Dissertation Submitted for the partial fulfillment of the B.Sc. as a part of M.Sc. (Integrated) Five Years Program AIML degree to the Department of AIML & Data Science.

IOT Device-Type Identification Using Supervised Machine Learning

Submitted to

-

By

Varsha Mishra
Semester-VI

M.Sc. (Integrated) Five Years Program AI&ML
Department of AIML & Data Science.
School of Emerging Science and Technology
Gujarat University

June, 2022

DECLARATION

This is to certify that the research work reported in this dissertation entitled "**IOT Device-Type Identification Using Supervised Machine Learning"** for the partial fulfillment of B.Sc. as a part of M.Sc. (Integrated) in Artificial Intelligence and Machine Learning degree is the result of investigation done by myself.

Place: Ahmedabad Varsha Mishra

Date:

ACKNOWLEDGEMENT

I would like to thank Dr. Ravi Gor, coordinator of School of Emerging Science and Technology at Gujarat University, for allowing me to work on this project, as well as Machine Learning faculties Mrs.Rashmi Kumari and Mr.Eric Shah, for their consistent encouragement, support and supervision. I'm also appreciative of the cooperation of the department administration.

Last but not least, I would like to express my appreciation towards my parents Indra Mishra and A.K Mishra along with my sisters Diksha Mishra and Esha Mishra and my friends Shruti H. Agarwal, Radhika Sharma and Prableen Sandhu for their help and encouragement throughout this effort.

-Varsha Mishra

Contents

List of Figures and Tables

Chapter 1

Abstract

Abstract

Iot devices are relatively new, therefore their software assurance may not be present, and they may be vulnerable to vulnerabilities. making them typically the weakest link in a corporate network. Cybersecurity teams need additional technology to keep track of all the devices and keep the network safe. A machine learning (ML) approach to IoT security can address some of these challenges. It solves the issue of identifying unknown devices on a network, ensuring they're included in the existing security framework and makes IoT management easier for busy IT teams. The first step is to identify these devices. For this , several pcap files with a lot of packet entries of seven D-Link IoT Devices each .Those packets with TCP protocol were filtered out and a single pcap file for each device was made and it was named with name of the device it belonged to ,then those names were taken as labels and features from all those files were extracted to obtain a single vector space matrix and that was used to train 4 different machine learning model namely logistic Regression, Random Forest, Decision Tree and Naive Bayes out of which Decision Tree predicted the device type with best accuracy of 96%.

Chapter 2

Introduction

2.1 Background

The Internet of Things (IoT) describes the network of physical objects—"things"—that are embedded with sensors, software, and other technologies for the purpose of connecting and exchanging data with other devices and systems over the internet. These devices range from ordinary household objects to sophisticated industrial tools

IoT networks have become an increasingly valuable target of malicious attacks due to the increased amount of valuable data they contain.

2.2 Problem statement

Classify the Iot devices on the network and see which model works the best

2.3 Objective

To Build a model that Classifies the records of packet on network to predict which device sends the packet.

2.4 Motivation and significance

Iot devices are relatively new, therefore their software assurance may not be present, and they may be vulnerable to vulnerabilities. making them typically the weakest link in a corporate network. Cybersecurity

teams need additional technology to keep track of all the devices and keep the network safe.

A machine learning (ML) approach to IoT security can address some of these challenges. It solves the issue of identifying unknown devices on a network, ensuring they're included in the existing security framework and makes IoT management easier for busy IT teamsand the first step for that is to identify these devices.

Chapter 3

Review of Literature

3.1 Online Articles

Read on how machine learning has been used to improve cyber security. Because of ML's ability to deal with massive amounts of data, it is ideal for dealing with cyber security. Unsupervised machine learning models are frequently used to detect anomalies in network traffic and alert cyber security systems. Many businesses have been protected from ransomware attacks. Financial institutions, such as banks, are increasingly relying on machine learning for cyber security.

Many cyber-attacks these days target IoT devices. Cybercriminals use compromised IoT devices to steal information and spy on their victims.

3.2 University Lectures on Youtube

Ricardo Calix lectures on Machine Learning for Cybersecurity and tells how it could be used to solve various problems like fishing,malware detection etc. he in one of his lectures explain how machine learning can be used to detect iot devices on the network .This project follows a somewhat similar approach where we extract data from pcap files representing data packets from seven different D-Link devices.

Chapter 4

Methodology

4.1 Environment

- A machine with Linux
- Python with some libraries such as numpy, pandas and sklearn.
- Wireshark

4.2 Files

Several pcap files of D-link Iot Device which are taken from IMPACT- Iot devices captures.

4.3 Algorithms Used

Supervised Machine learning was used because labels were present in the final dataset.

4.3.1 Logistic Regression

Logistic regression uses sigmoid function to predict the output of a categorical dependent variable. Therefore the outcome must be a categorical or a discrete value.This Algorithm is used here because though by default it classifies data into two categories. It can be changed to predict multiple classes by passing the '*multinomial*' to the `multi_class` parameters.

4.3.2 Decision Tree

It is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.This is used here because Decision trees are very easy as compared to the random forest and are fast and operate easily on large data sets.

4.3.3 Random Forest

It is mainly used for classification problems. It creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.This Algorithm is used here because a random forest combines several decision trees and so though it is a long and slow process and it is more reliable and stable.

4.3.4 Naive Bayes

It is based on Bayes theorem and used for solving classification problems.It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.It is used here because it is one of the simple and most effective Classification algorithms which helps in

building the fast machine learning models that can make quick predictions.

4.4 Workflow

4.4.1 Data Collection

- Pcap Files
    - D-Link IoT device ‘s pcap files were taken from IMPACT-Iot devices captures.Link in the Bibliography.
    - Dataset ID: DS-0941 Name:IoT devices capture

4.4.2. Feature Extraction

pcap file feature extraction

- A python script was written to extract features from pcap files and represent them in the vector space model.
- The labels were assigned based on the type of devices that sent the packets.
- There are 7 iot device folders (D-LinkCamera, D-Link DoorSensor, D-LinkHomeHub, D-LinkSensor, D-LinkSiren, D-LinkSwitch, D-LinkWaterSensor). So, in the end there were 7 labels or classes.
- The device type is judged by the folder the files were present in.

- 18 features were extracted from each of the packages from the 7 devices. The 18 features are:

Feature names	Field name in pcap file
IPLength	ip.len
IPHeaderLength	ip.hdr_len
IPFlags	ip.flags
TTL	ip.ttl
Protocol	ip.proto
IPID	ip.id
Ipchecksum	ip.checksum
SourcePort	ip.srcport
DestPort	ip.destport
SequenceNumber	tcp.seq
AckNumber	tcp.ack
WindowSize	tcp.window_size_value
TCPHeaderLength	tcp.hdr_len
TCPflags	tcp.flags
TCPChecksum	tcp.checksum
TCPStream	tcp.stream
TCPUrgentPointer	tcp.urgent_pointer

Table 1. Features and their name in pcap file

There were multiple pcap files in several folders for a device each which were filtered and merged to form only one pcap file for each device

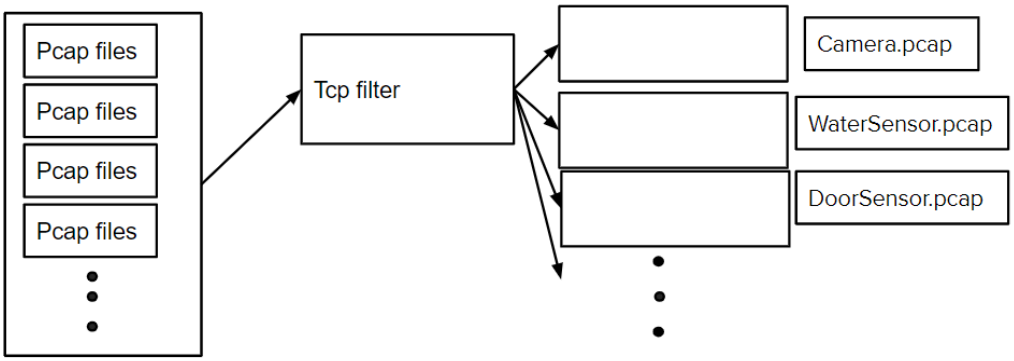

Figure 1. Feature extraction part 1

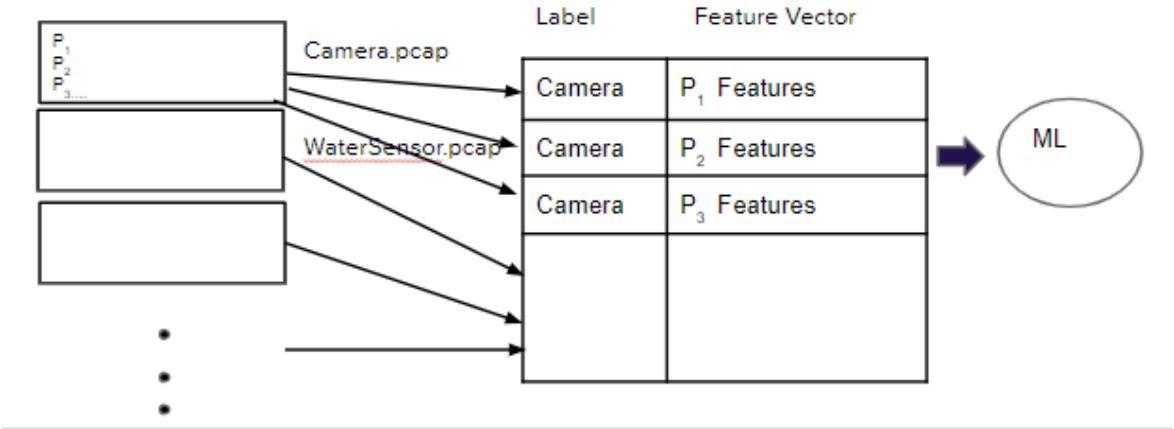

Figure 2. Feature extraction part 2

4.4.2.1. Data Type

The Final dataset contains 57166 records (D-LinkCamera: 2075, D-Link DoorSensor : 3634 , D-LinkHomeHub : 12252, D-LinkSensor : 9988, D-LinkSiren : 9174, D-LinkSwitch : 10783, D-LinkWaterSensor : 9260) with 18 features for each. It is in csv format.

4.4.3 Preprocessing and Model Training

Implemented various Classifiers using Sklearn. Following was done:

A. Imported the required libraries

B. The features and class values were read from dataset

C. Features that were hexadecimal digits were converted to decimal digits.

D. Labels for class values were encoded. Standard scaling of the feature values was done and the dataset was split into training and testing set

E. Specify the classifiers. Four models are trained, and algorithms used are Decision Tree, Random Forest, Logistic Regression and Naive Bayes.

F. Metrics were checked

4.4.4 Model Evaluation

4.4.4.1 Logistic Regression

Accuracy:0.327

Precision: 0.315

Recall: 0.327

F1-measure: 0.309

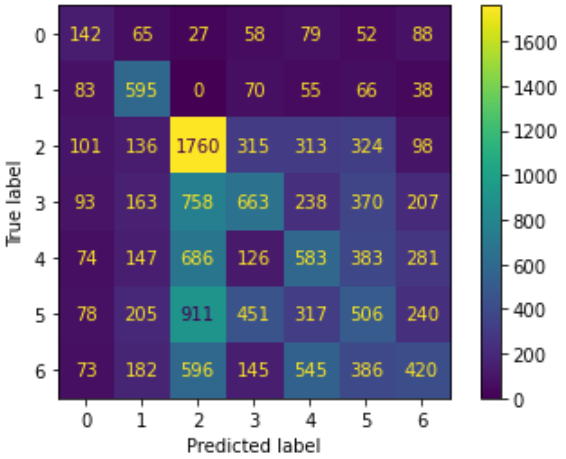

Figure 3: Confusion matrix of Logistic Regression model

4.4.4.2 Random Forest

Accuracy: 0.954

Precision: 0.954

Recall: 0.954

F1-measure: 0.954

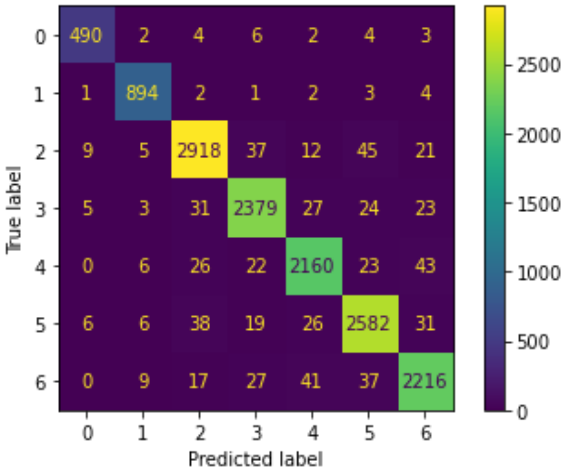

Figure 4: Confusion matrix of Random Forest model

4.4.4.3 Decision Tree

Accuracy: 0.960

Precision: 0.960

Recall: 0.960

F1-measure: 0.960

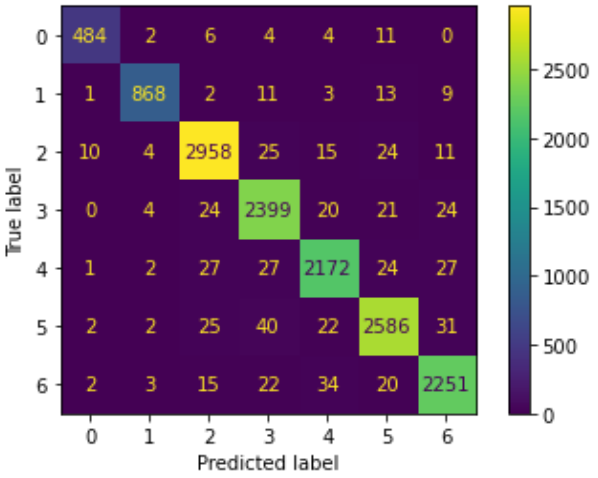

Figure 5: Confusion matrix of Decision Tree model

4.4.4.4 Naive Bayes

Accuracy:0.257

Precision:0.382

Recall: 0.257

F1-measure:0.278

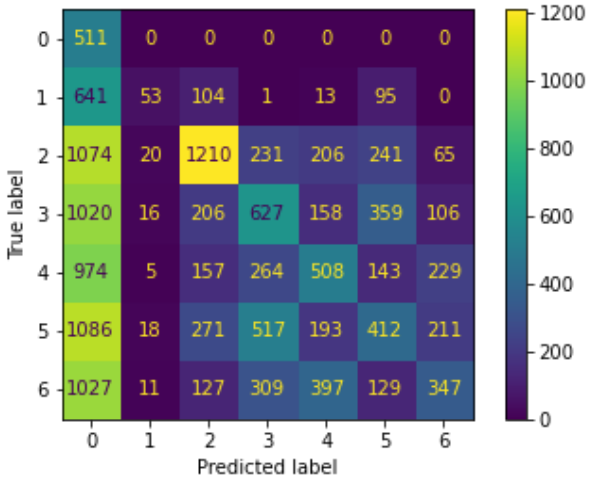

Figure 6: Confusion matrix of Naive Bayes model

Chapter 5

Results

A Decision Tree model and a Random Forest model classify IOT Devices with the accuracy of 96% and 95.4% respectively.

Model	Accuracy
Logistic Regression	32.7%
Random Forest	95.4%
Decision tree	96%
Naive Bayes	56%

Table 2 : Models and their accuracy

Chapter 6

Conclusion

The Decision Tree model had the highest accuracy in classifying IOT devices, but this could be due to overfitting, whereas the Random Forest model has slightly lower accuracy but lowers the risk of overfitting because it uses multiple decision trees. and is more stable and trustworthy. It is resolved by your needs. If you have a large dataset and limited time to work on a model, you will almost certainly choose a decision tree. However, because the dataset for this project is relatively small, we opt for stability and reliability and select the Random Forest model.

6.1 Limitations

This ML model can only classify devices that it has been trained on. So, it can only classify 7 D-link IOT devices .

6.2 Future Enhancements

1. Training model with more no. of IOT devices.
2. Integrating it with a system such that it can work live.
3. Integrating multiple models together to improve accuracy.
4. Build an unsupervised version of this model.

BIBLIOGRAPHY

IoT devices captures — Aalto University's research portal. (April 3 , 2017). Aalto Research.[webpage]. https://research.aalto.fi/en/datasets/iot-devices-captures

Aalto University.(April 3 , 2017).IoT devices captures.[Dataset]. https://research.aalto.fi/files/13004478/captures_IoT_Sentinel.zip

How to use Glob() function to find files recursively in Python? (2020, April 25). GeeksforGeeks. https://www.geeksforgeeks.org/how-to-use-glob-function-to-find-files-recursively-in-python/

IoT devices captures — Aalto University's research portal. (n.d.). Aalto Research. https://research.aalto.fi/en/datasets/iot-devices-captures

Pedamkar, P. (n.d.). *What is TCP Protocol? | How TCP Protocol Works?* eduCBA. https://www.educba.com/what-is-tcp-protocol/

Why is IoT Device Management Important? (n.d.). Cybersecurity Automation. https://www.cybersecurity-automation.com/why-is-iot-device-management-important/

Wireshark Cheat Sheet - Commands, Captures, Filters, Shortcuts & FAQs. (n.d.). Comparitech. https://www.comparitech.com/net-admin/wireshark-cheat-sheet/

Ricardo C. (June 25, 2019), *Machine Learning for Cyber Security: Labs* [Video Playlist], https://www.youtube.com/watch?v=lTge-G02Cis