Mini Project Report on

MALWARE DETECTION USING MACHINE LEARNING

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

Student Name –PEARL BHUTANI **University Roll No. -2018986**

Under the Mentorship of

Siddhant Thapliyal
(Assistant Professor)

Department of Computer Science and Engineering
Graphic Era (Deemed to be University)
Dehradun, Uttarakhand
July 2023

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the project report entitled "**MALWARE DETECTION USING MACHINE LEARNING**" in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering of the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of "**Siddhant Thapliyal "** Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name –Pearl Bhutani University Roll no - 2018986

Table of Contents

Chapter 1

Introduction

Malware:

It is malicious software written with the intent of deleting or damaging any file/data/resources. It is the tool that is commonly used by criminals, to gain unauthorized access to a Computer System .

Common types of malware that are used to exploit the resources are :

1) ***Adware*: is malicious software that delivers unwanted pop-ups to the user's computer system or device. In the worst scenario, it is used to track users' behavior as well as users's online activity .**
2) ***Bug :* any flaw is termed as bug which is generally caused by the human errors . Which leads in system crashing or malfunctioning .**
3) ***Viruses :* it is a type of malicious software which has a tendency of self-replication, It multiplies with the passage of time . Used to spread over the system and damage the files , and even modify the data which has been accessed by the virus.**
4) ***Worms* : is a self-replicating malicious software used to spread over the network of the computer system and exploit security vulnerabilities, gain unauthorized access over the computer system .**
5) ***Trojans:* is a type of malicious software which disguise itself as a normal, easily downloadable file . They create backdoor to gain unauthorized access over the computer system .**

6) ***Ransomware :*** **It locks the victim's computer system and demands a ransom amount of money from the user to get access to their own computer System. It is the type of malicious software that cannot be predicted unless and until some symptoms are felt like system crashing, slower down of the system and many more. The ransom amount is accessed by the attacker through cryptocurrency (bitcoin), so that they could not get traced .**

Significance of malware in todays digital landscape :

Cybersecurity Threats : With the on going growth of malware it creates constant challenge for the cybersecurity professionals, to mitigate these malware-caused threats. Malware auses a significant threat to various computer system , devices to gain unauthorized access .

1) **Financial Losses: Threats which has been caused by talware(malicious software) results in substantial financial loss of individual , organization .**

   Malware like ransomware takes a ransom amount of money from the organization , individual to get access to their own computer system, if ransom aamountof money is not received by the criminal then it could result in damaging of the personal dat , or selling that data which has been accessed by the attacker .

2) **Privacy Breaches : Data accessed by the malware is often sensitive information , that leads to privacy breaches , data that has been accessed by the malware is**

used for identity threat, fraud, blackmailing . That leads to reputational damage as well as financial loss .

3) **Exploits vulnerabilities : malware the malicious software often exploits vulnerabilities in computer system . It uses Social engineering techniques , dictionary attack etc. mechanism to get access to any computer system .**

4) **Spread Globally : Malware can spread rapidly due to interconnection of networks and because of networks of networks(Internet) . If one computer system is infected with the malware the computer system , devices which are connected through that infected system can infect as quickly as possible .**

Need for cybersecurity measures in malware infected landscape :

To protect our computer system , organizations , individuals from a wide range of cyberattacks . This can include all active as well as passive attacks (Social Engineering attack , Denial of service Attack , Distributed Denial of service attack , Eavesdropping and many more attacks) . Cybersecurity measures to protect consumer data , proactive threat detection , avoid unauthorized access to a computer system . It acts as a layer between the computer system and the malicious software .

1) **Increase in Number of Sophisticated Hackers :**

   Hackers are being active with the passage of time , along with the bad mindset and the skills they can get unauthorized access to the computer system . To protect our data , sensitive information cybersecurity measures plays a vital role .

2) **Increase in cost of breaches :**

When an organization gets attacked by the attacker , it not only causes the financial loss to an organization as it results in threat to data but also it causes reputational damage to the organization as people will think twice before investing into that particular organization .

Preventing attacks from your side :

To prevent any attack not at high scale but by being at your own level ->

1) **Always keep your software , computer system updated as with the growth in technology more and more security measures are being implemented to protect user's data at the level it can be protected , avoiding unauthorized access to any system . By updating your software sometimes it can be helpful in detecting any vulnerabilities in your computer system , which can be cured if observed .**
2) **The key to avoid any attack is " VIGILANCE " , by being vigilant no one with the malicious mindset can get access to your computer system , can never cause harm your personal data . As being vigilant Social Engineering attack and other active as well as passive attack will be reduced at much larger scale .**

3) **Back up your data on regular basis can also reduce any attack as having up to – date data can help in restoring any sensitive files without paying a ransom amount of money to the attackers .**
4) **Be cautious , think twice before clicking on any links specially if they are from unknown or suspicious sources . The Phishing attack is used by the attacker by creating a sense of emergency to the targeted organization or a random person.**
5) **Use strong and unique passwords for all your accounts , and its advisable not to use easy to get passwords . As commonly used passwords could be cracked by the attacker (the person with the malicious mindset) using dictionary attack or brute force attack or reverse brute force attack or keylogger attack .**

Some Common Attacks used by the attacker's :

1) **Dictionary Attack :**

   Attacker uses all the possible , easy to crack passwords to get access to an unauthorized computer system , to steal sensitive data .
2) **Brute force Attack :**

   Is the type of cyberattack where the attacker tries to get access to an unauthorized system by trying all possible combination of passwords , usernames . Time consuming attack but if the access is gained it can cause harm to the data .

3) Social Engineering Attack :

Is the type of cyberattack that tricks user's mind to reveal the username and all the important credentials , to get access to the targeted system .

Life without Machine Learning :

If there would be no Branch of Artificial intelligence or the AI tool as well then no facial recognition would be there , to look up for any words we just directly rush towards chatgpt or google for any lookups but what if there would be no such thing like search engine then searching or looking up for words would be next to impossible as it would be a tedious job to look for each particular word in the dictionary or in search materials .

Life with Machine Learning :

With the fast growing technology , Artificial intelligence along with all its branch has spread its roots in todays time . Artificial intelligence provide :

- **Gesture control gaming**
- **Products suggestion in any e commercial website**
- **Price of products which basically varies on increase in demand**
- **Customer segmentation**
- **Suggest drop – pickup location in any transportation service (uber)**

Machine Learning :

It is the branch of artificial intelligence that basically focuses on the development of algorithms, and provides the ability to automatically learn and improves from experience without being explicitly programmed. Computers learn from the data set provided, analyze the data, identify all the patterns and based upon that take the following decisions.

Steps of machine learning :

- **Analyze the data ->**

  The data set provided to the system, is firstly analyzed ,and then based upon the analyzed result observations are made.

- **Find Patterns ->**

  After analyzing the data , system observes pattern based upon the color , shape ,size . And then it segregates the input data based upon the following parameters .

- **Prediction / Decision ->**

  After recognizing all the patterns and performing segregation on the input data set , it is used to predict things as well tries to make decision based upon the analyzed results.

- **Learn from the feedback** ->

  With the each prediction as well as the decision made the system , it learns from it and remembers it for later use .

  If again the same input is provided to the system then it won't go through all the input data set or analyze the data , it directly recalls the stored result and gives the desirable output .

Several types of machine learning are :

1) **Supervised Learning :**

   In Supervised Learning, the algorithm is basically trained on the mentioned data , where the output which has to be obtained is known.

   The model is trained by passing / giving the known data along with the known response and model remembers the desired response and if later on the same input data will be given to the model it will not perform any algorithm it will basically give the stored result of that particular input , in such cases the model learns from experience.

2) **Unsupervised Learning :**

   In unsupervised Learning , the algorithm is trained on the unlabeled data . The algorithm , explores the data , discover the patterns and give the desirable output . On receiving the input data set , the model basically categorize or

separates the data set based upon their color, size , shape and then clustering of data takes place .

3) **Reinforcement Learning :**

   In this learning, an agent learns to interact with the environment to maximize the rewards and minimize the penalties. The model learns from its mistakes and then later on gives the desirable output .

   Eg : If the apple as the data set is given to the model / computer system then based upon the input received the model will categorize it into mango but if the response received is not as expected then it will learn from the feedback and then will store the result and will minimize the error's occurred.

Supervised Learning vs Unsupervised Learning :

1) **Supervised Learning operates on the labelled data , input as well output is provided to the model , while Unsupervised Learning operates on the non – labeled data .**
2) **Supervised Learning gets direct feedback, unsupervised learning gets no feedback .**

3) **Supervised Learning is used to predict output while Unsupervised Learning used to find hidden structure data .**
4) **Supervised Learning , learns from environment and then based upon that gives the desirable output while in unsupervised Learning it is used to separate the input data set based upon the patterns , size as well as shape .**

Deep learning

It is a subfield of machine learning algorithm , inspired by the structure and the function of the brain . Work similar to the neuron in the human body .

Deep learning is a subset of machine learning , that uses all the complex algorithms and deep neural network to train a model.

Deep learning is preferred much as compared to machine learning as deep learning handles large amount of unstructured data , above all uses neural network so it operates much faster than machine learning model .

Need for deep learning model :

a) **It can process a large amount of data but can work with enormous amount of structured as well as unstructured data .**
b) **Perform complex algorithms in less time as compared to machine learning model.**
c) **To achieve best accuracy with less error , on large amount of data for that purpose we use Deep learning model .**

Application of Deep Learning model :

a) **It helps to detect cancerous tumors in the human body. With the growing technology, and spreading roots of Deep learning as well, we can easily check for cancerous elements in the human body with the help of deep learning models.**

b) **It helps to train Robots so that they can perform the task that are being performed by the human's but in very efficient manner .**

c) **Autonomous Driving – self driving cars are slowly coming into picture as these cars are reliable as well as avoid any accidents . Deep learning model behind this strategy basically works in such sense that video of the surrounding is provided to the model , so that it can sense whether there are cars on the road , can detect traffic signals without human intervention .**

d) **Machine translation – Deep learning model behind this translation accepts / receives the data which is used to convert from one language into another language .**

e) **Music Composition – Deep neural networks can be used to Produce music by making computer learn the composition .**

1.2 Problem Statement

Project on malware detection using Artificial Intelligence and deep Learning .

The main objective behind this project is to develop a machine learning–based solution that can identify any type of malicious software in any computer system.

Traditional malware-detecting techniques were not up to the mark as they could simply categorize them on the basis of patterns and signature, above all they are inefficient to detect all new malicious software

Therefore Machine learning it trains the model / computer system to detect the presence of any malicious software with much more accuracy than the traditional malware detection methods .

Malware detection could not be as whole as sufficient than the subset of malware detection – Deep learning is further used in this project , to reduce the error and increase the efficiency of the model .

1.3 Objectives

The proposed work objectives are as follows :

1) **To build a model that predicts the presence of malware with less error and high accuracy .**
2) **As with rapid growth in technology , people with criminal mindsets also are increasing in number , in order to protect our data from attackers , this model is used to detect any type of malicious software present in the computer system.**
3) **This machine learning based model can analyze large volumes of data , and search for any malicious content which could not be easily detectable by the humans .**

4) **Machine learning model to detect malware , it basically learns from the environment and extract all the relevant information , that could be proved helpful in future. This model is basically error-free as the AI model fails to deliver the wrong information unless and until it is not recognized earlier .**

Chapter 2

Literature Survey / Background

Sno	Title	Year Published	Description	Authors
1)	"A survey of Machine-Learning Technique for Malware Detection " .	2009	This survey paper provides an overview of machine learning techniques used for malware detection along with some feature selection , classification of algorithms and evaluation metrices .	Published by : *Natraj et al.*
2)	" Using Machine Learning Algorithms for detecting all new Malicious Executables "	2006	This survey paper published by the author proposes a machine learning approach to detect all unknown malware executables by training models using various machine learning algorithm	Published by : *Kotler and Maloof*
3)	" Learning to detect and classify Malicious Executables in the wild "	2010	This survey paper presents a hybrid approach combining static and dynamic analysis to detect and classify malicious executables using machine Learning Techniques.	Published by : *Ye and Yi*

4)	**“ Deep- Droid : A Deep Learning Based Systems for Android Malware Detection “**	**2017**	**The authors introduce DeepDroid , a deep learning-based system that analyzes Android apps to detect malware .**	**Published by :** ***Tian et al.***
5)	**“Adversarial Examples for Malware Detection “**	**2017**	**This research was basically based, To explore adversial attacks on machine learning model for malware detection, used to generate malicious samples that can evade detection.**	**Published by :** ***Groose et al.***
6)	**“End to End Deep Learning Models for Malware Detection “**	**2017**	**The authors proposed an end-to-end deep learning framework for malware detection which differentiates the malicious software as benign or malicious .**	**Published by :** ***Saxe and berlin.***
7)	**“Malware detection Based on Dynamic Analysis of API Calls “**	**2012**	**The authors proposes a dynamic analysis approach for the malware detection which was based on monitoring the sequence of API calls .**	**Published by :** ***Canfora et al .***
8)	**“ Eureka : A framework for Enabling static malware Analysis“**	**2006**	**This paper proposes the Eureka Framework , which is a combination of static analysis as well as machine learning technique to facilitate the detection of malware .**	**Published by :** ***Rajab et al .***

Chapter 3

Methodology

Dataset :

- Data Collection:

The Source of the Data set used for the classification of legitimate and non-legitimate software using the Deep Learning model is " C:/Users/bhuta/Downloads/malware_data_2.csv " while the source of the data set used to classify the same using the Machine Learning model is " C:/Users/bhuta/Downloads/malware_data_2.csv ". All these sources already exist on the internet, they are just extracted and chosen in the way as the data present in the dataset was well aligned, which does not lead us into a problematic situation like searching for a particular word that exists in the data -set.

- Malware Samples :

Whereas the total number of data samples on which the process of categorization has occurre d using the deep learning model is : Total number of data samples in the data set that is evalu ated through Machine learning techniques is : Input sets:
(58596, 56)
Target sets:
(58596,)

- Data – Preprocessing :

After collecting the data – set pre- processing part was taken into consideration , Several Libraries were imported ,

- Import NumPy as np -> was used to implement NumPy arrays
- Import pandas as pd -> was used to create data set (data frames)
- Import tensorflow -> to split the data (legitimate or non legitimate data) into training and testing data .
- Import sklearn.feature extraction -> was used to convert mail data to numeric values so that machine learning model could identify feature vector (numerical value).
- Import Logistic regression -> was used to classify which type of software was it ? either legitimate or non legitimate .

- Import accuracy -> to know about the accuracy of our model which has been trained for any future recognition .

About Libraries which has been used ->

- from sklearn.ensemble import ExtraTreesClassifier :

It is the part of sklearn , which help us to provide the implementation of extra Trees classifier algorithm .

Commonly used for classification task including malware detection .

Implemented using decision tree.

- from sklearn.model_selection import train_test_split :

used to split the dataset , into testing and training subset , it splits the data randomly to measure much more accuracy .

- from sklearn.model_selection import cross_val_score :

Library provides cross validation , the technique of machine learning for evaluating machine learning models. Using this method data is splitted into multiple folds and hence the model is trained .

- from sklearn.feature_selection import SelectFromModel :

  It is used for the feature selection , based upon the features in machine learning model , helps in improving model efficiency .

- from sklearn.linear_model import LogisticRegression :

  It provides an implementation of logistic Regression which is a linear classification algorithm used for binary classification .
  Often used for binary classification problem's in malware detection .

- from sklearn.ensemble import RandomForestClassifier :

This library is another machine learning algorithm that builds multiple decision trees and combine their output to make the final predictions .

It is an extension of decision tree algorithm generally used for classification purpose .

- import pandas as pd :

It is a powerful library for data manipulation

provide data structures like data frame , series , which help in efficient handling .

- Various Types Of algorithms Used in this project :

1) Cross Validation Classifier :

From sklearn.model_selection import cross_val_score

It is basically a technique commonly used in machine learning algorithms to check for the accuracy as well the performance of the model . Estimates how well the model will perform on the unknown data – set .
First and the foremost step is to divide the data set into two or more subsets , (testing and training set) using train test split , used to split the data .
The training set , train the classifier model while the testing set is used to evaluate the model based upon its performance .
Extra feature of test_ size = 0.2 determines that , 20% data resides on the test set while the rest 80% resides on the training set .
Further k – fold Cross Validation is the common approach where the dataset is divided into k equal – sized folds . The model is evaluated number of times , the number equal to k folds used in the model , that allows for a throughout evaluation of the model's performance across the entire data set .
With each passing fold , the classifier model is trained to minimize the training error and maximize the accuracy , hence K folds are used in this project for better accuracy.
After training the model based upon the data set , model performance is evaluated on the testing set by measuring accuracy of that particular model .

Cross validation technique basically provides an overview , that how the trained model will perform on the unseen data – set .

2) Gradient Boosting Classifier :

From sklearn.ensemble import gradientBoostingClassifier

It is technique of machine learning , that involves ensemble of weak learners by minimizing the loss function of the whole model , with the passing of each iteration a new weak learner is added and taken into consideration which results in minimizing the errors made by the previous learners . Decision tree is primarily used in training the weak learner based upon the weighted training data , then model computes the error with each training of the weak learner , to minimize the errors caused by the weak learners . Once all the boosting process comes into completion all the weak learners are combined , Gradient Boosting Classifier uses the weighted sum of the weak learner's prediction to make the final classification . This technique produces highly accurate results , as it combines all the weak learners all together and learn from their predictions on each step of training data _Set .

While any type of data can be handled by this technique of machine learning , data such as numerical data or categorical data .

3) KNN (K Nearest neighbor) :

It is a type of machine algorithm used for both classification as well as regression purpose . The KNN algorithm , measures the similarity between the data points by relying on the distance metric, during the initial phase of this algorithm it simply stores the labelled training dataset in memory , without performing any explicit model training. When the data set needs to be classified or predicted , KNN algorithm it computes the distance between new data and the training data set using the chosen distance metric .
Hence the distances are selected based upon the computed distances , and with the selected distance the classification process comes into consideration, the algorithm assigns the class label which is most frequent among the K nearest neighbors to the new data point.
Moving on with the Regression tasks, the algorithm predicts the numerical value by taking the average or weighted average of the target values of the k nearest neighbors.
Thus algorithm is not used widely as with the large data set , the algorithm needs to compute distances for each instance, which can become time – consuming .

4) Random forest :

It is way all powerful machine learning algorithm , used to combine multiple decision 3wzztree to create a robust and accurate machine learning model , each decision tree is trained on a different subset of the training data , using random selection technique .
Random forest it operates on the technique named as bootstrap aggregating , it randomly replaces the training data with bootstrap samples, then each decision tress is trained on these bootstrap samples , which makes it more resilient to noise .
Once all the decision tress are trained , Random Forest combines all their predictions to make the final prediction, the class with the highest count is selected as the only final prediction .
Random forest can produce high accuracy , with much more accurate results , by combining multiple decision tress all together , as well as can handle complex relationships .

5) Decision Tree :

It a popular machine learning algorithm , used for classification and regression tasks. A tree like model is constructed which is used to provide a simple way of predictions , this decision tree looks like a hierarchical structure composed of nodes as well as edges , the tree is divided into internal nodes and those internal nodes represent decisions on the input features while leaf node represent the final outcome which is in all way predicted . The data splitted by this algorithm is totally based upon the threshold , to split the data at each internal node . Once the data is splitted it enters into the training phase, decision tree algorithm recursively partitions the training data based upon the criteria , to minimize the error at each splitting the threshold at each internal node is determined . The process stops or come to an hault state only when on splitting it reaches the maximum tree depth , minimum number of samples per leaf . The final step of prediction takes place with a trained decision tree , a new data traverse the tree from the root node to a leaf node.
Decision trees represents a proper sequence of iteration , making it easy to understand it as well .

6) Confusion Matrix :

It is a type of machine learning model , that summarizes the performance of a classification model by comparing the predicted labels with the true labels of the dataset ,

a) True positive (TP) -> The model basically predicts the positive class with much more accuracy .
b) True negative(TN)-> the model predicts the negative class
c) False Positive(FP)-> the model predicted the positive class incorrectly when the true label is negative .
d) False Negative(FN) -> model predicted the negative class incorrectly when the true label is positive .

this matrix insights into the model performance , it Is the essential tool for evaluating and comparing classification models in machine learning.

7) Linear Regression :

It is widely used machine learning algorithm , for solving regression problems . The main objective of this algorithm is to search for the fitting line that minimizes the difference between the predicted values as well as the actual values .
It involves a single independent , variable and a linear relationship between the independent and dependent variables . In the training phase , it estimates the values of coefficients , done by minimizing the sum of squared differences between the predicted values and the actual values . After training the model , it is evaluated using various techniques , squared error is one of them .

This technique of machine learning provides a simple model for understanding the relationship between variables and making predictions based on the observed data .

1) Cross Validation Classifier :

It is a technique which is used to train and evaluate our model , on a small portion of our data set which is broken down into testing data and training data , before portioning the data set and evaluating it on the new portions .
Training data is basically used by our model to learn where as the testing data is used by our model to make predictions on the unseen data – set .
On initial we use a different portion of the training data and the remaining as testing data , with each folds portion for testing and training get changed to measure the accurate results on the data set and then later on the average of all the portions is taken and is marked with the actual performance.
Types of cross validation are :

- Leave one out cross validation :
  In this type of cross validation , entire data set is used for training but the one last data point or singular data – point is left for testing data .
  On the repeat for n times different portion for the testing data and the training data will be selected to get much more accuracy .
  Final measure is calculated by taking average of all the portions divided by n times the portion has been chosen .
- K Fold Cross Validation :

  In this type of cross validation , data is divided into k number of different sections. The number of sections is selected based upon the size of the dataset (k) .

(Different Training data as well as different testing data is selected based upon the different number of rounds.)

- Stratifies K- Fold Cross Validation :

  In this type of cross validation ,the data split so that each portion has the same percentage of all different classes that exist in dataset , it is even beneficial for the prediction of the minority classes . Same percentage of data classes are maintained , classes can be chosen based upon the number of sections for division .

2) Gradient Boosting Classifier :

In order to solve complex computer problems , we require this algorithm , it is used to compute the accuracy of the model based upon the strong learner , as weak learner are not sufficient in predicting the accuracy of the model . It could lead to wrong predictions if only weak learner are lead to make classification . All the predictions / outcome of the weak leaner combine to form the strong learner , the predictions of the weak learner , are made through the majority rule or the weighted average .

Boosting ->

It is a process through which weak leaners are combined to form the strong learner in order to measure the accuracy of the model .

Ensemble Learning ->

It is used to enhance the performance of the machine – learning model by several weak learner's predictions .

Firstly equal weights are applied to the testing and training data with the completion of the first step , decision stump is drawn ., then it will check for all the false predictions which has been made through each possible iterations and with this higher weightage is assigned to the mis – classified patterns .

Types of Boosting :

a) Adaptive Boosting ->

   In every step , decision stump is drawn and is assigned with the maximum weight / higher weight to the mis matched objects / patterns
   Unless and until all the patterns come under same class .
   This is used for classification as well as regression purpose .

b) Gradient Boosting ->

Here base learners are generated sequentially , in such a way that the present base learners is always more effective and is best in making predictions as well as calculating accuracy for the model .
Maximum weight is not assigned to the mis - matched patterns , in spite of this an additional model is required to regularize the loss function from the previous weak learner .

c) XGBOOST ->

It Is the advanced version of Gradient Boosting method , that is designed to focus on the computational and model efficiency .
The only disadvantage Gradient Boosting classifier has that is slow in execution , whereas XGBOOST it the Extreme Gradient Boosting .

3) KNN (K Nearest neighbor) :

It is simple to implement as well as donot leads to a confusion state. When the value for the variable k is assigned , it will basically takes that particular number of nearest objects and according to this , the majority of the nearest object will be assigned to the new point .

Like for eg : k=3 that means in the plotted data set , a new point has to be determined and based upon its parameters it will be placed in the dataset as the value for the k variable is assigned as 3 then the nearest object of that particular point will be considered and majority of the objects that has been selected will be assigned to the new point and based upon that the accuracy is calculated ,which basically leads in increase accuracy .

4) Linear Regression :

It is a model which is used to find , relationship between one or more independent variables .

The main purpose behind this model is used to find the linear line , so that the model could be easily plotted on the selected line which further leads into the calculation of accuracy for the model . Here the squared of the distance is taken from each point to the line as the distance could be either greater than zero or it could be less than zero . whole data set is not taken into consideration , as it could lead to inaccurate results . hence the random data is chosen and splitted . we find the best fitted line for the model which has to be trained .

5) Confusion Matrix :

here the total number of values in the matrix is equal to the test_ dataset , it is all calculated manually .

calculations are calculated manually , higher the sum of diagonal elements higher will be the accuracy .

6) Decision Tree :

It is more powerful for classification as well as Regression , which is much more similar to random forest classification . Any Classification can be carried out including binary classification as well . in this algorithm each node represents a feature , each link represents a decision and each leaf node represents the outcome .

Is an inverted tree where root node is at the top and they decision is taken on the basis of several conditions and if the outcome leads to affirmation it is further divided into various decision branches otherwise it could lead to the leaf node which consists of the final outcome.

7) Random forest :

It is much similar to the decision tree on the basis of execution the only difference is in decision tree only one tree is generally used but here as the name suggest Random Forest here forest / group of trees are used for the purpose of classification .

With the usage of group of trees it increase efficiency of the model along with the accuracy .

Here final outcome is calculated by taking the average of the multiple forest which further leads to increase in the accuracy .

Code Template

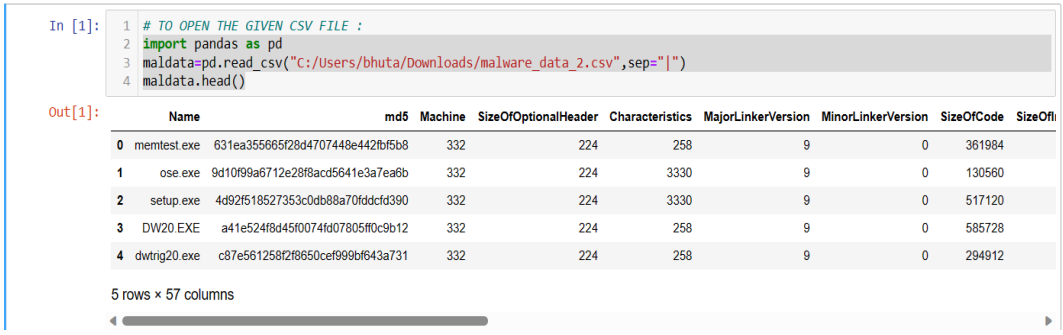
The brief explanation of the code is as follows :

1) Import pandas as pd
   It signifies to import the Pandas library and pd is the alias name of the corresponding import . Alias name is generally used for convenience purpose , and its utlity is throughout the code .
   Maldata = pd.read_csv(“ C:/Users/bhuta/Downloads/malware_data_2.csv”,sep=”|”)
   Note : This file already exist on the internet .
   This loads the csv file into the variable named as maldata , read function is provided by the Pandas Library to read the data from the csv file .
   Maldata.head()

   Maldata refers to the variable that holds the dataframe which has been loaded from the mentioned location .

Significance of . => it is used to access methods of an object.

head() -> it returns the first 5 rows of the Dataframe.



In [1]:
```
# TO OPEN THE GIVEN CSV FILE :
import pandas as pd
maldata=pd.read_csv("C:/Users/bhuta/Downloads/malware_data_2.csv",sep="|")
maldata.head()
```

Out[1]:

	Name	md5	Machine	SizeOfOptionalHeader	Characteristics	MajorLinkerVersion	MinorLinkerVersion	SizeOfCode	SizeOfI
0	memtest.exe	631ea355665f28d4707448e442fbf5b8	332	224	258	9	0	361984	
1	ose.exe	9d10f99a6712e28f8acd5641e3a7ea6b	332	224	3330	9	0	130560	
2	setup.exe	4d92f518527353c0db88a70fddcfd390	332	224	3330	9	0	517120	
3	DW20.EXE	a41e524f8d45f0074fd07805ff0c9b12	332	224	258	9	0	585728	
4	dwtrig20.exe	c87e561258f2f8650cef999bf643a731	332	224	258	9	0	294912	

5 rows × 57 columns

```
In [5]:
#TO PRINT ALL THE COLUMNS :
print(maldata.columns.tolist())
```

```
['Name', 'md5', 'Machine', 'SizeOfOptionalHeader', 'Characteristics', 'MajorLinkerVersion', 'MinorLinkerVersion', 'SizeOfCode',
'SizeOfInitializedData', 'SizeOfUninitializedData', 'AddressOfEntryPoint', 'BaseOfCode', 'BaseOfData', 'ImageBase', 'SectionAli
gnment', 'FileAlignment', 'MajorOperatingSystemVersion', 'MinorOperatingSystemVersion', 'MajorImageVersion', 'MinorImageVersio
n', 'MajorSubsystemVersion', 'MinorSubsystemVersion', 'SizeOfImage', 'SizeOfHeaders', 'CheckSum', 'Subsystem', 'DllCharacterist
ics', 'SizeOfStackReserve', 'SizeOfStackCommit', 'SizeOfHeapReserve', 'SizeOfHeapCommit', 'LoaderFlags', 'NumberOfRvaAndSizes',
'SectionsNb', 'SectionsMeanEntropy', 'SectionsMinEntropy', 'SectionsMaxEntropy', 'SectionsMeanRawsize', 'SectionsMinRawsize',
'SectionMaxRawsize', 'SectionsMeanVirtualsize', 'SectionsMinVirtualsize', 'SectionMaxVirtualsize', 'ImportsNbDLL', 'ImportsNb',
'ImportsNbOrdinal', 'ExportNb', 'ResourcesNb', 'ResourcesMeanEntropy', 'ResourcesMinEntropy', 'ResourcesMaxEntropy', 'Resources
MeanSize', 'ResourcesMinSize', 'ResourcesMaxSize', 'LoadConfigurationSize', 'VersionInformationSize', 'legitimate,,']
```

2) Print(maldata.tolist())
   Maldata Is the variable that contains the data / contents which has been loaded from the mentioned csv file , hence with this statement it prints out the columns present in the variable – maldata .

3)

```
MeanSize', 'ResourcesMinSize', 'ResourcesMaxSize', 'LoadConfigurationSize', 'VersionInformationSize', 'legitimate,,']
```

```
In [6]:
#TO RENAME THE CONTENTS IN THE CSV FILE:
#Preprocessing

X = maldata.drop(["legitimate,,"],  axis=1)
Y = maldata[41323::].drop(["legitimate,,"], axis=1)

In [7]:
#print(X)
```

Droping in machine learning is the important aspect as , machine learning model do not accept NAN values hence to resolve the presence of the not a number values droping is carried .

Drop() : it is used to drop the rows or columns from a dataframe .

X=maldata.drop([“legitimate,,”],axis=1)

It drops the column legitimate , to avoid the NAN values , axis=1 it denotes to drop all the rows for the particular column (legitimate,,) and these dropped values are stored into the variable X .

Y=maldata[41323::].drop[“legitimae,,”],axis=1)

It is used to drop the column named “ legitimate” from the dataframe , where rows before index 41323 are removed , axis=1 it is used to denote to drop the column (legitimate) . After dropping the column it is assigned into the Y variable

4)

```
In [10]: 1 #print(maldata.head(5))
         2 #seaborn matplotlib

In [11]: 1 data_in=maldata.drop(['Name','md5','legitimate,,'],axis=1).values
         2 labels=maldata['legitimate,,'].values
         3 trees= ExtraTreesClassifier().fit(data_in,labels)
         4 select=SelectFromModel(trees,prefit=True)
         5 data_in_new=select.transform(data_in)
         6 print(data_in_new.shape,data_in.shape)

(138047, 14) (138047, 54)
```

This code basically drop , the column named : Name , md5 , legitimate,, from the dataframe and then that modified dataframe is stored in the variable (data_in).

axis=1) .values it specifies to drop the data / values present in the respective columns (Name , md5 , legitimate) .

whereas label variable it holds the data of the legitimate ,, column .

trees= ExtraTreesClassifier().fit(data_in,labels)

It trains the ExtraTreesClassifier model on the data_in and the labels , fit() it is used to fit the model to the training data , which helps it in making predictions based upon the labels which has been provided , this Is then assigned to the variable trees .

select=SelectFromModel(trees,prefit=True)

SelectFromModel is used for feature selection based on the existing model's . trees is the variable that has been passed as the argument that hold the data that has been fitted , prefit=true it signifies that the trees model is already fitted or trained .

data_in_new=select.transform(data_in)

select.transform , it provide feature selection transformation to the data which has been provided , select it selects only the important features from the data_ in and all these important features are stored into the variable termed , data_in_new .

print(data_in_new.shape,data_in.shape)

for checking whether the shapes of the training and testing data matches , shape of data_in_new as well as data_in is verified.

5)

```
13 SectionsMinEntropy 0.02355487083860786
14 ResourcesMinSize 0.01978272446796368
```

In [14]:
```
# split the data (that's why imported train_test_split:)
X_train, X_test, Y_train, Y_test = train_test_split(data_in_new,labels, test_size=0.2)

classif=RandomForestClassifier(n_estimators=50)

classif.fit(X_train,Y_train)
#prred(x,y)
#accuracy_score
#crosss
#confusion matrix
```

Out[14]:
```
RandomForestClassifier(n_estimators=50)
```

X_train, X_test, Y_train, Y_test = train_test_split(data_in_new,labels, test_size=0.2

Train_test_split is a function from sklearn , that splits the data into random test data and train data , this function receives two arguments where one argument is data_in _new and other is labels .

Test_size=0.2 denotes that , 20 % of the data from the data set will be allocated for the testing purpose and rest will be allocated for the training purpose .

X_train, X_test, Y_train, Y_test these are the variables that will hold the result after training and testing of the inputted data .

X_train, this variable will hold the training subset of the inputted data :

X_test this variable will hold the testing subset of the input data

Y_train this variable will hold the labels for the training data

Y_test will hold the labels for the testing data .

classif=RandomForestClassifier(n_estimators=50)

RandomForestClassifier is the class from sklearn module , n_estimators=50 that specifies the number of decision trees to be used in the random forest, while this being the optional statement .

classif.fit(X_train,Y_train)

classif is the variable that holds the instance of RandomForestClassifier model , with this we can fit the model and can perform various operations related to randomforestclassification .

- clf = ExtraTreesClassifier()
  it is used to create instance of the ExtraTreesClassifierclass and assign it to the variable clf .

  scores = cross_val_score(clf, data_in_new, labels, cv=5)

  cross_val_score is a function from sklearn that perform cross validation on the inputted data set , along with the clf classifier it is used to perform cross validation on data_in _new and labels variables , cv=5 that demonstrates 5 cross fold for cross validation. The data will be splitted into 5 folds and can be adjusted according to the requirement .

  print("Cross-Validation Scores:", scores*100)

  variable that stores the result after passing the inputted data into number of k folds is finally stored into scores variable and for calculating the percentage / accuracy of the model through cross_validation will be printed .

  print("Average Score:", np.mean(scores)*100)

  to print the average score / average accuracy of the model through cross_validation it is calculated , by using the formula (np.mean) , scores is the variable that holds the result of inputted data by passing through several k folds .

- print(classif.score(X_train,Y_train)*100)
  by linear regression technique of machine learning the accuracy is calculated by passing X_train , Y_train as the arguments that holds :

X_train, this variable will hold the training subset of the inputted data :

X_test this variable will hold the testing subset of the input data

Y_train this variable will hold the labels for the training data

Y_test will hold the labels for the testing data .

- print("false positives is" ,conf_mat[0][1]/sum(conf_mat[0])*100)
  print("false negatives is" ,conf_mat[1][0]/sum(conf_mat[1])*100)

Conf_mat represents the confusion matrix , which is in the table form that summarizes the performance of classification model by showing the true positives , false positives , false negatives and true negatives .

Through this first line of the code 0th row as well as 1 st column of the conf_mat is accessed which represents the false positives , by diving the result obtained with the sum of the values in the 0th row from the conf_mat , for calculating percentage it is multiplied by 100. From the second line of the code , the value corresponding to the 1st row and 0 th column is accessed from conf_mat and is divided with the sum obtained of the values of 1st row from the conf_mat and later on is multiplied with 100 in order to calculate its percentage .

- grad_boost=GradientBoostingClassifier(n_estimators=50)
  grad_boost.fit(legit_train,mal_train)

GradientBoostingClassifier is a class from sklearn module , that represents the gradient boosting classifier model .
N_estimators=50 is an optional parameter which is passes which counts the number of decision trees used in the gradient boosting process , that particular number can be adjusted based upon the user's choice

grad_boost is the variable , that holds the instance of the GradientBoostingClassifier .

legit_train holds the : this variable will hold the training subset of the inputted data :

mal _train : this variable will hold the labels for the training data

.fit is used to train the model using inputted data .

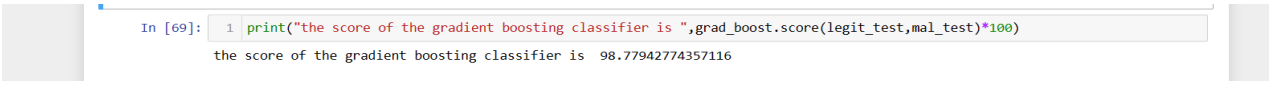
Chapter 4

Result and Discussion

In this project, various machine learning algorithms have been implemented to identify and mitigate malicious software threats, in order to increase the accuracy of the AI model and to lessen down the loss we have implemented the project using deep learning (double layer perceptron) as well .

Result obtained after applying different machine Learning Algorithms are :

1) Using Gradient Boosting Classifier :



Accuracy : 98.77 %

2) Using K Nearest Neighbour (KNN) :

Accuracy : 41.60 %

3) Using Random Forest Algorithm :

```
In [3]: 1 accuracy=(8439/27610)*100
        2 print("Accuracy using Random Forest : " , accuracy)

        Accuracy using Random Forest :  30.56501267656646
```

Accuracy : 30.56 %

4) Using decision Tree Algorithm :

```
In [4]: 1 accuracy=(8381/27609)*100
        2 print(" Accuracy using Decision Tree : " , accuracy)

        Accuracy using Decision Tree :  30.356043319207505
```

Accuracy : 30.35 %

5) Using Confusion – matrix :

```
In [5]: 1 accuracy=(176/27612)*100
        2 print(" Accuracy using Confusion _ Matrix " , accuracy)

        Accuracy using Confusion _ Matrix  0.6374040272345357
```

6) Using Cross – Validation – Score :

```
Cross-Validation Scores: [98.89532778 99.11264035 96.14618422 98.018762   99.42047883]
Average Score: 98.31867863400973
```

Accuracy : 98. 31 %

7) Using Linear Regression Algorithm :

```
Out[40]: 8381

In [41]: accuracy=(8381/27609)*100
         print(accuracy)

         30.356043319207505
```

Accuracy : 30.35 %

Chapter 5

Conclusion and Future work

Malware detection using machine learning algorithms has significantly advanced the identification as well as mitigation of malicious software. Machine learning models offer a wide range of algorithms that can be utilized to train models with different approaches. This allows us to obtain targeted and accurate results, facilitating the detection of various types of malware.

With the growing demand for Machine Learning as well as Deep learning , it can never go out of date .

a) Development of robust machine can be beneficial in the field of defense , training.

b) With the usage of deep learning architectures , such as CNN and RNN and transfer model it can provide improved Feature extraction and detection accuracy .

c) Fusion of machine learning techniques as well as deep learning techniques can yield better results in future.

d) As with increase in number of sophisticated hackers , malware too grow in number , hence there is a need for adaptive model that can lean and update from the real word scenario , hence this could lead in detection of new malware software as well.

e) With the proliferation of INTERNET OF THINGS devices and platform can too detect the presence of any malicious software , future work should focus on working upon things that could be really beneficial for the people as it will avoid data breaches along with privacy breaches along with it will safeguard people's information .

References :

https://www.ijraset.com/research-paper/malware-detection-using-machine-learning

Malware Detection with ML (openai.com)

https://www.researchgate.net/

https://www.sciencedirect.com/