**PROJECT REPORT**

**ON**

**MALWARE DETECTION AND ANANLYSIS**



Submitted By:

Raj Priyadarshi

B.Tech(CSE)

Sec: D RollNo: 38

U.RollNo: 2014795

**MOTIVATION:**

The interest in this project is because of my interest in cyber security. It is very useful if you know how to protect your computer and if it is hacked how to recover from it. At first I didn’t know it required machine learning but as I progressed through my project it started to make sense and I understood a little about Machine Learning as well.

**ACKNOLEDGEMENT:**

I want to thank the CSE department who gave this project as a choice. I also want to thank my teachers for clearing my doubts about the project. I am thankful for the internet that provided me with the datasets and the libraries needed for the project.

**INTRODUCTION TO MALWARE:**

In this age of computers where every computer system is connected to each other it’s difficult to trust one another.

Another person may send some harmful files that may cause damage to you. These files contain harmful programs called malwares. Malware are programs that run on your computer system and either harm your system or gather sensitive information about you i.e., your bank account number or login passwords.

There are different types of malwares which work in different ways according to the needs of the attacker. Some of whose objective is to destroy the computer system completely while others just steal information quietly.

**Types of Malware:**

* Spyware: It is a type of malware that functions by spying on user activity without their knowledge. The spying include activities monitoring, collecting keystrokes, data harvesting.

The information like account information, logins, financial data are most insecure as they are considered as sensitive data.

A spyware can enter the system through clicking and accessing data from various untrusted sources like pop-up ads, and downloading attachemnts from emails etc.

* Trojan Horse: It works by disguising itself as a normal file or program to trick users into downloading malicious software.

A trojan can give a malicious party remote access to an infected computer.

Once an attacker gains access to the infected system, they can steal data like login credentials and monitor user activity. Modify files etc.

* Virus: This malicious code or program alters the way a computer operates and is designed to spread from one computer to another. A virus operates by inserting or attaching itself to legitimate program or document that supports macros in order to execute its code. In the process, a virus has potential to cause unexpected or damaging effects, such as harming the system software by corrupting or destroying data.

A computer virus much like the real world virus replicates and spreads very fast from one computer to another.

They can spread through text messages, internet downloads from untrusted sources and internet scam links.

* Ransomware: It works by holding the infected system captive while demanding a ransom. The malware can also restrict user access to infected computer either by encrypting files on the hard drive or locking down the system and displaying messages to make the user pay ransom.

**LIBRARIES USED:**

* Pandas: It is python library for data analysis. It offers data structures and operations for manipulating numerical tables and time series.
* Numpy: It is a library for the python programming language, adding support for large, multi-dimensional array and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
* Sklearn: It stands for scikit learn. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the python numerical and scientific libraries.

**ALGORITHMS USED:**

* Random forest algorithm: It is Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Most important feature of random forest algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification.

* Gradient Boosting Algorithm: Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model.

**METHODOLODY:**

Import a dataset (.csv) containing legitimate files and malware files. Then with the help of scikit learn (sklearn) python library analyze the dataset and look for the malwares. Then differentiate the legitimate files from the malwares. After the output the accuracy of the algorithms used to find the malwares in the dataset.

**REFERENCES:**

* Book ‘Mastering Machine learning for Penetration Testing by Chiheb Chebbi.’
* Wikipedia