##### **A User-Centric Machine Learning Framework for Cyber Security Operations Center**

##### **CREATIVE AND INNOVATIVE PROJECT REPORT**

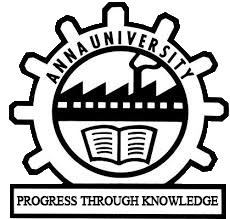
###### **Submitted by**

**G. Santhosh(2016103057)**

**S. Ben Stewart(2016103513)**

**P. Udaykumar(2016103622)**

**College of Engineering, Guindy**



**ANNA UNIVERSITY: CHENNAI 600 025**

OCTOBER 2019

# 

**ANNA UNIVERSITY: CHENNAI 600 025**

**BONAFIDE CERTIFICATE**

# Certified that this project report “**A User-Centric Machine Learning Framework for Cyber Security Operations Center**” is the bonafide work of “G.Santhosh (2016103057), S.Ben Stewart (2016103513) and P.Udaykumar (2016103622)” who carried out the project work under my supervision.

**Dr. AR. Arunarani**

**SUPERVISOR**

Teaching Fellow

Department of

Computer Science and Engineering, College of Engineering, Guindy Anna University.

**Dr. S. VALLI**

**HOD**

Professor and Head

Department of Computer Science and Engineering

College of Engineering, Guindy

Anna University

# **ACKNOWLEDGEMENT**

This project would not have been possible if not for the tireless efforts by the members of this team in all the divisions in which they were asked to work on. Right from the project ideation till the implementation and the documentation and presentation part. We would also thank our mentors and friends who helped us when the project hit a few roadblocks.

# **ABSTRACT**

To assure cyber security of an enterprise, typically SIEM (Security Information and Event Management) system is in place to normalize security events from different preventive technologies and flag alerts. Analysts in the security operations center (SOC) investigate the alerts to decide if it is truly malicious or not. However, generally the number of alerts is overwhelming capacity to handle all alerts. Because of this, potential malicious attacks and compromised hosts may be missed. Machine learning is a viable approach to reduce the false positive rate and improve the productivity of SOC analysts. In this paper, we develop a user- centric machine learning framework for the cyber security operation center in real enterprise environment. We discuss the typical data sources in SOC, their workflow, and how to leverage and process these data sets to build an effective machine learning system. We use the system using the key repository of information regarding the vulnerabilities that allow intruders to breach computer networks is the National Vulnerability Database (NVD). NVD is a product of the U.S. National Institute of Standards and Technology's (NIST) Computer Security Division and is also sponsored by the U.S. Department of Homeland Security’s Computer Emergency Readiness Team (US-CERT). We are implementing the below steps from data massaging, label creation, feature engineering, machine learning algorithm selection, model performance evaluations, to risk score generation.

The above implementation would help other teams with only knowledge of machine learning to get a better understanding of the domain of cyber security and the challenges it provides with the requirement of high accuracy models though the dataset is highly biased. It also helps the teams on the other side of the spectrum who are from cyber security to get to understand how machine learning models can be used to the greater benefit.

# **TABLE OF CONTENTS**

**CHAPTER NO. TITLE PAGE NO.**

**LIST OF FIGURES**

**1 PROBLEM STATEMENT** 1

**2 PROBLEM DESCRIPTION** 1

**3 LITERARY SURVEY** 1

**4 INTRODUCTION** 2

**5 WORKFLOW** 2

**6 SYSTEM ARCHITECTURE** 3

6.1 DATA COLLECTION 4

6.2 LABEL CREATION 4

6.3 FEATURE ENGINEERING 4

6.4 ALGORITHM SELECTION 4

6.5 PERFORMANCE EVALUATION 4

**7** **IMPLEMENTATION** 4

7.1 DATA LOADING 5

7.2 CLEANING 5

7.3 WEB SCRAPING 5

7.4 CWE CODE ANALYSIS 6

7.5 CVSS SCORE MAPPING 7

7.6 UNSUPERVISED LEARNING 8

7.7 PERFORMANCE ANALYSIS 9

7.8 CVE-2017-5638 11

**8 CONCLUSION** 11

**REFERENCES v**

iii

# **LIST OF FIGURES**

**FIGURE. TITLE PAGE NO.**

5.1 General outline of workflow 2

6.1 Detailed System Architecture 3

7.1 Primary CWE Code by Incidence in 6

2017 NVD Data

7.2 Secondary CWE Code by Incidence in 6

2017 NVD Data

7.3 Distribution of CVSS 3.0 Base Score in 7

2017 NVD Data

7.4 Primary CWE Code 8

7.5 Usefulness of Various Cluster Numbers in 9

Analyzing NVD Data

7.6 CVSS 3.0 score for Clusters 9

7.7 CVSS 3.0 impact and exploitability score 10

# 

# 

# 

# 

# 

# 

# 

iv

# **REFERENCES**

1. A user-centric machine learning framework for cyber security operations center <https://ieeexplore.ieee.org/document/8004902>
2. Downloading and unzipping a .zip file without writing to disk <https://stackoverflow.com/questions/5710867/downloading-and-unzipping-a-zip-file-without-writing-to-disk>
3. Parsing JSON Dataset with Pandas <https://www.geeksforgeeks.org/pandas-parsing-json-dataset>
4. Filtering Pandas DataFrames on dates <https://stackoverflow.com/questions/22898824/filtering-pandas-dataframes-on-dates>
5. Beautiful Soup to parse url to get another urls data <https://stackoverflow.com/questions/4462061/beautiful-soup-to-parse-url-to-get-another-urls-data>
6. Adding a y-axis label to secondary y-axis in matplotlib <https://stackoverflow.com/questions/14762181/adding-a-y-axis-label-to-secondary-y-axis-in-matplotlib>
7. Rotate axis text in python matplotlib <https://stackoverflow.com/questions/10998621/rotate-axis-text-in-python-matplotlib>
8. Become a Machine Learning Engineer <https://www.udacity.com/course/machine-learning-engineer-nanodegree--nd009t>
9. National Vulnerability DatabaseCVE-2017-5638 Detail <https://nvd.nist.gov/vuln/detail/CVE-2017-5638>

v