous model (Figure 1.1) this model had a 20% increase in performance overall. When comparing each class individually class 0 had a 3% increase, class 1 had a 27% increase, class 2 had a 25% increase, and class 3 had a 25% increase in detecting True positives. Confusion matrix showed us a visual representation of Precision and F1-score, except Recall.

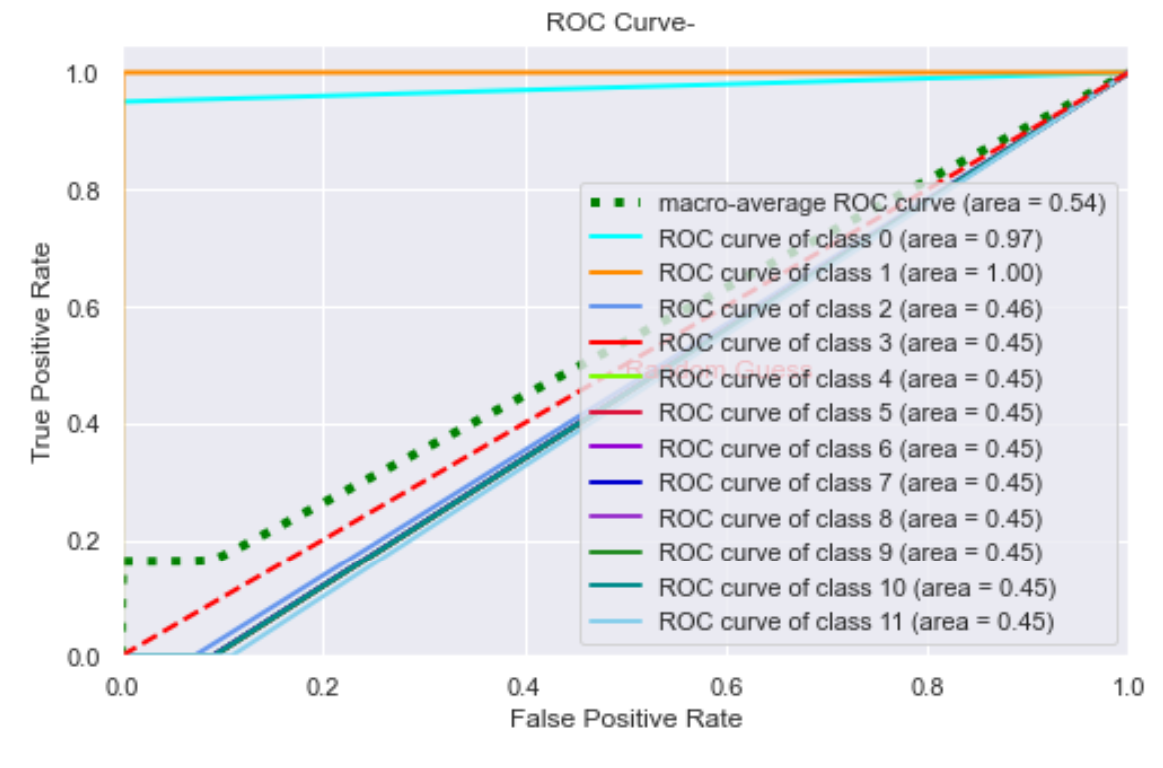


Figure 2.1 Multi-Layer Perceptron with Recursive Feature Elimination (ROC Curve)

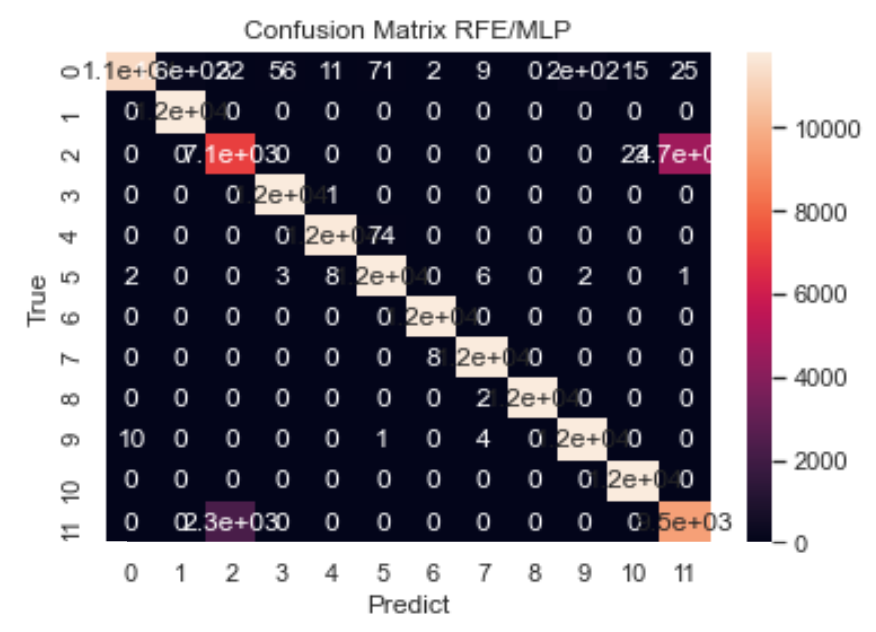


Figure 2.2 Multi-Layer Perceptron with Recursive Feature Elimination (Confusion Matrix)

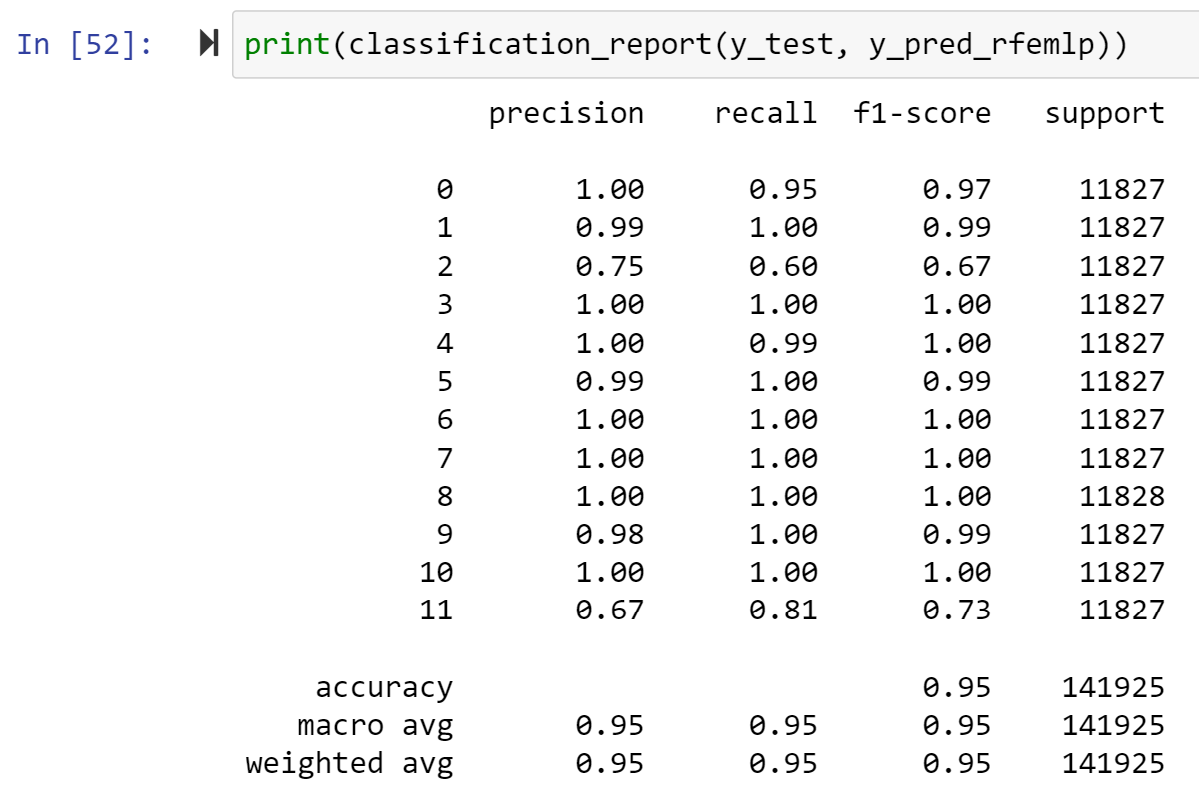


Table 2 Multi-Layer Perceptron with Recursive Feature Elimination (Classification Report)

**Conclusion**

This paper shows the design, implementation, and evaluation of a Multi-Layer Perceptron in combination with Recursive Feature Elimination or Principal Component Analysis with the intention of demonstrating their performance to the task of intrusion detection in computer networks. To accomplish this, we used the CIC-IDS 2017 data set corresponding to validation

and testing, to evaluate the performance of the model against known and unknown entries. The results indicate that MLP with PCA feature selection technique is not a good model for this dataset. Model MLP with RFE is the one that better adjusts to the task of

intrusion detection in computer networks since it shows the best results, making it the most reliable model for the task. Our future work will be to try other high performing algorithms using our EDA results. Lastly, we would also implement different hybrid models, making use of other algorithm combinations to compare results with our current models.

References

[1] P. Mell and T. Grance, “The NIST Definition of Cloud Computing,” 2011.

[2] Sep 2 7 Most Infamous Cloud Security Breaches IT News. (2021, September 2). *7 most infamous cloud security breaches - storagecraft*. StorageCraft Technology, LLC. Retrieved February 4, 2022, from https://blog.storagecraft.com/7-infamous-cloud-security-breaches/

[3] D. Bhattacharyya and J. Kalita, *Network Anomaly Detection: A Machine Learning Perspective*. CRC Press, 2013.

[4] Thaseen, I. S., Poorva, B., & Ushasree, P. S. (2020, February). Network intrusion detection using machine learning techniques. In *2020 International conference on emerging trends in information technology and engineering (IC-ETITE)* (pp. 1-7). IEEE.

[5] Çavuşoğlu, Ü. (2019). A new hybrid approach for intrusion detection using machine learning methods. *Applied Intelligence*, *49*(7), 2735-2761.

[6] Alqahtani, H., Sarker, I. H., Kalim, A., Hossain, M., Md, S., Ikhlaq, S., & Hossain, S. (2020, March). Cyber intrusion detection using machine learning classification techniques. In *International Conference on Computing Science, Communication and Security* (pp. 121-131). Springer, Singapore.

[7] Taher, K. A., Jisan, B. M. Y., & Rahman, M. M. (2019, January). Network intrusion detection using supervised machine learning technique with feature selection. In *2019 International conference on robotics, electrical and signal processing techniques (ICREST)* (pp. 643-646). IEEE.

[8] Maseer, Z. K., Yusof, R., Bahaman, N., Mostafa, S. A., & Foozy, C. F. M. (2021). Benchmarking of machine learning for anomaly-based intrusion detection systems in the CICIDS2017 dataset. *IEEE access*, *9*, 22351-22370.