Dissertation Submitted for the partial fulfillment of the B.Sc. as a part of M.Sc. (Integrated) Five Years Program AIML degree to the Department of AIML & Data Science.

Malware Detection Using Supervised Machine Learning

Submitted to

By

Esha Mishra
Semester-VI

M.Sc. (Integrated) Five Years Program AI&ML
Department of AIML & Data Science.
School of Emerging Science and Technology
Gujarat University

June, 2022

DECLARATION

This is to certify that the research work reported in this dissertation entitled "**Malware Detection Using Supervised Machine Learning"** for the partial fulfillment of B.Sc. as a part of M.Sc. (Integrated) in Artificial Intelligence and Machine Learning degree is the result of investigation done by myself.

Place: Ahmedabad Esha Mishra

Date:

ACKNOWLEDGEMENT

I would like to express my sincere gratitude towards Dr. Ravi Gor, co-ordinator School of Emerging Science and Technology at Gujarat University for giving me the opportunity to work on this project. Invaluable guidance under the supervision of Machine Learning faculties Rashmi Kumari and Eric Shah helped me a lot. I am also grateful towards department administration for their cooperation.

Last but not the least, I would like to express my appreciation towards my parents Indra Mishra and A.K Mishra along with my sisters Diksha Mishra and Varsha Mishra and my friends Shruti H. Agarwal, Radhika Sharma and Prableen Sandhu for their unparalleled support throughout this project.

-Esha Mishra

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

NLP - Natural Language Processing

VM - Virtual Machine

SVM - Support Vector Machine

CHAPTER I

ABSTRACT

Abstract

In this project supervised machine learning is used for malware detection in the system. Primary data has been used for this project. Dynamic malware analysis is performed, where malware is analyzed after executing it in the malware analysis lab. Flare VM on Windows 10 distribution is used as the sandbox. The log files are obtained by running both malware and goodware on Flare vm. Data is extracted from these log files using an NLP technique called bag of words and labeled afterward. This is followed by model training and evaluation. RandomForest, Decision Tree, Logistic Regression and SVM are the four different algorithms which were used to train the model. Among them, Random Forest gave the highest accuracy of 99.99806%.

CHAPTER II

INTRODUCTION

2.1 Background

Cyber security is essential for securing sensitive information, data, and patents in today's world of increasing digitization. Malware is short for malicious software that is designed to disrupt, damage or gain unauthorized access to any device or network. Examples of common malwares include viruses, worms, Trojan viruses, spyware, adware, and ransomware. Static and dynamic are two types of malware analysis. This project focuses on dynamic malware analysis, where malware files are executed in a controlled environment for analysis. The log files generated will be used to extract data and train the machine learning model.

2.2 Problem statement

With increasing cyber attacks, it's difficult for traditional programming methods to deal with them effeciently because of sheer volume and variety of malwares. ML can be used to detect malware in the system with greater accuracy.

2.3 Objective

The aim is to create a machine learning model to detect malware in the system based on generated log files.

2.4 Motivation and significance

Any compromise to cyber security has the potential to harm the organization both long and short term. Cyber security is an important component of a country's overall security. ML is good at dealing with huge amounts of data and overcomes limitations of conventional programming methods when dealing with cyber attacks.

CHAPTER III

REVIEW OF LITERATURE

3.1 Online Articles

Because of its ability to deal with large amounts of data, machine learning is ideal for boosting cyber security. The deployment of an unsupervised machine learning model to detect anomalies in network traffic and warn cyber security systems is common. Many businesses have escaped ransomware assaults. Financial organizations, such as banks, are increasingly relying on machine learning for cyber protection.

Malware is a term used to describe malicious software that is designed to disrupt, damage, or gain unauthorized access to a device or network. Viruses, worms, Trojan horses, spyware, adware, and ransomware are all examples of prevalent malware.

Malware analysis can be divided into two categories. The first is a static analysis, in which we examine a malware file without running it. We examine file signatures such as size, hash, and extension. Dynamic analysis is the second type. This entails running the malware file in a sandbox and observing how it behaves.

3.2 University Lectures on Youtube

Ricardo Calix used log file data to do virus detection. He ran roughly fifty samples of both goodware and malware in an isolated environment. Each program's log files were gathered. They were stored with good1, good2 or badrabbit1, badrabbit2 based on malware and goodware. He extracts data using a bag of words technique and labels it 1 (goodware) or -1 (malware). After that, model training takes place. This project takes a similar method, however the data is retrieved from two big log files from malware infected and non-infected systems, rather than several short log files each representing one goodware/malware program.

CHAPTER IV

METHODOLOGY

4.1 System Requirement

4.1.1 Software Tools

4.1.1.1 For Log File Generation

Malware Analysis Lab :

Kali Linux Host (preferable)

Oracle Virtual box

Window 10 VM

Flare VM

ProcMon

Malware samples from VirusBazaar and github

Goodware samples

4.1.1.2 For Data Extraction, Model Training And Evaluation

Google Collab

Code : Python

4.1.3 Hardware Specification

8 GB RAM for your system.

4.2 Modules

Pandas

Numpy

Seaborn

Matplotlib

Sklearn

4.3 System Diagram

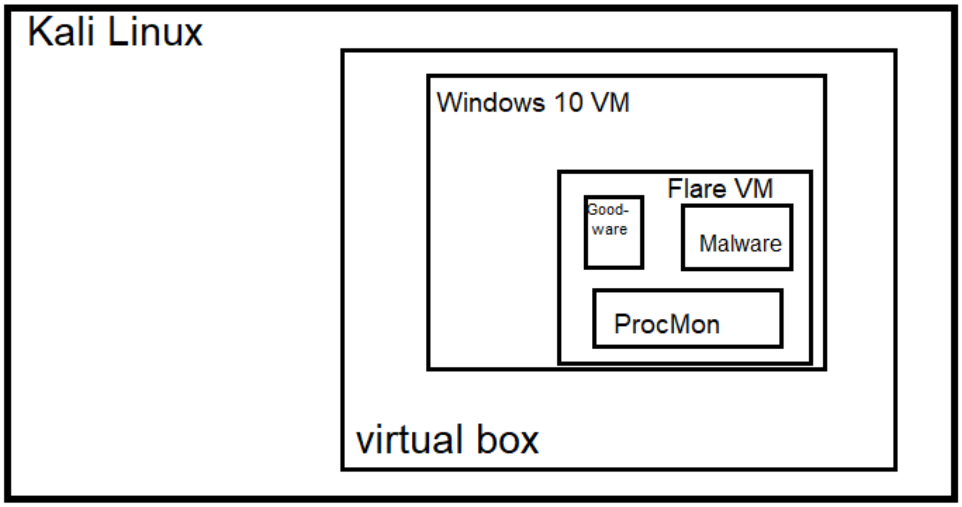

Fig 1 : System For Log File Generation

4.3 Algorithms Used

4.3.1 Decision Trees

A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. Decision Trees often give very good results.

4.3.2 Random Forest

Random forest is a supervised learning algorithm. It builds a forest with an ensemble of decision trees. It is an easy to use machine learning algorithm that produces a great result most of the time even without hyperparameter tuning. It combines the results of multiple decision trees thus it gives much better accuracy than the decision tree.

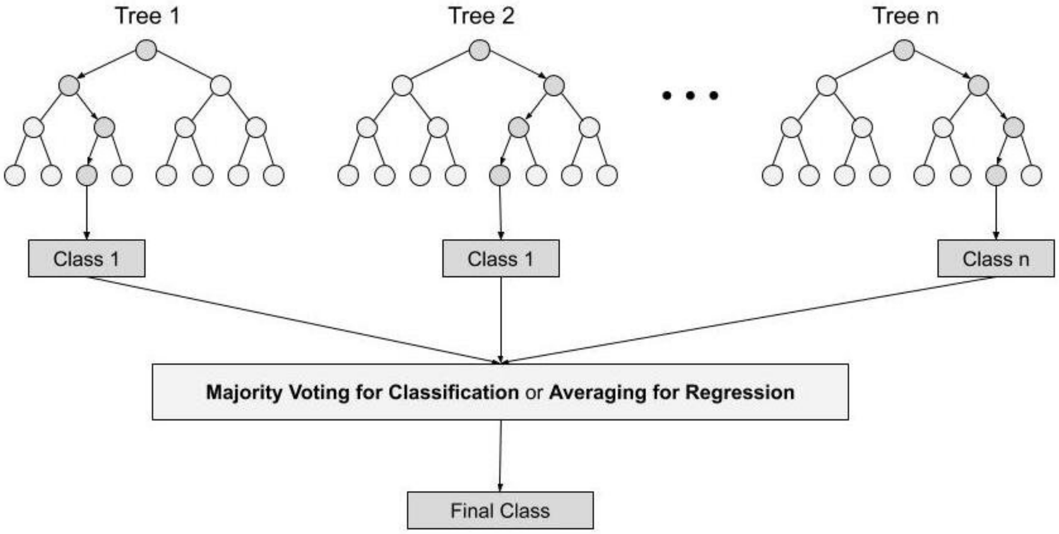
Tree 1
Tree 2
Tree n
Class 1
Class 1
Class n
Majority Voting for Classification or Averaging for Regression
Final Class

Fig 2: Random Forest Diagram

source: Analytics Vidhya

4.3.2.1 Pros

1. It overcomes the problem of overfitting by averaging or

combining the results of different decision trees.

2. Random forests work better for a large range of data items than a single decision tree does.
3. Random forest has less variance then single decision tree.
4. Random forests are very flexible and possess very high accuracy.
5. Scaling of data does not require a random forest algorithm.
6. Random Forest algorithms maintain good accuracy even if a large proportion of the data is missing.

4.3.2.2 Cons

1. Takes a lot of Storage to store the model.
2. It Took a lot more time and computation power.

4.3.3 Logistic Regression

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X. Since this is a binary classification, logistic regression is suitable for use.

4.3.4 SVM

The algorithm creates a line or a hyperplane which separates the data into classes. It shouldn't be used when there are overlapping classes or when there is too much noise in the data. This was thought suitable as there are a large number of features.

4.4 WorkFlow

4.4.1 Data Collection

4.4.1.1 Log file generation

1. Setting up a Malware Analysis lab and taking a snapshot.

2. Collecting Malware and Goodware Samples.

3. Run malware samples and save the log files as malware1.csv. Restore the lab to the previous snapshot.

4. Run goodware samples and save the log files as good1.csv.

4.4.1.2 Data Extraction from log files

Applying NLP technique bag of words for data extraction with the help of countvectorizer from sklearn to extract data. Also, labeling the dataset in the process.

4.4.2 Data Type

The final dataset obtained is in CSV format. It has 501 columns and 34,371 rows. The dataset is labeled 1 (goodware) and -1 (malware).

4.5 Model Training

Four models are trained and algorithms used are Decision Tree, Random Forest, Logistic Regression and SVM.

4.6 Model Evaluation

4.6.1 Model Accuracy

Model	Accuracy
Decision tree	99.83514%
Random Forest	99.99806%
Logistic Regression	99.9806%
SVM	99.71877%

Tabel 1: Models and Their Accuracy

4.6.2 Confusion Matrix And Classification Report of different Models

4.6.2.1 Decision Tree

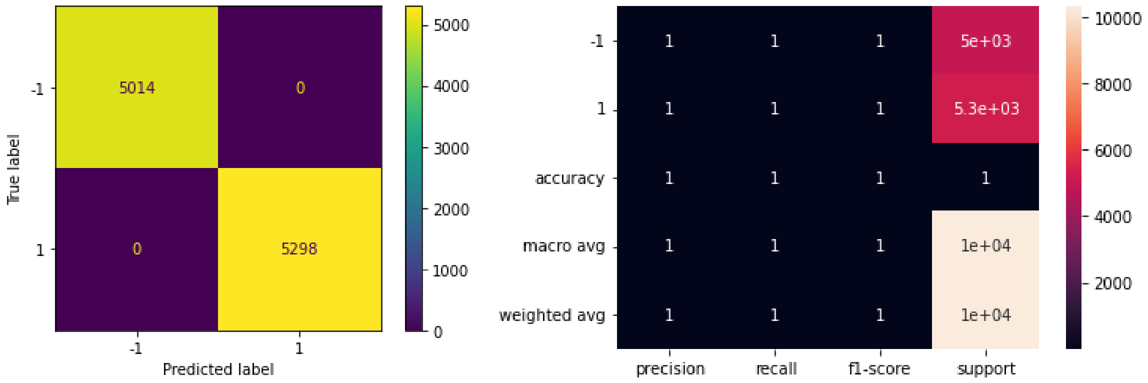

Fig 3: Confusion Matrix and Classification Report of Decision Tree Model

4.6.2.2 Random Forest

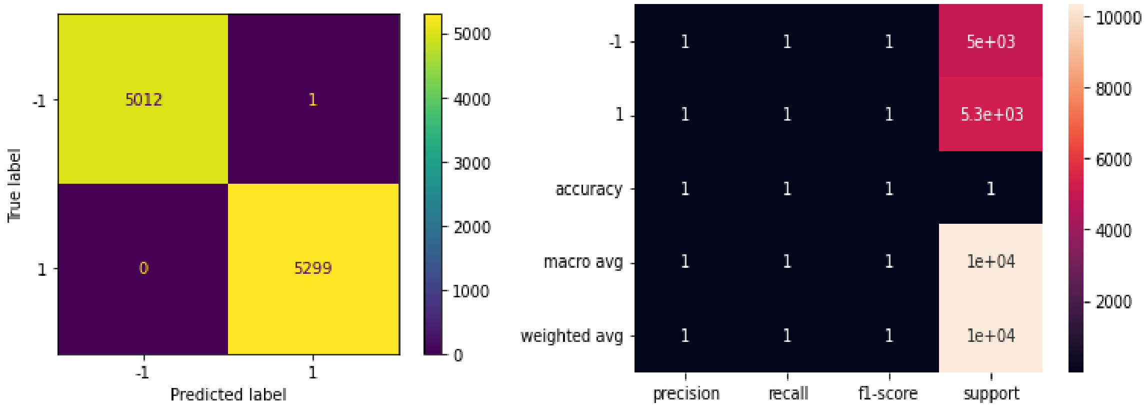

Fig 4: Confusion Matrix and Classification Report of Random Forest Model

4.6.2.3 Logistic Regression

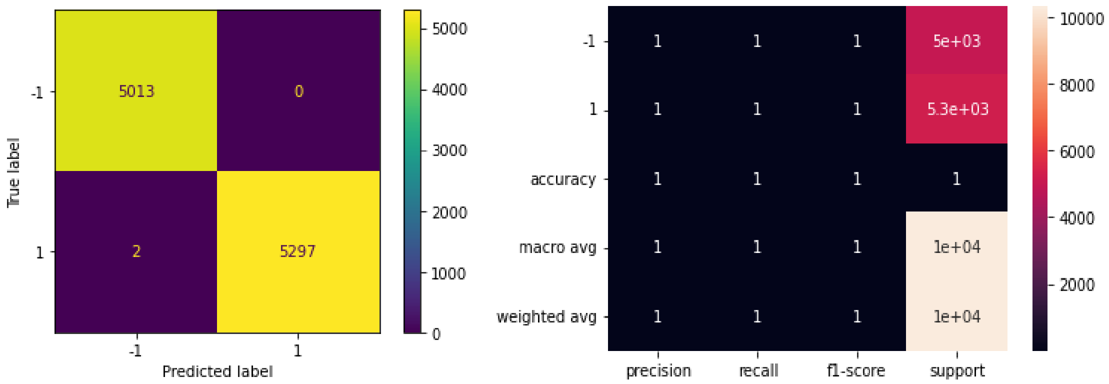

Fig 5: Confusion Matrix and Classification Report of Logistic Regression Model

4.6.2.4 SVC

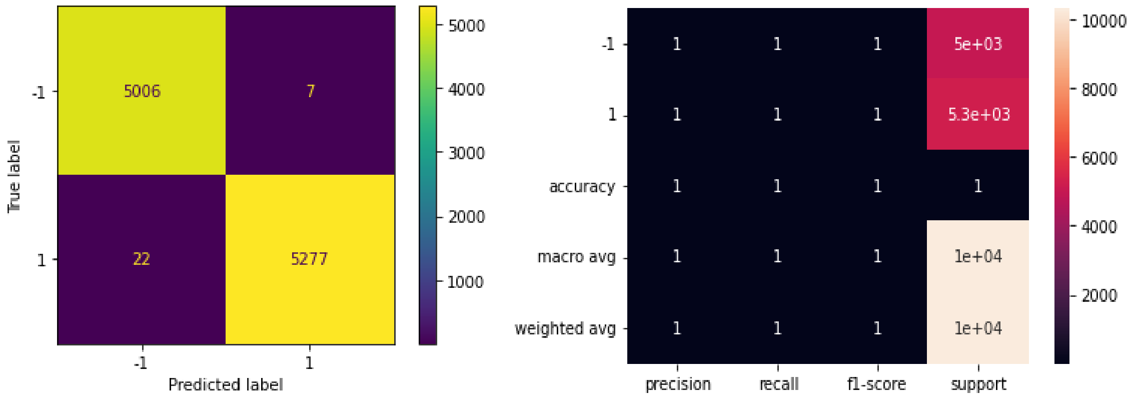

Fig 6: Confusion Matrix and Classification Report of SVM Model

CHAPTER V

RESULTS

Results

We have obtained four different machine learning models and all four models show accuracy of 99 per cent with minute differences. Random Forest has the highest accuracy followed by Decision Tree model, Logistic Regression model and SVM model.

CHAPTER VI

CONCLUSION

Conclusion

The aim of this project is to use ML to detect malware in the system. This helps us identify infected systems and improve cyber security at individual or organizational level.

6.1 Limitations

This malware detection model can only detect malware on which it has been trained on. This model also only helps in detection of malware and does not provide protection against malware or its removal. The data could have been trained on much more varied kinds of malware. This model only works for windows operating systems.

6.2 Cons of Random Forest Model

1. Complexity is the main disadvantage of Random forest algorithms.
2. Construction of Random forests is much harder and time consuming than decision trees.
3. More computational resources are required to implement the Random Forest algorithm.
4. It is less intuitive when we have a large collection of decision

trees.

5. The prediction process using random forests is very time-consuming in comparison with other algorithms.

6.3 Future Enhancements

1. Training model on more kinds of malware.
2. Integrating it with a system such that it can work live.
3. To improve accuracy, numerous models can be combined.
4. Built similar models for different operating systems.
5. Built an unsupervised version of this model.

BIBLIOGRAPHY

Baker K. (2022, January 4). *Malware Analysis Explained.* crowdstrike. https://www.crowdstrike.com/cybersecurity-101/malware/malware-analysis/

Hackesploit. (2019, Aug 10). *Malware Analysis Bootcamp - Setting Up Our Environment.* [video]. YouTube. https://www.youtube.com/watch?v=F1LE56QQ7iA

MalwareBazaar. [Malware Samples]. https://bazaar.abuse.ch/

Mandiant. (2021, Oct 23). *Flare-vm version 3.0. [software].* Github. https://github.com/mandiant/flare-vm

Ricardo C. (2019, Aug 5). *Machine Learning for Cyber Security: Lectures.* [Video Playlist], YouTube.https://www.youtube.com/watch?v=JxcBm7CRtI0&list=PL74sw1ohGx7GHqDHCkXZeqMQBVUTMrVLE

Ricardo C. (JUne 25, 2019), *Machine Learning for Cyber Security: Labs* [Video Playlist], Youtube. https://www.youtube.com/watch?v=lTge-G02Cis&list=PL74sw1ohGx7FE-DI18bOfi2X61zRE-wMd

Vectra AI. (2018, June 15). *Machine Learning Fundamentals for Cyber Security Pros. [Video].* YouTube. https://www.youtube.com/watch?v=uPSgfNhd2qY

Virus-Samples. (2021, Feb 6). *Malware-Sample-Sources.I[malware samples] Github. https://github.com/Virus-Samples/Malware-Sample-Sources*

What is malware ?. [webpage]. mcafree. https://www.mcafee.com/en-in/antivirus/malware.html

What Is Malware?.[webpage]. cisco.https://www.cisco.com/c/en_in/products/security/advanced-malware-protection/what-is-malware.html