Colab Cloud Based Portable and Shareable Hands-on Labware for Machine Learning to Cybersecurity

Dan Lo[1], Hossain Shahriar[2], Kai Qian[1], Michael Whitman[3], Fan Wu[4]

[1]Department of Computer Science, [2]Department of Information Technology, [3]Institute for Cyber Workforce Development, Kennesaw State University, Marietta, GA, USA
[4]Department of Computer Science, Tuskegee University, AL, USA
{dlo2, hshahria, kqian, mwhitman}@kennesaw.edu, fwu@tuskegee.edu

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671541

Abstract- **Machine Learning (ML) analyze, and process data and develop patterns. In the case of cybersecurity, it helps to better analyze previous cyber attacks and develop proactive strategy to detect, prevent the security threats. Both ML and cybersecurity are important subjects in computing curriculum but ML for security is not well presented there. We design and develop case-study based portable labware on Google CoLab for ML to cybersecurity so that students can access, share, collaborate, and practice these hands-on labs anywhere and anytime without time tedious installation and configuration which will help students more focus on learning of concepts and getting more experience for hands-on problem solving skills.**

Keywords: ML, Algorithm, Cybersecurity, Education

I. INTRODUCTION

Security threats are evolving and getting more hidden and complicated. Detecting malicious security threats and attacks have become a huge burden to our cyberspace. We should apply proactive prevention and early detections of security vulnerabilities and threats rather than patching security holes afterwards. To analyze the huge amount of data to find out suspicious behaviors, threat patterns, and vulnerabilities and to predict and prevent future cybersecurity threats is a challenge. Today the Machine Learning plays a very important role in cybersecurity. According to Information Data Corporation (IDC), artificial intelligence (AI) and ML grow very fast from $8 billion in 2016 to $47 billion in 2020. As shared by Google, 50-70% of emails on Gmail are spam. With the help of ML algorithms, Google is making it possible to block such unwanted communication with 99% accuracy. Apple is also taking advantage of ML to protect its users' personal data and privacy. Here, we cover the applications of ML in cyber security [1]. Machine Learning (ML) is a powerful instrument to take up such challenge. Both ML for big data analysis and cybersecurity are widely taught in colleges and universities in these days.

However, there is a shortage of teaching and learning materials on ML for Cybersecurity. We observe the scarce open source portable hands-on handy labware for ML for cybersecurity. The challenges in offering hands-on labs include: the configuration of open source hands-on real labs; scarce dedicated staff and faculty in this field; and the excessive time needed for developing open source materials and projects. To overcome the above difficulties, we present an open sourced, portable, modular, and easy-to-adopt approach to enhance ML for Cybersecurity education through the development of online enable portable hands-on labware which consists of multiple modules covering spam email filtering, financial fraud prediction, network Denial of Service (DoS) prediction, website phishing detection, malware classification and prevention, m, intrusion detection, and others with various machine learning techniques such as statistical based learning and deep learning. Each module supports hands-on engagement learning cycle which consists of pre-lab activity for conceptualization and getting started with a Hello World example, hands-on lab activity for doing via concrete hands-hands experience with real-world data sets, and post add-on lab activity for creative enhancement.

Students will not only have the opportunity to look insight the typical and common security cases and analyze the flaws and threats, but also can learn ML key concepts and get hands-on problem solving experience in preventing and predicting suspicious security attacks and threats with suitable ML techniques.

The portable labware are designed, developed, and deployed on the open source Google CoLaboratory (CoLab) environment where learners can access, share, and practice all labs interactively and collaboratively with browser anywhere and anytime without tedious installation and configuration. Also, the hands-on lab modules support a wide audience to effectively learn

Authorized licensed use limited to: Kennesaw State University. Downloaded on August 05,2022 at 16:53:24 UTC from IEEE Xplore. Restrictions apply.

the subjects and result in more efficient student learning and engagement which help to enhance the cybersecurity curricula across computing discipline integrated with data science, engage student active learning and problem solving capability.

Many schools offer ML and cybersecurity courses in their computing curriculum but application of ML for cybersecurity is not well presented in our current curriculum. Not only the ML is an important technique to cybersecurity but also learning it will help to understand both subjects better and gain knowledge and skills well.

The hands-on learning is to gain knowledge and experience by actually doing something rather than just lessoning lectures or reading books. The hands-on learning is the preferred method in cybersecurity education where Problem-solving skills need to analyze real-world problems which involve critical thinking and hands-on practice. Lab environment will lead to effective education and training which help student gain hands-on experience and prepare students for cybersecurity workforce [2].

II. LABWARE DESIGN

The primary goal of the learning model is to create an engaging and motivating learning environment that encourages all students in learning emerging technologies. It provides students with hands-on experiences in solving real-world security problems. Each topic consists of pre-lab, hands-on activity lab, and post-lab (Pre/Lab/Post) activities.

The pre-lab for conceptualization and getting started. It presents foundamental concepts on a subject which iintroduce a specific cybersecurity study case with the root of security threats and overviews ML solutions such as prevention and detection. A simplified ”Hello world” example with corresponding ML solution is demonstrated such that students can watch, observe and get perspective insight processing which prepares students with a simple case for conceptual understanding and getting started experience with ML solutions for such cybersecurity case.

The hands-on activity labs are designed, developed, and deployed on the open source Google CoLab collaboration platform which is an in-the-browser environment with free Google cloud service. It provides learners with an interactive and easy to use platform for deep learning researchers and engineers to work on their data science projects such as cybersecurity project. Students only need a Google account to access CoLab and use GPU facility and works will be saved on Google drives. The Colab allows students to practice the lab without tedious installation and configuration and run the lab anywhere and anytime with any mobile devices or laptops. Upon completion the hands-on activity lab based on real world case study students will get perception into the ML solution for cybersecurity and gain hands-on experience for problem solving. Students will get insight in problem solving practice hands-on lab to learn how to use ML to solve a specific cybersecurity problem via a real world study case.

The post add-on lab extends lab practice, requiring independent analysis and assessment in developing further solutions. It promotes reflective and active thinking on the given case and hands-on doing for enhancement of problem solution such as to improve the prediction and detection accuracy rate with new creative ideas and active testing and experiments. It can help students build self-efficacy by observing peer performance, strengthening their confidence, and promoting their creativity. Students are encouraged to apply ML in new cases and share their works with others.

This learning model is summarized as following steps:

1. Pre-lab: Initiate/overview concepts through pre-lab instructions
2. Hands-on activity lab: Engage/explore problem solving with the hands-on activity lab on real-world problems
3. Post add-on lab: Reflect on the real-world problem and proposed solution in post-lab
4. Repeat steps 1-3 on various algorithms or scenarios.

III. SAMPLE MODULE ON NAIVE BAYES FOR SPAM EMAIL FILTERING

We use Naive Bayes algorithm, a popular and powerful algorithm in machine learning, to detect email spam as an example module of ML for cybersecurity.

A. Pre-Lab

The Pre-Lab shown in Fig. 1 gives some introductions about cybersecurity concern, attacker’s strategy and

Authorized licensed use limited to: Kennesaw State University. Downloaded on August 05,2022 at 16:53:24 UTC from IEEE Xplore. Restrictions apply.

method, the consequences after attacking and helps students build a basic concept of why these cybersecurity issues need to fix by using the machine learning algorithm.

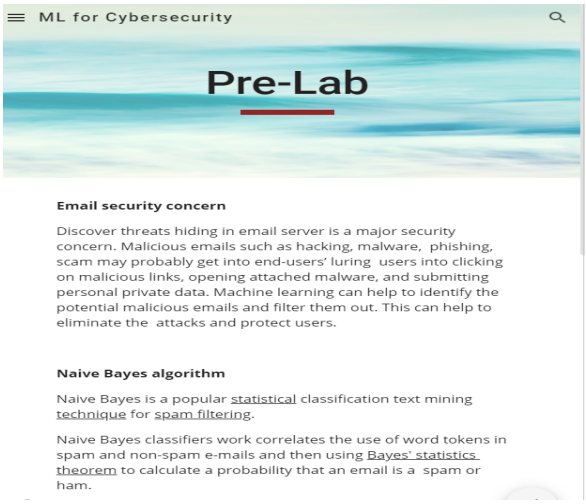


Figure 1: Pre-lab of Naive Bayes for email spam detection

The definition and brief synopsis are provided in the following paragraph of the website page.

We introduce the Naive Bayes Theorem which is the core of the Naive Bayes Algorithm. The formula of the Naive Bayes theorem is:

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

in this above equation, P(B) is the probability of the evidence P(A) is the prior probability(The probability of hypothesis A is true) P(B|A) is the probability of the evidence B given that hypothesis A is true P(A|B) is the probability of the hypothesis A given that the evidence B is true. Naive Bayes Theorem is used in Naive Bayes classifier. The probabilities for each class in the given dataset will be predicted and the highest probability will be the prediction. This process is called Maximum A Posteriori (MAP). We assume everything is independent given the class label.

We provide a "Hello-world" example for Naive Bayes Algorithm to let students have getting started the experience with ML solutions for such cybersecurity case. A small training dataset with 3 columns(B:go party, C:play game, D: study exam) and one target column(A: pass/fail exam)consists of few student samples of exam records in the past. Assume we want to predict a student whether he will pass the exam. He will not go to party(A)=0, not play video games(B)=0, and will study for the exam (D) = 1). First, we calculate the passing exam probability for this student:

P(A=1|B=0,C=0,D=1) = P(B=0|A=1)*P(C=0|A=1)*P(D=1|A=1)*P(A=1)/P(B=0,C=0, D=1) = 0.75*0.5*0.75*0.5/P(B=0,C=0, D=1) = 0.140625/P(B=0,C=0, D=1)

Then we calculate the probability for this student to fail the exam:

P(A=0|B=0,C=0,D=1) = P(B=0|A=0)*P(C=0|A=0)*P(D=1|A=0)*P(A=0)/P(B=0,C=0, D=1) = 0.5*0.5*0.25*0.5/P(B=0,C=0, D=1)= 0.03125/P(B=0,C=0, D=1)

As we can see, these two calculations have the same denominator. The value of P(A=1|B=0,C=0,D=1) is greater than P(A=0|B=0,C=0,D=1).With the Naive Bayes algorithm, we predict that this first student will pass the exam. The output by Python is shown in Fig 2. We also predicted the second student who go to party(B=1), pays games(C=1), and without studying(D=0) will fail the exam.

```
The two examples are predict(1means pass, 0 means fail on exam)
The first example we predict it is: 1
The second example we predict it is: 0
```

Fig 2. "Hello-World" for Naive Bayes Algorithm

We also introduce the advantages and disadvantages of the Naive Bayes algorithm to help students understand the algorithm as well as implement the algorithm for a suitable case. Our coding environment for the Naive Bayes algorithm is Google co-lab, which provides a setup free environment for students to practice.

B. Hand-on activity Lab

The Hand-on Activity Lab has the step by step instruction and some explanation for the coding phase. Students could practice the hands-on activity lab on their laptop and get the perception of ML solution implementation for cybersecurity and gain hands-on experience for using machine learning to solve the cybersecurity problem.

Authorized licensed use limited to: Kennesaw State University. Downloaded on August 05,2022 at 16:53:24 UTC from IEEE Xplore. Restrictions apply.

Figure 3 Hand-on activity lab

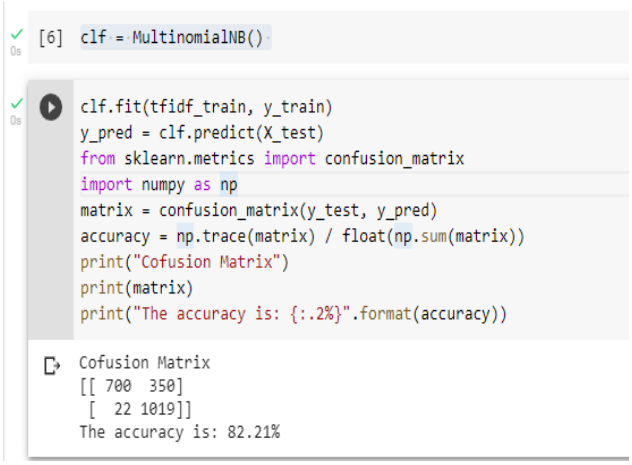
We also have screenshots for each step which help student practice more directional with the visual indication. The Email Spam Dataset has 5728 (4360 ”ham” and 1368 “spam”) instances. We split the dataset into two parts: the training set of 4296 examples and the test set of 1432 examples about 25% of the whole data set. The following text is a specific example of the spam email.

Subject: claim a free $ 25 kmart (r) gift card! you are receiving this mailing because you are a member of sendgreatoffers . com and subscribed as : jm @ netnoteinc . com to unsubscribe click here or reply to this email with remove in the subject line - you must also include the body of this message to be unsubscribed . any correspondence about the products / services should be directed to the company in the ad . % em % jm @ netnoteinc . com % / em %

This dataset we provided for testing is from Kaggle. We do some data pre-processing before fitting the dataset to the classifier and split the data into the training set and test set. Finally, we calculate the accuracy 98.95% and making the confusion matrix.

C. Post add-on Lab

This lab promotes reflective thinking on the email spam detection and hands-on doing for enhancement of Naive Bayes algorithm. We give students some questions such as how to improve the accuracy of the email spam detection by using pre-processing method. The learners are encouraged to find more effective and powerful algorithms in machine learning. With creative new ideas and active testing and experiments, post-add on lab will enhance students active learning, problem-solving learning, and life long learning for their creativity. Here is a student add-on lab on Naive Bayes text classification for fake news detection. The fake news often imposes false and/or exaggerated claims on social media and machine learning can effectively detect a fake news. The dataset has columns of news title and news text, and the target column with labels of FAKE or REAL. The task is to classify a news item into true real news or false/fake news. Figure 4 shows part of python code and execution results.



```
[6] clf = MultinomialNB()

clf.fit(tfidf_train, y_train)
y_pred = clf.predict(X_test)
from sklearn.metrics import confusion_matrix
import numpy as np
matrix = confusion_matrix(y_test, y_pred)
accuracy = np.trace(matrix) / float(np.sum(matrix))
print("Cofusion Matrix")
print(matrix)
print("The accuracy is: {:.2%}".format(accuracy))

Cofusion Matrix
[[ 700  350]
 [  22 1019]]
The accuracy is: 82.21%
```

Figure 4: Results from python code

As we can see the accuracy is 82.21% for this fake news detection with Naive Bayes machine learning.

IV. CONCLUSION

The overall goal of this labware is to address the needs and challenges of building capacity with ML for cybersecurity and the lack of pedagogical materials and real-world hands-on practice learning environment through effective, engaging case study based learning approaches. The project will help students and faculty to know what should be considered in proactive cybersecurity problem solving with ML. The student preliminary feedbacks on some modules are positive that students learn from the concepts and practice the skills through the hands-on labs.

ACKNOWLEDGMENT

The work is partially supported by the U.S. National Science Foundation Awards #2100134, #2100115, #1723578, #1723586. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Authorized licensed use limited to: Kennesaw State University. Downloaded on August 05,2022 at 16:53:24 UTC from IEEE Xplore. Restrictions apply.

REFERENCE

[1] APPLICATIONS OF ML IN CYBER SECURITY YOU NEED TO KNOW ABOUT, TECHNOLOGY INDUSTRY TREND, https://apiumhub.com/tech-blog-barcelona/applications-machine-learning-cyber-security/, 2018

[2] Why Hands-on Skills are Critical in Cyber Security Education, https://www.cybintsolutions.com/hands-on-skills-in-cyber-security-education/, 2018,

[3] The Role of Machine Learning in Cybersecurity, https://www.securitymagazine.com/articles/90064-the-role-of-machine-learning-in-cybersecurity, 04/2019

[4] Dan Liebermann, MACHINE LEARNING IN CYBER SECURITYFact, Fantasy, and Moving Forward, cybersecurity Summit, 12/2018

[5] Rishabh Das ; Thomas H. Morris, Machine Learning and Cyber Security, 2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE), 2017,IEEE Explore, https://ieeexplore.ieee.org/document/8526232

Authorized licensed use limited to: Kennesaw State University. Downloaded on August 05,2022 at 16:53:24 UTC from IEEE Xplore. Restrictions apply.