Comparative Performance Evaluation of Random Forest on Web-based Attacks

Oluwaseye Abayomi ADEYEMI

LCU/PG/001144

Being a MSc Thesis Submitted to the Department of Computer Science, Faculty of Natural and Applied Science, Lead City University, Ibadan, Oyo State, Nigeria

In Partial Fulfillment of the Requirements for the Award of Master of Science Degree (MSc) in Computer Science

2023

Certification

This is to certify that Oluwaseye Abayomi ADEYEMI with Matriculation Number LCU/PG/001144 carried out this research work title 'Comparative Performance Evaluation of Random Forest on Web-based Attacks' in the Department of Computer Science, Faculty of Natural and Applied Science, Lead City University, Ibadan, Oyo State, for the award of Master Degree (MSc) in Computer Science and that this has not been previously submitted.

……. ……..

Dr. Ajani Azeez WAHEED **Date**

Supervisor

……. ……

Dr. Wilson SAKPERE **Date**

Head of Department

Dedication

This project work is dedicated to Almighty God, the author and finisher of my faith who in his infinity mercy has kept me thus far.

Acknowledgement

First and foremost, I want to express my gratitude to Almighty God for the commencement and completion of this Degree and for making this research work a success; all glory belongs to Him. Special thanks goes to my humble and hardworking supervisor, Dr. Ajani Azeez WAHEED, who is also the Postgraduate Coordinator of Computer Science Programme, for his tireless effort, effective guidance, constant encouragement, constructive criticism, and willingness to listen and provide suggestions whenever needed, all of which contributed to the success of this work. I also have to appreciate Head of Department Computer Science, Dr. Wilson SAKPERE, for his guidance, encouragement, and support, toward the success of this work. I am very grateful. My humble appreciation goes to all the academic and non-academic staff of the department for teaching, nurturing, and influencing knowledge that will be of extreme value to me in my future endeavours.

Unforgettably, I would love to thank my entire family, most especially my mother- Mrs. Ogunkemi Oluwafunmilayo, my wife- Mrs. Adeyemi Blessing, my son's- Adeyemi Eriifeoluwa, Adeyemi Ireayomide, for their prayers, love and unflinching support in all aspects of my life, you are wonderful. My appreciation will be incomplete without mentioning my brothers and sisters, colleagues, friends, and classmates (Computer Science M.Sc. Class of 2021/2022), who have contributed in one way or the other toward the success of this work. Thank you, and God bless you all.

'' Though the above-mentioned institutions and persons have assisted in the process of this research work, I alone stand responsible for the errors, if any, found in the work''.

Abstract

As human resources try to break into networks, control systems, and steal information with the help of expanding data communication paths and protocols, cyber intrusions are currently on the rise. The majority of typical online attack methods are thoroughly researched and documented. Countries, corporations, people, and vital infrastructures that depend on information technology for daily operations have suffered financial losses, the loss of personal information, and economic harm as a result of web-based intrusion. However, foreseeing an attack before it happens can aid in its prevention. This research proposes a predictive model for web-based attacks and a performance comparison of random forest with and without feature selection to secure the availability, integrity, and secrecy of networks, computer systems, and their data. The CIC-Bell-IDS2017 dataset, which includes typical and contemporary intrusion attacks, served as the raw data source for the proposed model. A python-based programming environment and interface for Anaconda Navigator, Jupyter Notebook, was used to create the predictive models. Performance evaluation and comparative analysis were conducted, and the results demonstrate that, once big data analytics (feature scaling and feature selection) were applied to the dataset, the models' prediction accuracies improved, creating a potential intrusion detection system. The outcome yielded excellent accuracy and model development times in both cases, with 97% and 98% precision for both sets and model development times of 35 seconds for the raw set and 15 seconds for the reduced set, which is an important factor when deploying machine learning models in a real-time setting. Random Forest is more computationally expensive than Correlation feature Selection-based classifiers, but having higher predictive accuracy, according to a comparison. Both of these methods work well and each has advantages and disadvantages. The use of big data analytics (PySpark) was found to help machine learning models perform better, resulting in better intrusion detection system.

Keywords: Web Based Attacks, Random Forest, Correlation Feature Selection,

Word Count: 300

Tables of Contents

List of Tables

List of Figures

List of Acronyms

Abbreviation	Meaning
A I	Artificial Intelligence
M L	Machine Learning
ISTR	Internet Security Treat Report
SNSes	Social Networking Services
RF	Random Forest
DT	Decision Tree
BOF	Buffer Over Flow
XSS	Cross Site Scripting
RSMT	Robust Software Modeling Tool
CFS	Correlation Feature Selection

List of Appendix