Global Journal of Computer Science and Technology: D
Neural & Artificial Intelligence
Volume 22 Issue 2 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Machine Learning-based Model for Implementing the Cybersecurity for Organizations 'Assets

By Dr. Mohammad Mahmoud Abu Omar, Mohammad Awawdeh & Ramze Sanea

An-Najah National University

Abstract- In the current era, cybersecurity problems are one of the critical problems that threaten the organizations' assets, since they may cause a big financial and moral loss. and in the parallel the with the advent of the Machine Learning and Artificial Intelligence techniques, It is important and wise to use these technologies to help in achieving cybersecurity for organizations' assets, due to accurate work of these systems and saving time, effort, and cost. So, this research develops a model that uses machine learning technology to detect the vulnerability in the information security of the organizations' assets to avoid as possible the lack of the information security in organizations' assets and thus avoid the financial and moral loss that such organizations may face.

Index Terms: information security policy, ISP, machine learning, dataset, vulnerable, CART.

GJCST-D Classification: DDC Code: 004.62 LCC Code: T58.5

Strictly as per the compliance and regulations of:

A Machine Learning-based Model for Implementing the Cybersecurity for Organizations 'Assets

Dr. Mohammad Mahmoud Abu Omar [α], Mohammad Awawdeh [σ] & Ramze Sanea [ρ]

Abstract- In the current era, cybersecurity problems are one of the critical problems that threaten the organizations' assets, since they may cause a big financial and moral loss. and in the parallel the with the advent of the Machine Learning and Artificial Intelligence techniques, It is important and wise to use these technologies to help in achieving cybersecurity for organizations' assets, due to accurate work of these systems and saving time, effort, and cost. So, this research develops a model that uses machine learning technology to detect the vulnerability in the information security of the organizations' assets to avoid as possible the lack of the information security in organizations' assets and thus avoid the financial and moral loss that such organizations may face.

Index Terms: information security policy, ISP, machine learning, dataset, vulnerable, CART.

I. Introduction

Machine learning aims to answer the question of how to create machines that make a decision on their own depending on the learned dataset. It is a mix between computer science and statistics, as well as it is at the heart of artificial intelligence and data science, and is one of today's fastest expanding technological topics.

The information security policy is one of the most critical information security measures. This important directiongiving document, however, is not always easy to create, and its writers struggle with issues such as what defines a policy. As a result, policymakers are forced to rely on existing sources for direction. The many worldwide information security standards are one of these sources.

A dataset is a set of data, usually represented in the form of a table. Each column in the table represents a specific feature, and each row returns to an element of the dataset that is used in the learning phase when building a model.

CART stands for Classification and Regression Tree, can easily handle both numerical and categorical

Author α: Assistant Professor, Faculty of Engineering and Information Technology, An-Najah National University- Nablus, Palestine.
e-mail: mmdabuomar@yahoo.com
Author σ ρ: Student and Cyber Security and Networking expert Faculty of Engineering and Information Technology An- Najah National University- Nablus, Palestine.
e-mails: mohamad.awawdeh2000@gmail.com, ramzesane854@gmail.com

CART stands for Classification and Regression Tree, can easily handle both numerical and categorical variables, but there is tow common disadvantages the first one is CART split one variable at a time, the other one is CART may create unstable decision tree.

II. Research Objectives

- The main goals of the research are to suggest best practices to cover the vulnerabilities that belong to user assets.
- Assisting the user in identifying unknown vulnerabilities that could be destructive and recommending appropriate solutions while keeping the budget in mind.

III. Research Methodology

A suitable dataset is required for any machine learning model, so before implementing any model, it is necessary to provide a suitable dataset that is used to train the machine, and the next step is to split the dataset into two subsets using an appropriate splitting algorithm, the first one being used to train the machine. After that, the appropriate ML algorithm must be chosen and used for dataset exploration and pattern recognition with minimum human intervention. Finally, after the algorithm chooses the second subset, it is used to test and evaluate the model.

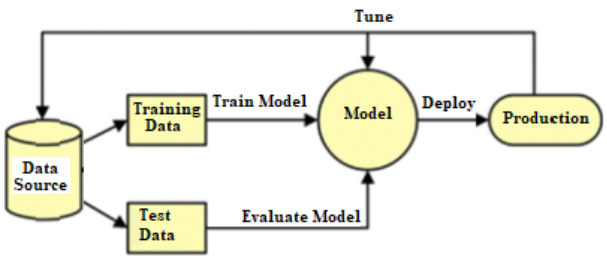

Fig. 1: Supervised learning Workflow

IV. How Does the Model Work

a) Input

Before input, the dataset should be converted to a numerical value so that the algorithm can deal with it and add a new column that contains the unique numerical key points to the solution. Then the numerical dataset will be entered as an input to train and test the

model, in addition to the user asset and the budget that he has.

b) Processing

After the dataset enters the model, it should be split into two subdatasets. The first one is to train and the other is to test, and to do so, the machine uses a k-fold algorithm to divide the origenal dataset into test dataset with 0.30 of the dataset at all and the remain data should be the train dataset. The random forest algorithm uses the training dataset by building a number of decision trees. Then combine them to one tree.

c) Output

The final step is to predict the solution key for the user asset, but before that, the machine uses the testing subdataset to evaluate the usage algorithm and the model. So the input data from the user enters into the model and takes a suitable path on the tree that has already been built in the processing phase, and when the machine reaches the leaf that is the prediction that the user needs, the result is converted to the related solution that maps to the prediction key value.

Fig. 2: Work summary

V. Literature Review

ISP is one of the essential documentation in any organization because it o define the rights and responsibilities of information resource users, so to make ISP effective is should be contain the those activity as shown in the figure.

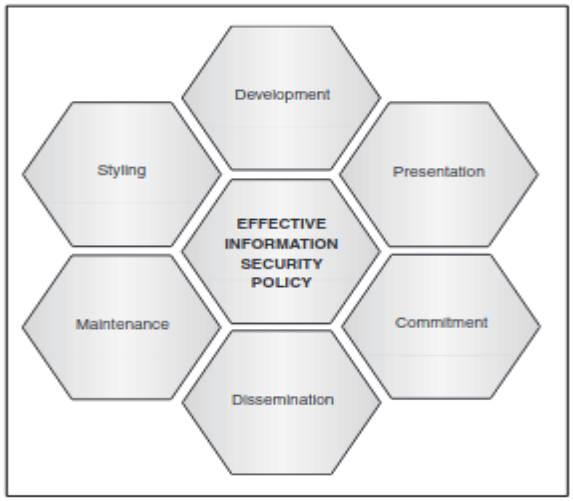

Fig. 3: Supporting activities for an effective information security policy

Styling one of the important factor to develop an efficient ISP so it should be write in clear and consist manner in addition to fit with the organizational culture, the second factor is Development so it should be updated from time to time to be suitable with the organization requirement. the other factor for develop in effective ISP is commitment because when the employee in the organization see the header and all the manger commit to the ISP this means all the employee should Committed to it.

VI. Application Example

After splitting phase the machine calculate information gain for each feature in the splitting datasets to indicate the impurity for each one of them using the following equation.

$$Gain(T,X) = Entropy(T) — Entropy(T,X)$$

- T = target variable
- X = Feature to be split on
- Entropy(T,X) = The entropy calculated after the data is split on feature X

Fig. 4: Information gain

When the machine using random force algorithm it must calculate nodes importance for each created decision tree using the following equation.

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

- ni sub(j)= the importance of node j
- w sub(j) = weighted number of samples reaching node j
- C sub(j)= the impurity value of node j
- left(j) = child node from left split on node j
- right(j) = child node from right split on node j

Fig. 5: Nodes importance

And this lead to calculate the feature importance on a decision tree using the following equation.

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k}$$

- fi sub(i)= the importance of feature i
- ni sub(j)= the importance of node j

Fig. 6: Feature importance

Then the machine normalized to a value between 0 and 1 by dividing by the sum of all feature importance values as the equation bellow.

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

Fig. 7: Normalized feature importance

Because the random force algorithm was used, the machine sum all feature's importance value for all decision trees and divided by the total number of trees.

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T}$$

Fig. 8: Actual feature importance

After the machine learn the dataset that entered it well be ready to suggest the key of the solution for the most important vulnerability that the machine see depending on the taken path that the machine take so when the key predicted the check if this key is exist in the interred dataset so if it exist the machine return the solution that related to this key, on other hand if the predicted key does not exist in the dataset the machine try to approximate this key to the nearest value that exist in the dataset and return the solution that related with it.

VII. Conclusion

Machine learning is a wondrous method by which to solve a critical problem. One of them is organization security, so it can help to build an information security policy in a short time with high accuracy.

Assets are the most important thing that the organization has so ot will be secure as possible we can and we can let it more secure by cover the vulnerability for all one of this assets by choose the best solution for all one of it.

In this research the machine learning used to suggest a solution for one vulnerability that belong to the user asset, so the machine depending on both mathematical equation and suitable path in the tree that built in learning phase to choose what one of the vulnerability should choose to suggest the solution for protect against it.

References Références Referencias

1. Mathias Riechert, "Research Information Standardization as a Wicked Problem: Possible Consequences for the Standardization", In Proceedings of The International Conference of Current Research Information Systems, At Rome, 2014, Volume 12.
2. Ram L Kumar, Antonis C Stylianou, " A process model for analyzing and managing flexibility in information systems," European Journal of Information Systems, 2013.
3. Valdis Vizulis, Edgars Diebelis, "Self-Testing Approach and Testing Tools", Journal of University of Latvia, 2012, Volume 787.
4. Lionel Briand, Yvan Labiche, "A UML-Based Approach to System Testing ", Journal of Carleton University, 2002, Version 4.
5. Lejk, M. and Deeks. An Introduction to Systems Analysis Techniques, 2nd Ed., Perason Education Limited, 2002.
6. Satzinger, J., Systems Analysis and Design, 2nd ed., Thomson Learning, 2002
7. Dennis, A. et al, Systems Analysis and Design: An-Object-Oriented Approach with UML. John Wiley Sons Inc., 2002
8. Maciaszek, A., Requirments Analysis and System Design, Developing Information Systems with UML, Addison-Wesely, 2001.