International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

MALWARE DETECTION USING ML

Aniket Gupta, Akashdeep Arora, Yukta Anand

[1] *CSE & ABES ENGINEERING COLLEGE*
[2] *CSE & ABES ENGINEERING COLLEGE*
[3] *CSE & ABES ENGINEERING COLLEGE*

Abstract - Research shows that malware has been growing aggressively, and so is the danger of harm caused by them, so different anti-malware companies have been proposing elucidation to defend attacks from these malware, and we propose a machine learning based approach in order to overcome the same issue, thus we propose a framework in which we use an approach that tries out different classification algorithms before deciding which one to use for prediction by comparing their results. The best algorithm is then used to produce a predictive model which helps to classify the files as Malicious or Legitimate.

After having successfully tested the approach on a small dataset to classify the malware and legitimate files we propose the process must be scaled up in order to get more accuracy.

1.INTRODUCTION

We as a whole understand the significance of malware around us and it has developed quickly in late decades, and alongside the development of the web, there has additionally been the development of programmers and psychological militants those having a dedication to doing violations are making malware. Likewise, with countless instruments and data accessible online there has been a noteworthy lessening in the measure of ability required to make malware. Any program or document that is hurtful to PC clients is named as malware, with the quick development of Internet malware that has gotten to a greater degree of danger to the life of clients. The customary strategies for malware recognition are not fit enough of perceiving current malware which can influence the PCs, as the conventional methodologies utilized straightforward physically created pre-execution rules to identify malware which was sufficient before in light of the fact that the number of malware dangers was close to nothing, however, with the fast utilization of web and a perceptible development in types and impacts of Malwares, the physically made discovery rules were not, at this point pragmatic - and new, propelled assurance innovations were required.

As indicated by the conversations above, unmistakably malware assurance is a significant assignment as there can be an immense misfortune from one single assault along these lines hostile to malware organizations that have been working rigorously in the field of advancement of new and proficient procedures for malware discovery.

When all is said in done, more often than not we can examine that new malware that emerges all the time are only a little changed adjustment of more established malware utilizing some smart procedures. Thus, this ought to be reasonable at this point strategies dependent on the conduct of malware are an extraordinary decision so as to isolate or identify malware.

2. Body of Paper

In this project we will follow the given steps:-

Step1:- Firstly we train our model after data collection and interpret which model will be used for classification. In this project we will be considering following data models:-

- **Decision Tree**
- **Random Forest**
- **Adaboost**
- **Gradient Boosting**
- **Linear Regression**.

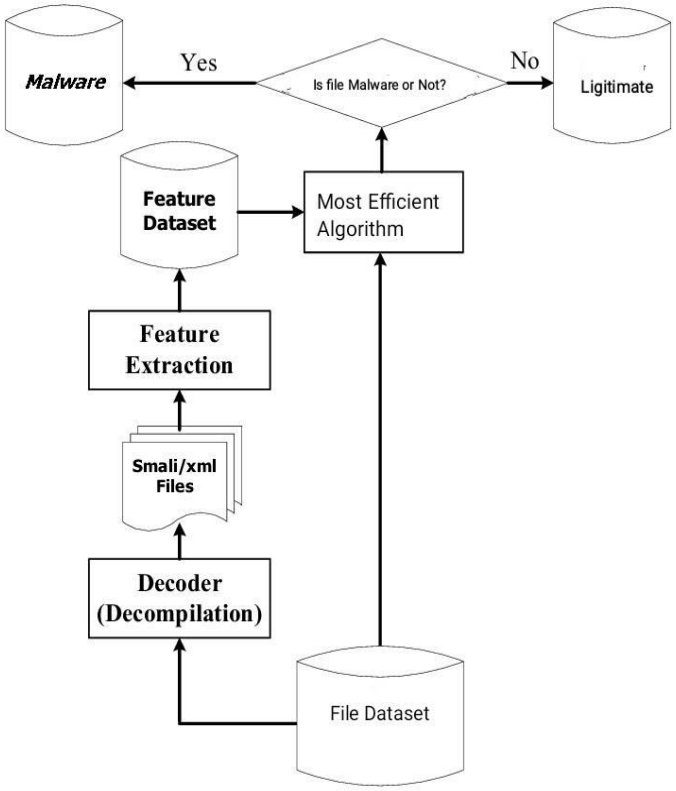
Step2:- After this we select our file which needs to be tested and then it gives the result that whether it contain malicious content or not.

Advanced or Modern techniques for Malware Detection:
A Machine Learning approach for classifying a file as Malicious or Legitimate. In this approach we use a large dataset of pre-classified malicious and legitimate files that help to train and develop our predictive model which then helps to classify unknown files to Malicious or Legitimate, based on training regarding the behavior of malware files in the host computers.
This works in two phases:

B. Training Phase:
The training phase comprises of training the machine learning model using a large dataset of various files which are already classified as legitimate or Malicious by a trusted source, the model is trained using this dataset, which helps to generate a predictive model which can work with unknown files and help to classify them as Malicious or Legitimate.

C. Prediction / Protection Phase
In this phase, the trained model then further which is now named as a predictive model helps to classify unknown files as Malicious or Legitimate based on their behavior in the host computer. This is an advanced technique that can replace the traditional techniques to identify malware files which are more accurate and reliable.
The efficiency of this model is increased by prompting the user with weather the classification was correct or not because in some cases the user is smart enough to trust some sources

which are sometimes classified as malware by the detection techniques and even by advanced ones as there is always a scope of improvement, and then log of the user responses is created is then further used to train the model in order to improve its performance.



PROPOSED METHOD

The proposed methodology or approach tries out 6 different classification algorithms before deciding which one to use for prediction by comparing their results. Different Machine Learning models tried are Linear Regression, RandomForest, DecisionTree, Adaboost, Gaussian, Gradient Boosting.

Different Classification Algorithms:

1.Linear Regression:

It is a direct way to deal with planning the connection between a scalar reaction i.e subordinate variable) and at least one logical factors for example free factor

.2.RandomForest:

Random Forest Classifier is a group calculation. In the following a couple of posts we will investigate such calculations. Ensembled calculations are those which consolidate more than one calculation of the equivalent or distinctive kind for characterizing objects. For instance, running forecasts over Naive Bayes, SVM and Decision Tree and afterward taking a decision in favor of definite thought of class for the test object.

3.DecisionTree:

A tree has numerous similarities, in actuality, and turns out that it has impacted a wide zone of AI, shielding both characterization and relapse. In choice investigation, a choice tree can be utilized to outwardly and expressly speak to choices and dynamic. As the name goes, it utilizes a tree-like model of choices.

4.Adaboost:

AdaBoost, short for Adaptive Boosting, is an AI meta-calculation detailed by Yoav Freund and Robert Schapire, who won the 2003 Gödel Prize for their work. It very well may be utilized related to numerous different sorts of learning calculations to improve execution.

5.Gaussian:

A Gaussian classifier is a generative approach in the sense that it attempts to model class posterior as well as input class-conditional distribution. Therefore, we can generate new samples in input space with a Gaussian classifier.

6.Gradient Boosting:

Gradient boosting classifiers are a gathering of AI calculations that intertwine numerous feeble learning models together to make a solid prescient model. Choice trees are typically utilized while doing slope boosting.

In order to test the model on an unseen file, it's required to extract the characteristics of the given file, these characteristics are passed on to the predictive model which has been trained based on a large data set of pre-classified malware and legitimate files, this model then classifies the unknown files to legitimate or malware files which helps in protection of computers from malware files.

The efficiency of this model is increased by prompting the user with weather the classification was correct or not because in some cases the user is smart enough to trust some sources which are sometimes classified as malware by the detection techniques and even by advanced ones as there is always a scope of improvement, and then log of the user responses is created is then further used to train the model in order to improve its performance.

We use the given six classification techniques in order to train the classifier and produce a predictive model based on one classifier which produces the best result, which is capable enough classifying the unknown files into Malware and legitimate based on its features that depend on the behavior of a malware and a legitimate file in a computer.

3. CONCLUSIONS

We might want to reach a conclusion of report by saying that further research is required around there of Malware recognition and characterization since web is contacting an ever increasing number of individuals consistently additionally malware the malware creation is enhancing straightforward information by day and I indicated utilizing

three unique strategies that AI can be of extraordinary assistance in shielding us from that

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B.Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Professor Birendra Kumar, Department of Computer Science & Engineering, ABESEC Ghaziabad for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance has been a constant source of inspiration for us. It is only his cognizant efforts that our endeavours have seen light of the day.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

REFERENCES

1. Santos, Y. K. Penya, J. Devesa, and P. G. Garcia, “N-grams-based file signatures for malware detection,” 2009.

2. K. Rieck, T. Holz, C. Willems, P. D¨ussel, and P. Laskov, “Learning and classification of malware behavior,” in DIMVA ’08: Proceedings of the 5th international conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 108–125.

3. E. Konstantinou, “Metamorphic virus: Analysis and detection,” 2008, Technical Report RHUL-MA-2008-2, Search Security Award M.Sc. thesis, 93 pages.

4. P. K. Chan and R. Lippmann, “Machine learning for computer security,” Journal of Machine Learning Research, vol. 6, pp. 2669–2672, 2006.

5. J. Z. Kolter and M. A. Maloof, “Learning to detect and classify malicious executables in the wild,” Journal of Machine Learning Research, vol. 7, pp. 2721–2744, December 2006, special Issue on Machine Learning in Computer Security.