TCP/IP Network Analyzer

Patel Yagnesh
Ahmedabad University
Ahmedabad, India
yagnesh.p@ahduni.edu.in

Kavan Gondalia
Ahmedabad University
Ahmedabad, India
kavan.g@ahduni.edu.in

Tej Thakar
Ahmedabad University
Ahmedabad, India
tej.t@ahduni.edu.in

Nityam Dixit
Ahmedabad University
Ahmedabad, India
nityam.d@ahduni.edu.in

Abstract – Many research have lately been conducted on network traffic analysis and forecasting, which has a wide range of applications. Many experiments are conducted, reported on, and used to address a variety of concerns with contemporary computer network applications. In this project, the TCP network dataset is the main emphasis, and we use a variety of techniques to try to forecast the application. Application prediction is a critical task in network security and monitoring since it aids managers in determining the nature of traffic flowing over the network and can be further implemented for selecting an appropriate congestion control protocol. The goal of this project is to assess the effectiveness of various algorithms for predicting the application using TCP network data, including Random Forest, Logistic Regression, and Decision Tree. Accuracy of the models will be examined. Also, the measure of different features will be measured. The project's objective is to gain a deeper understanding of the algorithms' effectiveness in predicting the application from TCP network data and to identify the features that significantly affect the accuracy of the prediction model.

Keywords - TCP/IP, Computer Networks, Classification, Applications

I. INTRODUCTION

The objective of the project is to propose a machine learning approach for application prediction based on network traffic patterns. The prediction model in the dataset used is based on DPI (Deep Packet Inspection) with ntop. The features we observed from the flows included time, transport-layer data, packet and byte counts, and so on. Overall, it has 87 features and 75 classes of applications and a total 35,77,296 instances. Random Forest and Logistic Regression are two machine learning models that have been built in this study to predict the application. Accuracy and F1 scores are two of the metrics used to determine the accuracy of the models.

II. LITERATURE SURVEY

This research suggests a machine learning-based method for identifying Applications. The paper proposes the use of several linear and non-linear machine learning methods, such as decision trees, support vector machines, and k-nearest neighbours, on a dataset of network traffic recorded from a local network for classification. Analysis and Prediction of Real Network Traffic suggest the use of a different linear model that can be used for predicting the flow of the network. According to the results, the suggested method outperforms standard signature-based network classification.

III.METHODOLOGY

The project aims to predict application names from the TCP network dataset. This is mainly divided into four stages: data preprocessing, feature selection, model building and model evaluation.

1. Data preprocessing - Data cleaning and preprocessing are done to make it usable for the machine learning algorithms. In this removal of irrelevant or missing data, also conversion of non-numeric data to numeric form.

2. Feature Selection - We selected some features to improve our machine learning algorithms in order to get more accuracy.We removed some features such as Flow.ID,Source.IP,Label,Timestamp and Destination.IP from our dataset to enhance the accuracy.

3. Model Building - We used different machine learning algorithms such as logistic regression, random forest for prediction.Preprocessed dataset is used to train the models. For binary classification we used logistic regression and for multi-class classification we used random forest in order to predict if a packet belongs to a specific protocol or not.

4. Model Evaluation - We used two different criterias for evaluation i.e. accuracy and F1-score, to measure the performance of each model. We then compared the performance of the algorithms and checked the impact of different features on the accuracy of the models.

IV. IMPLEMENTATION

1. Linear Regression - Logistic Regression is a supervised learning approach. It uses a logistic function to calculate the likelihood of a binary (two-class) outcome given one or more predictor variables. Based on a variety of network characteristics, the application name is predicted. It models binary classification to determine whether a given network traffic data belongs to a certain application or not. It also helps to understand the relation between different classes.

2. Random Forest - Random forest is a machine learning algorithm used to predict the application names by combining the outputs of multiple decision trees, wherein each tree is trained on a subset of the data and its features.

3. Accuracy - Accuracy is one of the criteria used to measure the performance of the model. Basically, it shows the correctness of a model. It is defined as the ratio of the number of correctly classified instances to the total number of instances in the dataset.

IV. RESULTS

- After dimensionality reduction, we got 36 columns from 87 columns.
- The accuracy of the Random Forest Classifier is 0.9837 with an F1 Score of 0.982
- The accuracy of Logistic Regression is 0.85 with an F1 score is 0.825

V. CONCLUSION

We tried to predict the application name by using the TCP network dataset using different algorithms. We applied logistic regression and random forest algorithms to build models and evaluated their performance by comparing their accuracy and F1-score. By comparing the accuracy of the two models, we found that random forest shows higher accuracy while logistic regression shows a lower one. Furthermore, the achieved accuracy score for different classes varies. We will try to get an overall balanced score or better performance for all classes in our future work. Also, we will understand the different congestion control protocols and try to use the gained learning from this to implement a predictive model.

VI. REFERENCES

[1] Rojas, J.S. (2018) IP network traffic flows labeled with 75 apps, Kaggle. Available at: https://www.kaggle.com/datasets/jsrojas/ip-network-traffic-flows-labeled-with-87-apps (Accessed: March 11, 2023).

[2] Joshi, M., & Hadi, T. H. (2015). A review of network traffic analysis and prediction techniques. arXiv preprint arXiv:1507.05722. https://arxiv.org/abs/1507.05722