Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

AN EFFICIENT SPAM DETECTION ON IOT DEVICES USING MACHINE LEARNING

Ch.Drakshayani, K.Tejaswi, L.Hima Gayathri, L.Tejaswini, Btech Student of IT Dept, Vijaya College, Vijayawada, Andharapradesh, Email:- drakshareddy06@gmail.com
CH.V.RAO,,Assistant Professor **,** IT DEPT, Vijaya College**,** Vijayawada, Andharapradesh
Email:- chvrao89@gmail.com

ABSTRACT
The Internet of Things (IoT) is a group of millions of devices having sensors and actuators linked over wired or wireless channel for data transmission. IoT has grown rapidly over the past decade with more than 25 billion devices are expected to be connected by 2020. The volume of data released from these devices will increase many-fold in the years to come. In addition to an increased volume, the IoT devices produces a large amount of data with a number of different modalities having varying data quality defined by its speed in terms of time and position dependency. In such an environment, machine learning algorithms can play an important role in ensuring security and authorization based on biotechnology, anomalous detection to improve the usability and security of IoT systems. On the other hand, attackers often view learning algorithms to exploit the vulnerabilities in smart IoT-based systems. Motivated from these, in this paper, we propose the security of the IoT devices by detecting spam using machine learning. To achieve this objective, Spam Detection in IoT using Machine Learning framework is proposed. In this framework, five machine learning models are evaluated using various metrics with a large collection of inputs features sets. Each model computes a spam score by considering the refined input features. This score depicts the trustworthiness of IoT device under various parameters. REFIT Smart Home dataset is used for the validation of proposed technique. The results obtained proves the effectiveness of the proposed scheme in comparison to the other existing schemes.

K**eywords:** —spam detection,IOT,websites, features, RandomForest, REFIT Smart Home dataset

1 INTRODUCTION

Machine learning and data mining often employ the same methods and overlap significantly, but while machine learning focuses on prediction, based on known properties learned from the training data, data mining focuses on the discovery of (previously) unknown properties in the data (this is the analysis step of knowledge discovery in databases). Data mining uses many machine learning methods, but with different goals; on the other hand, machine learning also employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy. Much of the confusion between these two research communities (which do often have separate conferences and separate journals, ECML PKDD being a major exception) comes from the basic assumptions they work with: in machine learning, performance is usually evaluated with respect to the ability to reproduce known knowledge, while in knowledge discovery and data mining (KDD) the key task is the discovery of previously unknown knowledge. Evaluated with respect to known knowledge, an uninformed (unsupervised) method will easily be outperformed by other supervised methods, while in a typical KDD task, supervised methods cannot be used due to the unavailability of training data. The difference between optimization and machine learning arises from the goal of generalization: while optimization algorithms can minimize the loss on a training set, machine learning is concerned with minimizing the loss on unseen samples. Characterizing the generalization of various learning algorithms is an active topic of current research, especially for deep learning algorithms

2 RELEATED WORK

1) Iot security: ongoing challenges and research opportunities

AUTHORS: Z.-K. Zhang, M. C. Y. Cho, C.-W. Wang, C.-W. Hsu, C.-K. Chen, and S. Shieh

The Internet of Things (IoT) opens opportunities for wearable devices, home appliances, and software to share and communicate information on the Internet. Given that the shared data contains a large amount of private information, preserving information security on the shared data is an important issue that cannot be neglected. In this paper, we begin with general information security background of IoT and continue on with information security related challenges that IoT will encountered. Finally, we will also point out research directions that could be the future work for the solutions to the security challenges that IoT encounters.

2) Communication security in internet of thing: preventive measure and avoid ddos attack over iot network

AUTHORS: C. Zhang and R. Green

The idea of Internet of Things (IoT) is implanting networked heterogeneous detectors into our daily life. It opens extra channels for information submission and remote control to our physical world. A significant feature of an IoT network is that it collects data from network edges. Moreover, human involvement for network and devices maintenance is greatly reduced, which suggests an IoT network need to be highly self-managed and self-secured. For the reason that the use of IoT is growing in many important fields, the security issues of IoT need to be properly addressed. Among all, Distributed Denial of Service (DDoS) is one of the most notorious attacking behaviors over network which interrupt and block genuine user requests by flooding the host server with huge number of requests using a group of zombie computers via geographically distributed internet connections. DDoS disrupts service by creating network congestion and disabling normal functions of network components, which is even more disruptive for IoT. In this paper, a lightweight defensive algorithm for DDoS attack over IoT network environment is proposed and tested against several scenarios to dissect the interactive communication among different types of network nodes.

3) The dark side of the internet: Attacks, costs and responses

AUTHORS: W. Kim, O.-R. Jeong, C. Kim, and J. So

The Internet and Web technologies have originally been developed assuming an ideal world where all users are honorable. However, the dark side has emerged and bedeviled the world. This includes spam, malware, hacking, phishing, denial of service attacks, click fraud, invasion of privacy, defamation, frauds, violation of digital property rights, etc. The responses to the dark side of the Internet have included technologies, legislation, law enforcement, litigation, public awareness efforts, etc. In this paper, we explore and provide taxonomies of the causes and costs of the attacks, and types of responses to the attacks.

4) Conditional privacy preserving security protocol for nfc applications

AUTHORS: H. Eun, H. Lee, and H. Oh

In recent years, various mobile terminals equipped with NFC (Near Field Communication) have been released. The combination of NFC with smart devices has led to widening the utilization range of NFC. It is expected to replace credit cards in electronic payment, especially. In this regard, security issues need to be addressed to vitalize NFC electronic payment. The NFC security standards currently being applied require the use of user's public key at a fixed value in the process of key agreement. The relevance of the message occurs in the fixed elements such as the public key of NFC. An attacker can create a profile based on user's public key by collecting the associated messages. Through the created profile, users can be exposed and their privacy can be compromised. In this paper, we propose conditional privacy protection methods based on pseudonyms to solve these problems. In addition, PDU (Protocol Data Unit) for conditional privacy is defined. Users can inform the other party that they will communicate according to the protocol proposed in this paper by sending the conditional privacy preserved PDU through NFC terminals. The proposed method succeeds in minimizing the

update cost and computation overhead by taking advantage of the physical characteristics of NFC 1 .

5) Neural network based secure media access control protocol for wireless sensor networks

AUTHORS: R. V. Kulkarni and G. K. Venayagamoorthy

This paper discusses an application of a neural network in wireless sensor network security. It presents a multilayer perceptron (MLP) based media access control protocol (MAC) to secure a CSMA-based wireless sensor network against the denial-of-service attacks launched by adversaries. The MLP enhances the security of a WSN by constantly monitoring the parameters that exhibit unusual variations in case of an attack. The MLP shuts down the MAC layer and the physical layer of the sensor node when the suspicion factor, the output of the MLP, exceeds a preset threshold level. Backpropagation and particle swarm optimization algorithms are used for training the MLP. The MLP-guarded secure WSN is implemented using the Vanderbilt Prowler simulator. Simulation results show that the MLP helps in extending the lifetime of the WSN.

2 PROPOSED WORK AND ALOGRITHAM

- The digital world is completely dependent upon the smart devices. The information retrieved from these devices should be spam free. The information retrieval from various IoT devices is a big challenge because it is collected from various domains. As there are multiple devices involved in IoT, so a large volume of data is generated having heterogeneity and variety. We can call this data as IoT data. IoT data has various features such as real-time, multi-source, rich and sparse.
- The proposed scheme of spam detection in IOT is validated using machine learning model. An algorithm is proposed to compute the spamicity score of the model which is then used for detection and intelligent decision making. Based upon the spamicity score computed in previous step, the reliability of IoT devices is analyzed using different evaluation metrics.
- To protect the IoT devices from producing the malicious information, the web spam detection is targeted in this proposal. We have considered the machine learning algorithm for the detection of spam from the IoT devices.
- The dataset used in the experiments, contains the data recorded for the span of eighteen months. For better results and accuracy, we have considered the data of one month. Considering the fact, the climate is the important parameter for the working of IoT device, the month with maximum variations has been taken into the consideration.

3.1 ADVANTAGES OF PROPOSED SYSTEM:

- Machine learning techniques help to build protocols for lightweight access control to save energy and extend the IoT systems lifetime.
- The efficiency IoT data increases, if stored, processed and retrieved in an efficient manner. This proposal aims to reduce the occurrence of spam from these devices.

3.2 **Random Forest Algorithm**

Random forest is a popular Machine Learning algorithm that is used to solve a variety of problems. It's a supervised algorithm that can solve classification and regression problems. It is made up of a parallel decision tree that takes in data and generates a particular class. As a result, a large number of trees generate various groups. Finally, as the final output class, the sum of all classes is used.

WHAT IS THE RANDOM FOREST ALGORITHM AND HOW DOES IT WORK?

The random forest algorithm is carried out in the following general steps.

1. Select N columns at random from the dataset.
2. Build a decision tree using these N records.
3. Repeat steps 1 and 2 with the number of trees you want in your algorithm.
4. Finally, the sum of all the available classes is selected as the final class.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

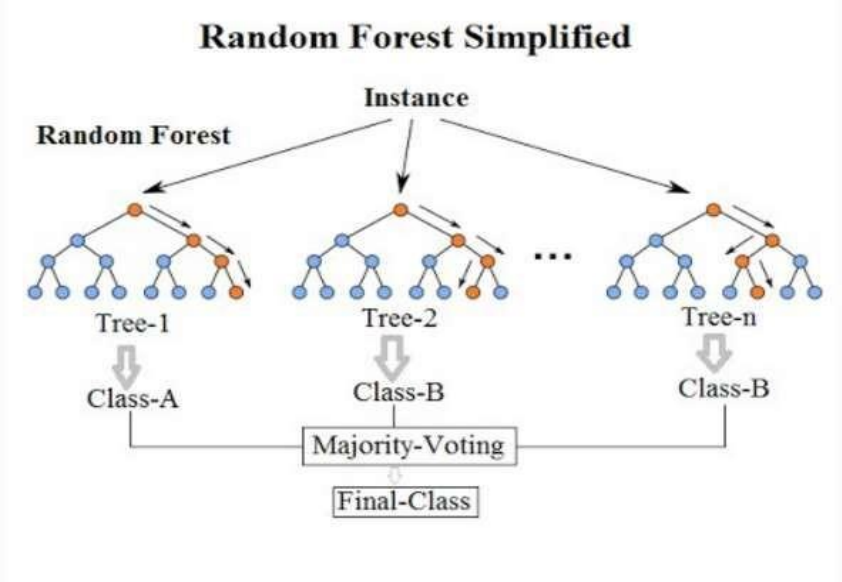

Fig1:- Random Forest Architure

3.3 Support Vector Machine:

The Support Vector Machine is one of the most widely used Machine Learning algorithms. The main goal of this algorithm is to find the best data split possible. It is used to solve problems involving classification and regression. It can solve both linear and nonlinear separable data, which is one of its main advantages. The separation line is known as the Hyper plane. Support vectors are the points on which the margins are built. The svm algorithm is depicted in the diagram below.

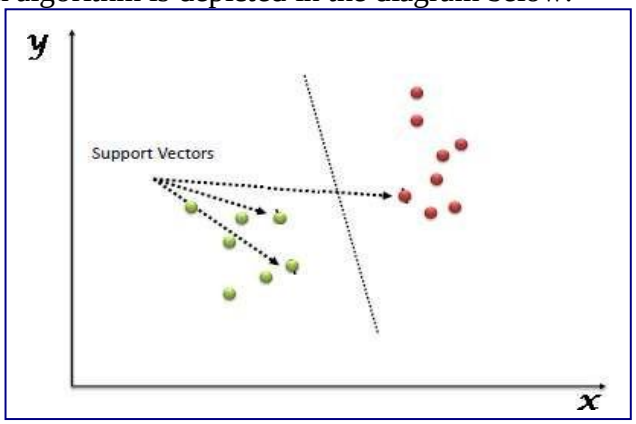

Fig2:- SVM Architure

4 METHODOLOGY

4.1 Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc. REFIT Smart Home dataset Link: https://www.refitsmarthomes.org/datasets/

4.1.1 Dataset:

In this data set we are taken 32 columns and 503910 rows in the dataset, which are described below.

- **gen [kW]**
  - Total energy generated by means of solar or other power generation resources
- **House overall [kW]**

    - overall house energy consumption
    - **Dishwasher [kW]**
    - energy consumed by specific appliance
- **Furnace 1 [kW]**
    - energy consumed by specific appliance
- **Furnace 2 [kW]**
    - energy consumed by specific appliance
- **Home office [kW]**
    - energy consumed by specific appliance
- **Fridge [kW]**
    - energy consumed by specific appliance
- **Wine cellar [kW]**
    - energy consumed by specific appliance
- **Garage door [kW]**
    - energy consumed by specific appliance
- **Kitchen 12 [kW]**
    - energy consumption in kitchen 1
- **Kitchen 14 [kW]**
    - energy consumption in kitchen 2
- **Kitchen 38 [kW]**
    - energy consumption in kitchen 3
- **Barn [kW]**
    - energy consumed by specific appliance
- **Well [kW]**
    - energy consumed by specific appliance
- **Microwave [kW]**
    - energy consumed by specific appliance
- **Living room [kW]**
    - energy consumption in Living room
- **Solar [kW]**
    - Solar power generation

Weather

- **temperature**:
    - emperature is a physical quantity expressing hot and cold.
- **humidity**:
    - Humidity is the concentration of water vapour present in air.
- **visibility**:
    - Visibility sensors measure the meteorological optical range which is defined as the length of atmosphere over which a beam of light travels before its luminous flux is reduced to 5% of its original value.
- **apparentTemperature**:
    - Apparent temperature is the temperature equivalent perceived by humans, caused by the combined effects of air temperature, relative humidity and wind speed. The measure is most commonly applied to the perceived outdoor temperature.
- **pressure**:
    - Falling air pressure indicates that bad weather is coming, while rising air pressure indicates good weather
- **windSpeed**:

- Wind speed, or wind flow speed, is a fundamental atmospheric quantity caused by air moving from high to low pressure, usually due to changes in temperature.

- **cloudCover**:
  - Cloud cover (also known as cloudiness, cloudage, or cloud amount) refers to the fraction of the sky obscured by clouds when observed from a particular location. Okta is the usual unit of measurement of the cloud cover.
- **windBearing**:
  - In meteorology, an azimuth of 000° is used only when no wind is blowing, while 360° means the wind is from the North. True Wind Direction True North is represented on a globe as the North Pole. All directions relative to True North may be called "true bearings."
- **dewPoint**:
  - the atmospheric temperature (varying according to pressure and humidity) below which water droplets begin to condense and dew can form.
- **precipProbability**:
  - A probability of precipitation (POP), also referred to as chance of precipitation or chance of rain, is a measure of the probability that at least some minimum quantity of precipitation will occur within a specified forecast period and location.
- **precipIntensity**:
  - The intensity of rainfall is a measure of the amount of rain that falls over time. The intensity of rain is measured in the height of the water layer covering the ground in a period of time. It means that if the rain stays where it falls, it would form a layer of a certain height.

Others

- **summary**:
  - Report generated by the by the data collection systm (apparently!).
  - Including:
  - Clear, Mostly Cloudy, Overcast, Partly Cloudy, Drizzle,
  - Light Rain, Rain, Light Snow, Flurries, Breezy, Snow,
  - Rain and Breezy, Foggy, Breezy and Mostly Cloudy,
  - Breezy and Partly Cloudy, Flurries and Breezy, Dry, Heavy, Snow.
- **icon**:
  - The icon that is used by the data collection systm (apparently!).
  - Including:
    cloudy, clear-night, partly-cloudy-night, clear-day, partly-cloudy-day, rain, snow, wind, fog.

Class = Spam 1 or no spam 0

4.2 Data Preparation:

we will transform the data. By getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain.

Next we drop or remove all columns except for the columns that we want to retain.

Finally we drop or remove the rows that have missing values from the data set.

4.3 Model Selection:

While creating a machine learning model, we need two dataset, one for training and other for testing. But now we have only one. So lets split this in two with a ratio of 80:20. We will also divide the dataframe into feature column and label column. Here we imported train_test_split function of sklearn. Then use it to split the dataset. Also, *test_size = 0.2*, it makes the split with 80% as train

dataset and 20% as test dataset. The *random_state* parameter seeds random number generator that helps to split the dataset. The function returns four datasets. Labelled them as *train_x, train_y, test_x, test_y*. If we see shape of this datasets we can see the split of dataset. We will use Random Forest Classifier, which fits multiple decision tree to the data. Finally I train the model by passing *train_x, train_y* to the *fit* method.

Once the model is trained, we need to Test the model. For that we will pass *test_x* to the predict method.

Random Forest is one of the most powerful methods that is used in machine learning for classification problems. The random forest comes in the category of the supervised classification algorithm. This algorithm is carried out in two different stages the first one deals with the creation of the forest of the given dataset, and the other one deals with the prediction from the classification.

4.3 Analyze and Prediction:

In the actual dataset, we chose only 22 features :

- **gen [kW]**
  - Total energy generated by means of solar or other power generation resources
- **House overall [kW]**
  - overall house energy consumption
- **Dishwasher [kW]**
  - energy consumed by specific appliance
- **Furnace 1 [kW]**
  - energy consumed by specific appliance
- **Home office [kW]**
  - energy consumed by specific appliance
- **Fridge [kW]**
  - energy consumed by specific appliance
- **Wine cellar [kW]**
  - energy consumed by specific appliance
- **Garage door [kW]**
  - energy consumed by specific appliance
- **Kitchen 12 [kW]**
  - energy consumption in kitchen 1
- **Barn [kW]**
  - energy consumed by specific appliance
- **Well [kW]**
  - energy consumed by specific appliance
- **Microwave [kW]**
  - energy consumed by specific appliance
- **Living room [kW]**
  - energy consumption in Living room
- **Solar [kW]**
  - Solar power generation

Weather

- **temperature**:
  - emperature is a physical quantity expressing hot and cold.
- **humidity**:
  - Humidity is the concentration of water vapour present in air.
- **visibility**:

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

  - Visibility sensors measure the meteorological optical range which is defined as the length of atmosphere over which a beam of light travels before its luminous flux is reduced to 5% of its original value.
- **apparentTemperature**:
  - Apparent temperature is the temperature equivalent perceived by humans, caused by the combined effects of air temperature, relative humidity and wind speed. The measure is most commonly applied to the perceived outdoor temperature.
- **pressure**:
  - Falling air pressure indicates that bad weather is coming, while rising air pressure indicates good weather
- **windSpeed**:
  - Wind speed, or wind flow speed, is a fundamental atmospheric quantity caused by air moving from high to low pressure, usually due to changes in temperature.
- **cloudCover**:
  - Cloud cover (also known as cloudiness, cloudage, or cloud amount) refers to the fraction of the sky obscured by clouds when observed from a particular location. Okta is the usual unit of measurement of the cloud cover.
- **windBearing**:
  - In meteorology, an azimuth of 000° is used only when no wind is blowing, while 360° means the wind is from the North. True Wind Direction True North is represented on a globe as the North Pole. All directions relative to True North may be called "true bearings.

Class = Spam 1 or no spam 0

4.3 Accuracy on test set:

We got a accuracy of 99.1% on test set.

4.4 Saving the Trained Model:

Once you're confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or . pkl file using a library like pickle .Make sure you have pickle installed in your environment. Next, let's import the module and dump the model into . pkl file

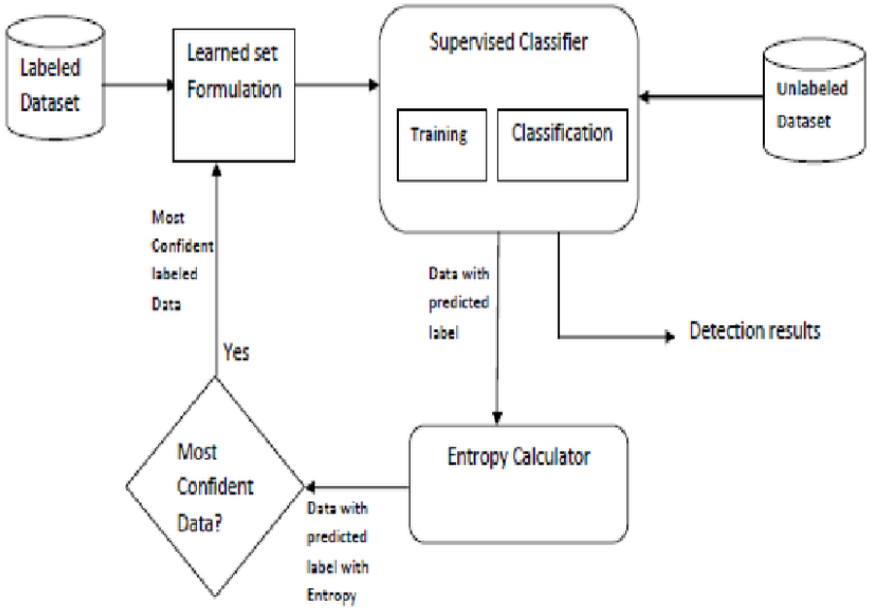

Fig. 1: proposed System Flow Diagram

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

5 RESULTS AND DISCUSSION

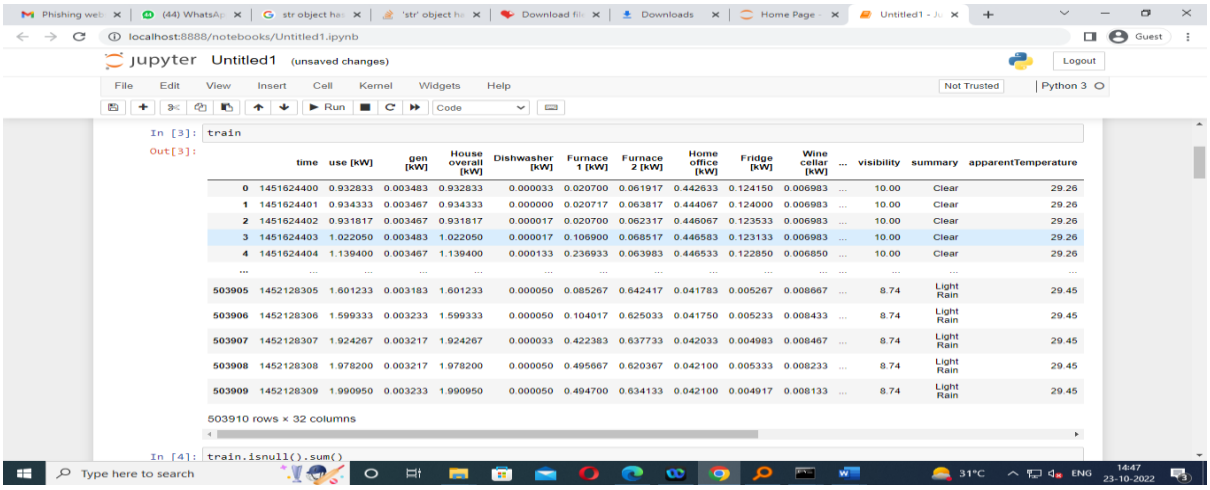

Fig2:-spam Iot dataset

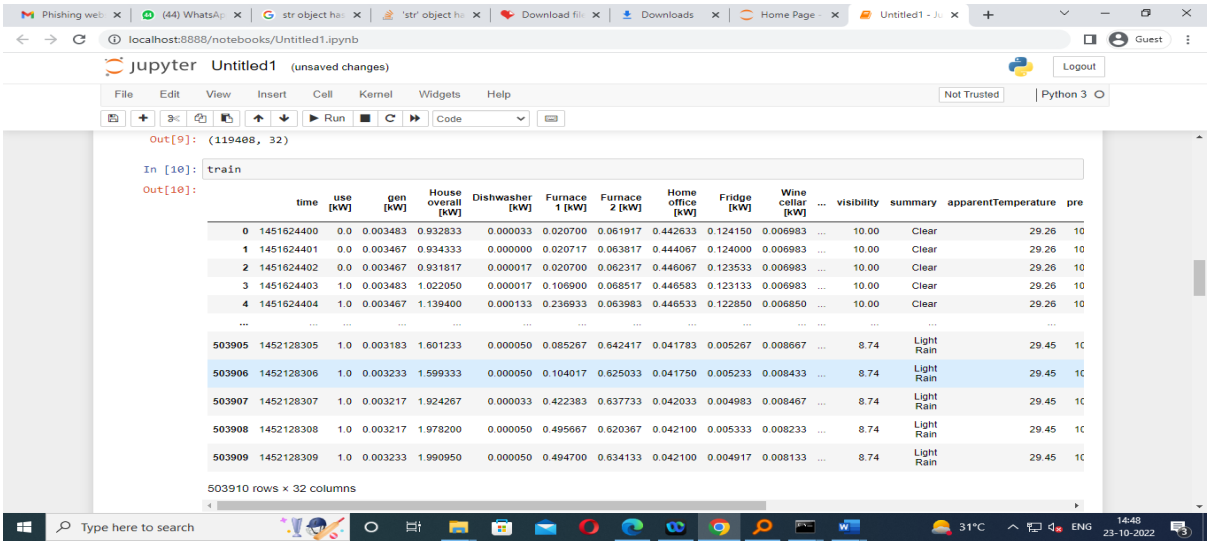

Fig3: IOT Training Dataset

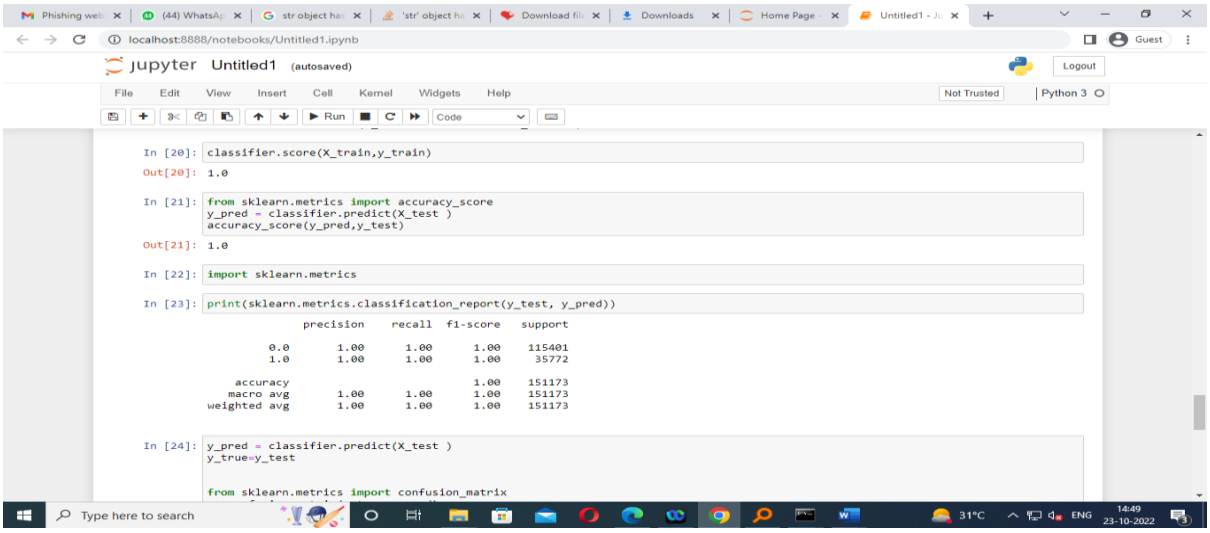

Fig4: Evolution Performance Result

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

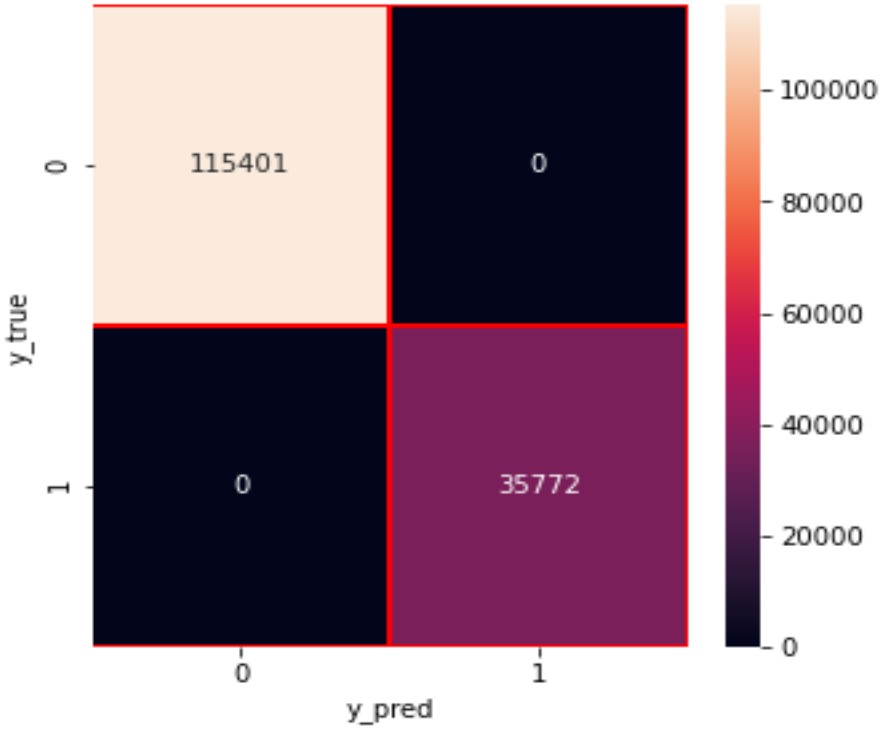

Fig5:-Confusion Matrix

6. CONCLUSION AND FUTURE WORK

The proposed framework, detects the spam parameters of IoT devices using machine learning models. The IoT dataset used for experiments, is pre-processed by using feature engineering procedure. By experimenting the framework with machine learning models, each IoT appliance is awarded with a spam score. This refines the conditions to be taken for successful working of IoT devices in a smart home. In future, we are planning to consider the climatic and surrounding features of IoT device to make them more secure and trustworthy.

7. REFRENCES

[1] Z.-K. Zhang, M. C. Y. Cho, C.-W. Wang, C.-W. Hsu, C.-K. Chen, and S. Shieh, "Iot security: ongoing challenges and research opportunities," in 2014 IEEE 7th international conference on service-oriented computing and applications. IEEE, 2014, pp. 230–234.

[2] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, "Blockchain for iot security and privacy: The case study of a smart home," in 2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops). IEEE, 2017, pp. 618–623.

[3] E. Bertino and N. Islam, "Botnets and internet of things security," Computer, no. 2, pp. 76–79, 2017.

[4] C. Zhang and R. Green, "Communication security in internet of thing: preventive measure and avoid ddos attack over iot network," in Proceedings of the 18th Symposium on Communications & Networking. Society for Computer Simulation International, 2015, pp. 8–15.

[5] W. Kim, O.-R. Jeong, C. Kim, and J. So, "The dark side of the internet: Attacks, costs and responses," Information systems, vol. 36, no. 3, pp. 675–705, 2011.

[6] H. Eun, H. Lee, and H. Oh, "Conditional privacy preserving security protocol for nfc applications," IEEE Transactions on Consumer Electronics, vol. 59, no. 1, pp. 153–160, 2013.

[7] R. V. Kulkarni and G. K. Venayagamoorthy, "Neural network based secure media access control protocol for wireless sensor networks," in 2009 International Joint Conference on Neural

Networks. IEEE, 2009, pp. 1680–1687.
[8] M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 1996– 2018, 2014.
[9] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2015.
[10] F. A. Narudin, A. Feizollah, N. B. Anuar, and A. Gani, "Evaluation of machine learning classifiers for mobile malware detection," Soft Computing, vol. 20, no. 1, pp. 343–357, 2016.
[11] Z. Tan, A. Jamdagni, X. He, P. Nanda, and R. P. Liu, "A system for denial-of-service attack detection based on multivariate correlation analysis," IEEE transactions on parallel and distributed systems, vol. 25, no. 2, pp. 447–456, 2013.
[12] Y. Li, D. E. Quevedo, S. Dey, and L. Shi, "Sinr-based dos attack on remote state estimation: A gametheoretic approach," IEEE Transactions on Control of Network Systems, vol. 4, no. 3, pp. 632–642, 2016.
[13] L. Xiao, Y. Li, X. Huang, and X. Du, "Cloud-based malware detection game for mobile devices with offloading," IEEE Transactions on Mobile Computing, vol. 16, no. 10, pp. 2742–2750, 2017