A Project Report
on
DETECT MALICIOUS ACTIVITY TO STOP ATTACKS USING MACHINE LEARNING

by

ACCESS DENIED

TEAM MEMBERS

KEERTHIKA S
KENISHIYA S
SHAKITHYA M
AARKA CHRISTAL SUJAA G

(Bachelor of Engineering in Computer Science and Engineering)

ABSTRACT

This project aims to present the functionality and accuracy of five different machine learning algorithms to detect whether an executable is infested or clean. So, we have created a extension file for this problem statement. It will present a description of the phenomenon of Malware, software programs or pieces of code that aim to hijack computer systems to steal information or to destroy it. We will delve deeper into this topic in order to have some understanding of these malicious programs. After a brief introduction to this phenomenon, we will present the evolution of malware over time. The following is the presentation of the different protection techniques. We will further discuss the importance of machine learning in addressing this situation. The algorithms used in this project as well as their benefits will be presented. Machine learning is widely used in this field by antivirus and antimalware programs as well as by these malicious programs, for example SQL Injections uses machine learning algorithms to encrypt itself in a differently each time it infests a new mass, becoming increasingly difficult to detect. At last, it deals with the implementation of complete project. The design of templates and static files that is HTML, CSS, python and JavaScript.

TABLE OF CONTENT

1) INTRODUCTION

1.1 MALWARE DEFINITION:

"Malware" is an abbreviation for "malicious software", it is used as a single term to refer to Viruses, Trojans, Worms, etc. These programs have a variety of features, such as stealing, encrypting or deleting sensitive data, modifying or hijacking basic computer functions, and monitoring computer activity. show user permission.

1.2 MACHINE LEARNING DEFINITION:

Machine Learning is a category of algorithms that allow software applications to predict much better results without being specifically programmed. The basic premise of machine learning is to build algorithms that receive input data and use statistical analysis to predict output data while output data is updated like many input data become valid. The processes involved in machine learning are similar to the processes of data mining and predictive modelling. Both require searching for certain patterns by date, and adjusting program actions accordingly. Many people are also familiar with machine learning from internet shopping and the advertisements that are shown to them depending on what they are buying.

1.2.1 HOW MACHINE LEARNING WORKS:

Machine learning algorithms are categorized as both supervised and unsupervised.

1.2.1.1 SUPERVISED ALGORITHMS

They require a data researcher, or data analyst, who has the knowledge of machine learning to supply the desired input and output data, in addition to delivering feedback on the accuracy of the predictions; acute during algorithm training. Some popular examples of supervised machine learning algorithms are: Linear regression for regression problems.

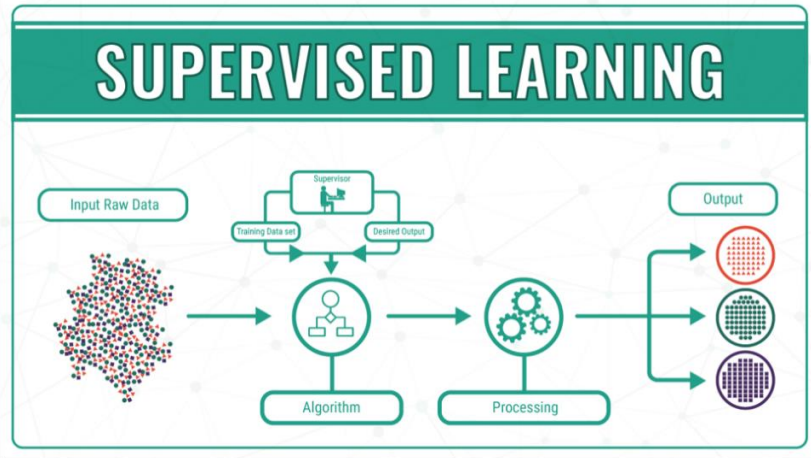


Figure 1 Supervised learning

1.2.1.2 UNSUPERVISED ALGORITHMS

They do not need training with output data. Instead, they use a method called deep learning to review the date and come to conclusions. Unsupervised and learned algorithms, also known as neural networks, are used for more complex processes than supervised algorithms, which include image recognition, speech-to-text, and natural language generation Some popular examples of unsupervised learning algorithms are :k-means for clustering problems ,Apriori algorithm for association rule learning problems.

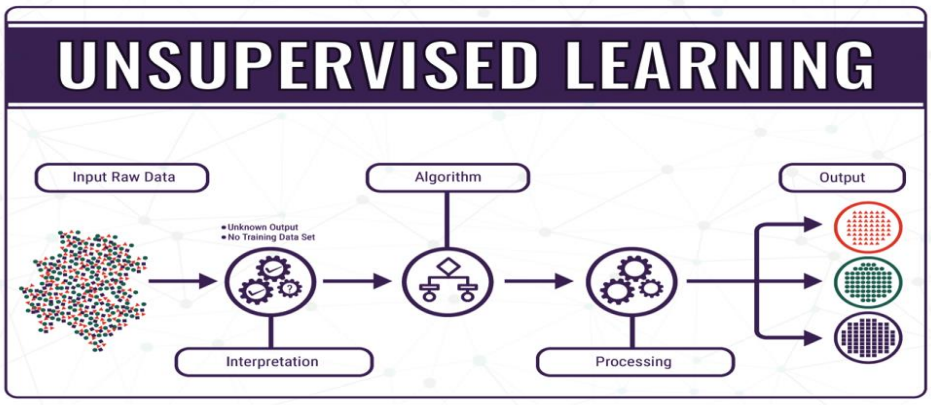


Figure 2 Unsupervised Learning

1.3 ALGORTIHM USED IN THIS PROJECT

DECISION TREE

- It is a supervised learning technique that can be used for both classification and Regression Problems but mostly it is preferred for solving Classification problems
- It is graphical representation for getting all the possible solutions to a problem/ decision based on given conditions.

RANDOM FOREST

- It is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset
- Using this the greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting

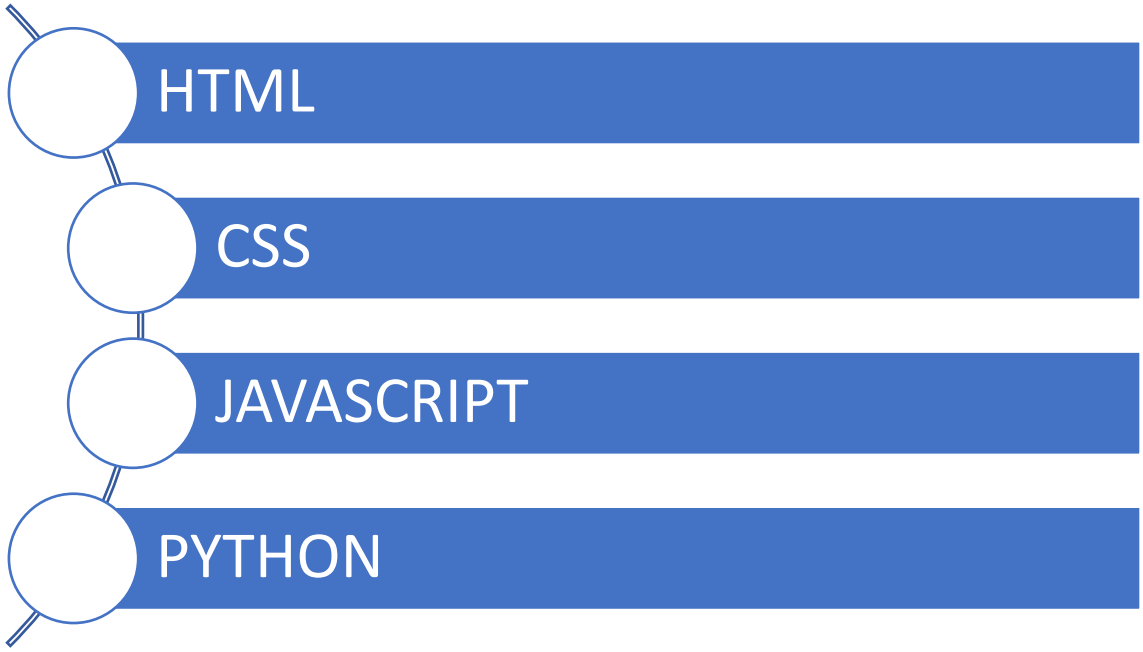
SVM

- It is Support Vector Machine is to create the best line or decision boundary that can segregatte n-dimensional space into classes so that we easily put the new data point in the coorect category in the future.
- SVM Algorithm can be used for Face detection, image classification, text categrization etc.

XG BOOST

- which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree(GBDT) machine learning library.
- It provides parallel tree boosting and is the leading machine learning library for regression, classification and ranking problems.

1.4 LANGUAGES USED IN THIS PROJECT



1.5 PURPOSE:

1.Alert message will be displayed while malicious attacks are excepted.

2.To track and warn about malicious attacks using JavaScript in a web browser to target user at application layer.

3. It is used to secure identity from phishing

4. It is used to detect malicious activity in web pages

5. If some malwares are detected then precautions are notified to avoid malicious activities.

2) LITERATURE SURVEY

2.1 EXISTING PROBLEM

- URL Checker will check the website link and displays its result as whether this is a scam website or a safe website but there is no such need in our problem solution it automatically gives an alter message any URL scanner.

- The transparency Report Service offers a standard field into which only can enter the URL you're concerned about. A few seconds later, the results- captured by Google's web crawlers- will tell you if the site can be trusted.

- Every website must be registered in the name of the individual or legal entity. Companies like Whois Lookup allow you to search, in the Whois field, for the ownership of a websites.

2.2 PROPOSED SOLUTION

- In this project will be able to add more dataset according to our use for detecting the malware.

- This is extension project which can be installed in every computer.

- Then using this project will be able to find where the websites is safe or not without any phishing of personal details.

- It can be installed easily and no background details or cost is required.

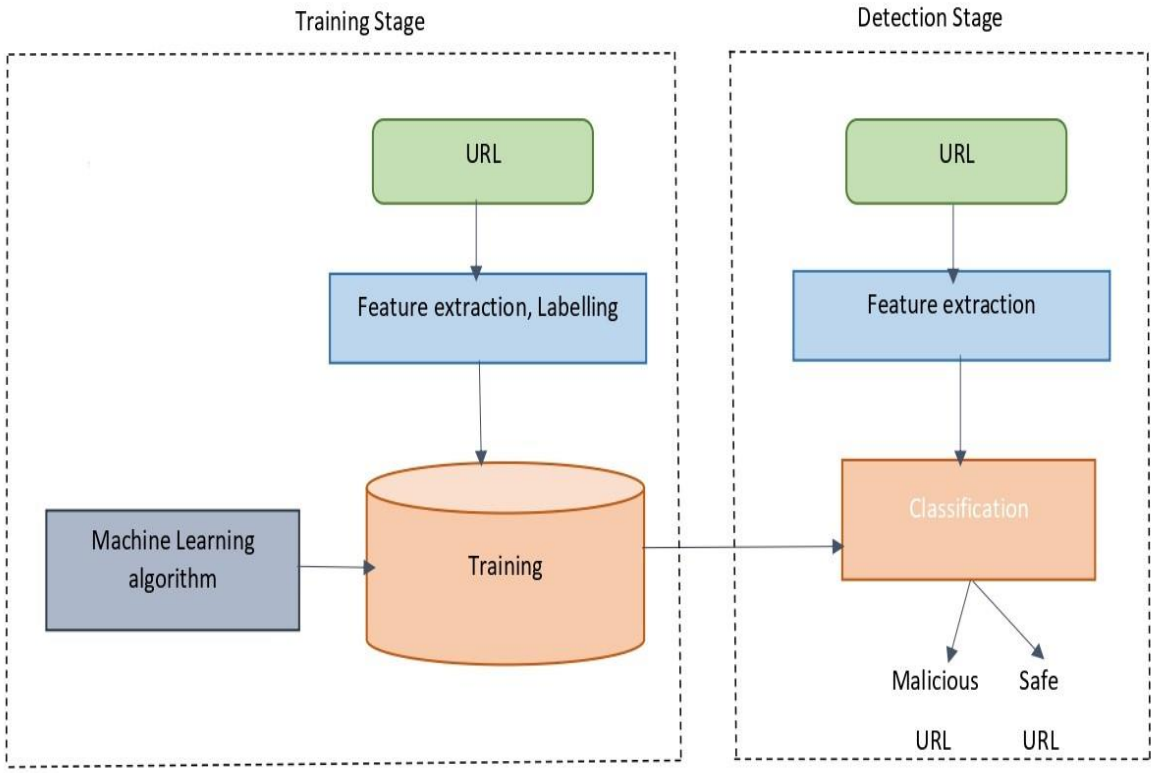
- This will help us to identify where the websites are safe or not in that time before malware attacking.

- An alert message will be notified to the users while entering to any websites.

- It is extension which can be installed in every personal system and it will not affect their system which is portable.

3) THEORITICAL ANALYSIS

3.1 BLOCK DIAGRAM.



3.2 HARDWARE / SOFTWARE DESIGNING

3.2.1 MATERIALS REQUIRED:

1) COMPUTER:

A base machine with windows 8, 8 GB RAM, 250 GB of hard drive with core i5, installed

2) OPERATING SYSTEM:

The main OS to be used is Windows 10

3) REPORTING TOOLS:

1) VScode as IDE
2) Github for version control and remote work.

3.2.2 OVERALL PROJECT IS PARTITIONED INTO VARIOUS PHASES.

a) PHASE 1: Downloading 1,38,000+ dataset of malicious and legitimate PE files

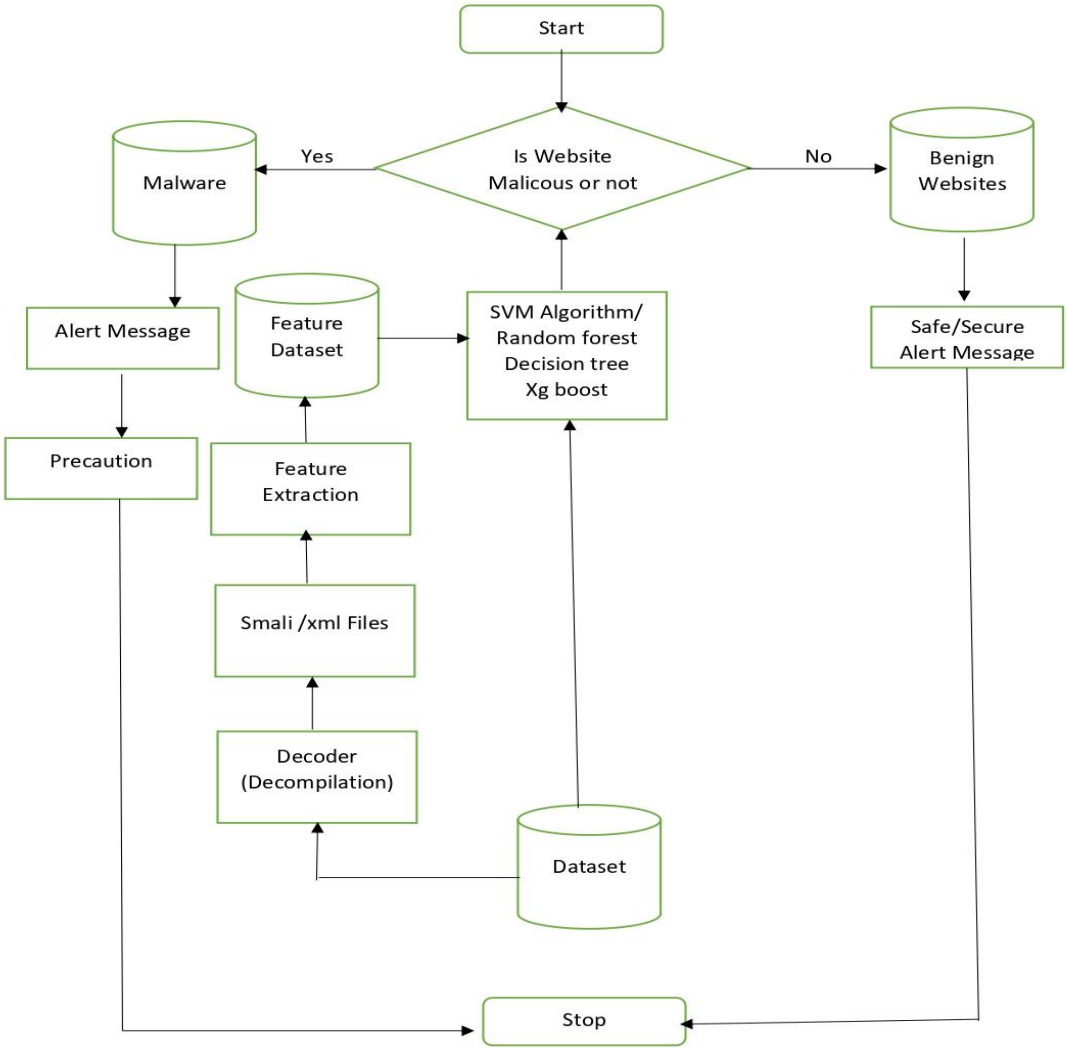
b) PHASE 2: Performing Malware analysis and classification, choosing best out of 4 ML algorithms and feature classification

c) PHASE 3: Deploying Full stack Flask framework using HTML, Python, CSS, JavaScript for static analysis

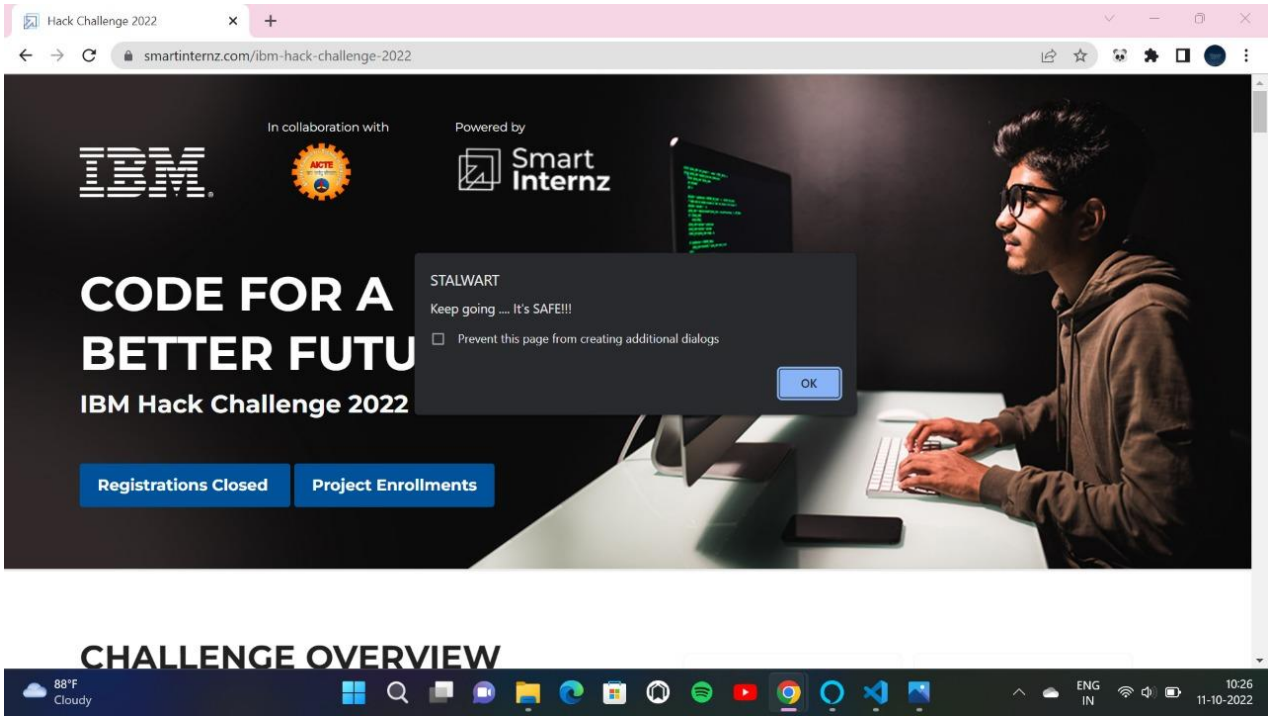
4)EXPERIMENTAL INVESTIGATIONS

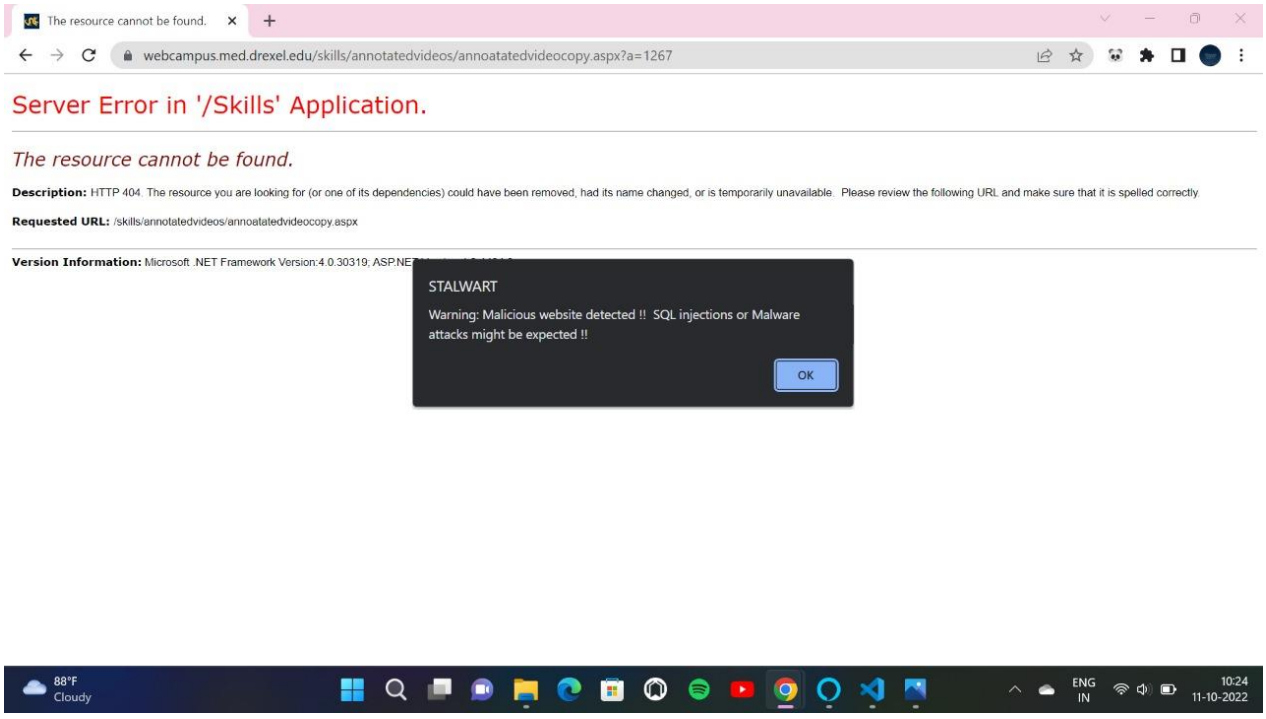
- While working on this project we came across some of the problems where we were not able to detect the URL of the certain websites. Then we made some changes on the code to detect the URL.
- Initially we had very smaller number of datasets, which resulted in limited activities. Then we added some more datasets to get the desired output.
- Through this project we came across many malicious attacks where we have selected only few and main malicious attacks.
- And choosing algorithm for this project was quite difficult because of accuracy of detecting the attacks.
- Then alert message was overlapped again and again to avoid that we made some changes in code then it worked well.

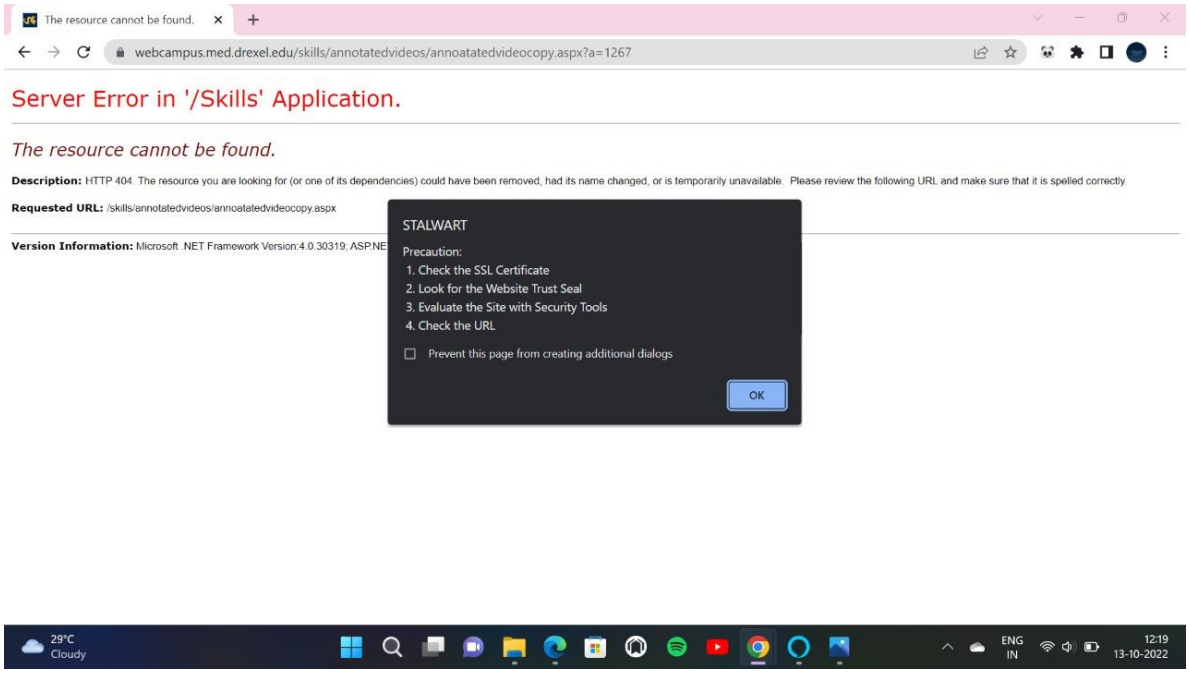
5) FLOWCHART


Start
Is Website
Malicous or not
Yes
No
Malware
Benign
Websites
Alert Message
Feature
Dataset
SVM Algorithm/
Random forest
Decision tree
Xg boost
Safe/Secure
Alert Message
Precaution
Feature
Extraction
Smali /xml Files
Decoder
(Decompilation)
Dataset
Stop

6) RESULT


Hack Challenge 2022
smartinternz.com/ibm-hack-challenge-2022
In collaboration with
Powered by
Smart Internz
CODE FOR A BETTER FUTU
IBM Hack Challenge 2022
Registrations Closed
Project Enrollments
STALWART
Keep going .... It's SAFE!!!
Prevent this page from creating additional dialogs
OK
CHALLENGE OVERVIEW
88°F Cloudy
ENG IN
10:26 11-10-2022


The resource cannot be found.
webcampus.med.drexel.edu/skills/annotatedvideos/annoatatedvideocopy.aspx?a=1267
Server Error in '/Skills' Application.
The resource cannot be found.
Description: HTTP 404. The resource you are looking for (or one of its dependencies) could have been removed, had its name changed, or is temporarily unavailable. Please review the following URL and make sure that it is spelled correctly.
Requested URL: /skills/annotatedvideos/annoatatedvideocopy.aspx
Version Information: Microsoft .NET Framework Version:4.0.30319; ASP.NET
STALWART
Warning: Malicious website detected !! SQL injections or Malware attacks might be expected !!
OK
88°F Cloudy
ENG IN
10:24 11-10-2022



7) ADVANTAGES

- The advantage of using machine learning is that it can determine whether a code or a file is malicious or not in a very small time
- Machine Learning is able to detect previously unknown malware with predictive capabilities and especially useful for the detection
- ML works because it can understand and identify malicious intent based solely on the attributes of a file — without prior knowledge of it, without signatures and without needing to execute the file to observe its behaviour.

DISADVANTAGES

- Machine learning displays a risk of running inefficient algorithms and making limited predictions when not trained properly.

8) APPLICATIONS

- Security, machine learning continuously learns by analysing data to find patterns so we can better detect malware in encrypted traffic, find insider threats, predict where "bad neighbourhoods" are online to keep people safe when browsing, or protect data in the cloud by uncovering suspicious user behaviour.
- Using this we can avoid using fake websites.
- This will secure personal computers and avoid phishing of data.

9) CONCLUSION

The aim of this project is to present a machine learning approach to the malware problem. Due to the sudden growth of malware, we need automatic methods to detect infested files. In the first phase of the work, the data set is created using infested and clean executables, in order to extract the data necessary for the creation of the data set, we used a script created in Python. After creating the data set, it must be ready to train machine learning algorithms. The algorithms used are: decision trees, Random Forest, SVM, Decision Tree and XG Boost presented comparatively. After applying the best accuracy algorithms, it had a Random Forest algorithm with an accuracy of 99.406012 %. This work demonstrates that Random Forest is the best algorithm for detecting malicious programs. In the future, this accuracy can be improved, if we add a much larger number of files in the data set to drive the algorithms. Each algorithm has several parameters that can be tested with different values to increase their accuracy. This project can reach the application level with the help of a library called pickle, to save what the algorithm has learned and then we can test a new file to see if it is clean or infected. Static analysis has also proven to be safer and free from the overhead of execution time. This project will helps in avoiding phishing and malicious attacking websites.

10) FUTURE SCOPE

- This can be made more accurate with adding more data set
- More algorithms with better performance can add on to accuracy
- It can be hosted on web for real time analysis of exe files on the cloud
- It can be also developed further using cyber security and store the list of attacks happened in time lap.
- Using this we can tell what attack is taking place
- In future we can also develop this to solve the malicious attack.
- This extension can be modified in later.

11) BIBILOGRAPHY

11.1 REFERENCES

https://www.sciencedirect.com/topics/computer-science/malware-detection#:~:text=Heuristic%2Dbased%20malware%20detection%20focuses,rather%20than%20patterns%20or%20signatures.

https://towardsdatascience.com/malware-detection-using-deep-learning-6c95dd235432

https://link.springer.com/chapter/10.1007/978-981-16-7618-5_53

https://en.wikipedia.org/wiki/Malware

11.2 SOURCE CODE

https://github.com/smartinternz02/SBSPS-Challenge-9347-Detect-malicious-activity-to-stop-attacks-using-Machine-Learning