BALTIC JOURNAL OF LAW & POLITICS
A Journal of Vytautas Magnus University
VOLUME 15, NUMBER 4 (2022)
ISSN 2029-0454

Cite: *Baltic Journal of Law & Politics* 15:4 (2022): 341-348
DOI: 10.2478/bjlp-2022-004036

Higher Accuracy of Malicious Websites Prediction using Logistic Regression Algorithm Comparing with Decision Tree Algorithm

Ch. Sai Venkatesh
Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode:602105.

L. Rama Parvathy
Project Guide, Corresponding author, Department of Computer Science and Engineering Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode:602105.

Received: August 8, 2022; reviews: 2; accepted: November 29, 2022

ABSTRACT:

Aim: The fundamental goal of the research study is to work on the accuracy of a prediction of malignant sites utilizing the Logistic Regression (LRA) machine learning algorithm against the Decision Tree Algorithm (DTA) .**Materials and Methods:** The review utilized 20 samples with two groups of algorithms with the G-power worth of 85% percent and the malicious attack information were gathered from different web sources with late findings and threshold 0.05% and confidence interval 97% with mean and standard deviation. To anticipate the vindictive assaults by further developing the Logistic Regression Algorithm has been viewed as 97% of precision, consequently this concentrate needs to find the better exactness for noxious Attack expectation with the Decision Tree Algorithm Algorithm machine learning algorithm. **Result**: This examination concentrated on saw as 85% of precision for sites utilizing the Decision Tree calculation with a critical worth of two tailed tests is 0.001($p<0.05$) with 97% confidence interval. **Conclusion:** This study presumes that the Logistic Regression calculation on Innovative malevolent site Prediction is essentially better compared to the Decision Tree Algorithm.

Keywords

Innovative Malicious Website Prediction, Machine Learning, Logistic Regression Algorithm, Decision Tree Algorithm, Statistical Analysis, Supervised Learning.

INTRODUCTION

Malicious web site threatening is usage of codes within the style of address by attackers to gather personal info and unauthorized access of user info(Vundavalli et al. 2020) Innovative Malicious Website Prediction. It involves aggregation information regarding passwords registered for email, checking account details, and number for either credit or positive identification and small alternative necessary info (Rani et al. 2020). Attackers might trick users to get their info while not making sense. As like hacking, this technique conjointly takes management over user pc within the kind of breaking weapons system employed in pc. (El-Din, Hemdan, and El-Sayed 2021). The malicious links square measure unfold through email that has details about organization, job vacancies, and on-line buying offers and conjointly it's like legitimate websites. that the user can simply

attract far better than what square measures all the items bestowed in lexical (Manjeri et al. 2019). For every year, Innovative Malicious Website Prediction Supervised Learning the rise of malicious content websites is increasing and it's unbeatable. As per the small print gathered from banking society, malicious attacks of zero 47% were raised early once (Wu and Yang 2011). Attackers feel ease to attack unsuspected users and UN agencies aren't awake to it. In an exceedingly following approach the attackers explored their address no doubt (Lavreniuk and Novikov 2020). The popular web content login portal is targeted by attackers to hide users and it seems to be a legitimate website. Once an unknown user visits the link, the script running behind (Raja et al. 2021; Singh and Goyal 2019) extracts information and makes use of it by the attacker (Yan et al. 2020). In figure one the steps dispensed by assaulter to thieving info from the user is clearly pictured (Chiramdasu et al. 2021). A vindictive application has contaminated a PC, there's actually trust in eliminating it to help any further harm. There's an enormous assortment of spyware and malware throwing out instruments accessible for download on the Internet. Before you leap out and begin downloading tasks, be troubled that there are many phony and awful malware trashing programs.
Malicious is the most unsafe criminal Supervised Learning exercise in cyberspace. Since most of the users go browsing to access (M et al. 2021)the services provided by government and monetary establishments, there has been a big increase in Malicious attacks for the past few years. Several researches are going on to forestall malicious attacks by totally different communities around (Raja et al. 2021) the world Malicious attacks will be prevented by detecting the websites and making awareness to users to identify Innovative Malicious Website Prediction. Machine learning algorithms have been one in every of the powerful techniques in detective work malicious websites. During this study, varied strategies of detective work malicious websites are mentioned.(Rayala et al., n.d.) The Web has Statistical Analysis become a platform for supporting a good variety of criminal enterprises like spam-advertised commerce. Supervised Learning These visits are driven by email, internet search results or links from alternative web content, however all need the user to require some action, like clicking, that specifies the specified Uniform Resource surveyor (URL). The Best study of prediction malicious websites (Mondal et al. 2021).

Previously our team has a rich experience in working on various research projects across multiple disciplines (Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020).The drawbacks of the prediction of a malicious website is If the Internet connection fails, this system won't work and Loss of Customers. Loss of Data and all websites related data will be stored in one place. The accuracy percentage of the Logistic Regression Algorithm is 97.11% individually and the average accuracy of the Decision Tree is 85%. There are more relative articles with a precision score from the DATA classifier for advancement of Anti-malevolent to foresee vindictive sites assaults. Hence the point of this study is to expand the precision of Innovative Malicious Website Prediction weakness and further develop the forecast model utilizing the LRA.

MATERIALS AND METHODS

This exploration review was completed at the DBMS Laboratory, Department of Computer Science and Engineering, Saveetha School of Engineering, Chennai. The two Supervised Learning gatherings of order calculations utilized for the review. Group 1 and Group 2 are the Logistic Regression calculations and choice tree separately als their reaches are displayed in the Fig.1. Each example size was anticipated utilizing the G-power apparatus with rendition 3.1.10 and bringing about 20 example sizes with 97% of G-power values and the limit two tailed significant values is set to 0.05 and the confidence interval as 97%.(McGahagan et al. 2021))

The malicious, Anti- noxious dataset that will be credited for the arranged work is gathered from the (Urcuqui n.d.) one among a ton of inescapable on-line networks for data researchers and machine learning . grants them to go looking and acknowledge entirely unexpected datasets that they require; . It conjointly gives an adjustable individual Google co-lab with a free on-line GPU. The dataset used here consists of forty four attributes and contains five options that can be wont to predict the website malicious attacks. The dataset has 11044 rows that consists of knowledge for the symptoms of that area unit associated with malicious Attack and conjointly includes several sites in the dataset that shows in Fig. 1. Nearly 5.1 billion active net users will be there in 2020, a record for that year throughout the planet.

Logistic Regression Algorithm:
Logistic Regression is a supervised learning algorithm. It provides accurate results when new data is given to the trained model. It is a predictive analysis algorithm based on the concept of probability. The sigmoid function is a mathematical function used to map the predicted value to probabilities. The value of Logistic Regression must be between 0 and 1 which can be calculated using the below equation (1).

Value(V)=1/(1+e^-value) (1)

Where, e is base of the natural algorithms

Pseudocode
INPUT:Training dataset()
OUTPUT:Accuracy

1. Read the training dataset into the classifier
2. Calculate cost function, gradient descent
3. Repeat
4. Calculate sigmoid function for each iteration
5. While the condition satisfy
6. Define class

```
define Logistic Regression
        if(condition satisfy)
            return accuracy
    else
        return previous step
    end
```

7. Classifiers predicted accuracy

Decision Tree Algorithm:
Decision tree classifiers are utilized as a commonly known grouping method. A decision tree could be a flowchart-like tree structure anywhere an inside node addresses a feature or attribute, the branch addresses a decision rule, and each leaf node addresses the outcome. The highest level of node in a call tree is perceived in view of the root node. It figures out how to segment upheld the attribute worth. Also, call trees are ideal for coping with nonlinear relationships between attributes and categories. The following pseudocode comes under the Decision Tree Algorithm recipe to use on the middle pictures dataset and moreover works with the tree model. The pseudocode can take the datasets as info and thus the last result of the pseudocode is sent through the parameters Accuracy and the classification.

Pseudocode
INPUT: Training dataset()
OUTPUT: Accuracy

1. Read the training dataset as input
2. Preprocess the dataset and split to train and test
3. Define class

Decision Tree(test attribute)

if(condition satisfy)

return accuracy

else

return previous step

end

4. Classifiers predicted accuracy

In the proposed system the training and testing of the data is made in the Jupyter notebook and having used the SPSS software to predict the graph and also G-power software to calculate and pretest for the algorithm to get better percentage of the algorithm.In this proposed system 50 gb hard disk and 8 gb RAM is used for execution of the algorithm. The framework type utilized was a 64-digit OS, intel i5 and the operating system in windows.

Statistical Analysis:

The investigation done by IBM SPSS adaptation 23 for both proposed and existing calculation cycle was finished with the 20 examples and for every predicted accuracy was noted for analyzing accuracy for breaking down exactness with esteem obtained from the Independent Sample T-test. Independent variable is Time which is there in the dataset for prediction and the dependent variable is the input text for prediction(Brintha, Preethi, and Winowlin Jappes 2021).

RESULTS

In Table 1**,** It was seen that the Logistic Regression algorithm is essentially better than the decision Tree algorithm. In the Logistic Regression algorithm, Dataset saw that the accuracy and performance of Logistic Regression algorithm is better than decision Tree algorithm. In Descriptive statistics the Accuracy and Algorithm values contain the upto 20 values. Standard Deviation of Accuracy of Logistic Regression algorithm 96.08 And accuracy of decision Tree algorithm is 93.46

In Table 2, The group statistics of Algorithms of Both Logistic Regression algorithm and Decision Tree algorithm. Number of Logistic Regression algorithms are 10 and decision Tree Algorithms are 10. Mean of Logistic Regression algorithm value is 96.08 and Decision Tree is 93.46. and Standard deviation of Both the algorithms are 2.47488 and 0.97673. Standard error 0.78263 and 0.30887.

In Table 3 two tailed significance values less than 0.001($p<0.005$) showed that our hypothesis holds good. When contrasted and different calculations, execution of the Logistic Regression proposed classifier accomplished preferable execution over the Decision tree.

In Figure 1, the independent sample test accuracy Equal variance of sig value is 0.034 and the equal variance not assumed of sig value is null. From Figure 1, both the Logistic Regression algorithm and Decision Tree technique the accuracy value of The Logistic Regression algorithm model Accuracy is 96.08 and Decision Tree Algorithm is 93.46.

DISCUSSION

Based on the above it is observed in the Logistic Regression algorithm that 96.08 has better accuracy than Decision tree 93.46 in prediction of malicious websites. There is a statistical 2-tailed significance in exactness for calculations is 0.001($p<0.05$) by independent t-test.

In the Existing system the accuracy for the Logistic Regression algorithm is 96% and the Decision Tree 93% respectively.(Wang et al. 2022) This analysis makes use of machine learning to predict the accuracy of prediction of malicious websites. The accuracy values for Classifier are 96% and 93%. and compared with another model of Prediction,(El-Din, Hemdan, and El-Sayed 2021) malicious websites for Logistic Regression Algorithm and Decision Tree Algorithm 96.08% and 93.46%(Prabakaran et al. 2022). The Factors affecting the algorithm are sample size of the dataset and test size of the dataset Based on the above finding the Existing Algorithm was chosen to improve the accuracy.

The limitations, That the research attributes that the dataset contains are not many to anticipate accuracy(%) for Innovative malevolent sites expectation. The more the independent and dependent variables the more precision will be gotten to the next level. After performing the statistical analysis and independent sample test in the IBM SPSS tool the significance is $p<0.05$. The future, in the event that the dataset contains many attributes, the classifier can work effectively and can further develop the forecast precision. Attributes like profile, source, and verifications can bring about better precision and definite accuracy upsides of Innovative Malicious Website Prediction.

CONCLUSION

The methodology of ordering the malevolent site expectation physically requires more information on the domain. In this research, It discussed the problem of classifying Innovative Malicious Website Prediction articles using machine learning models. The outcome of the Logistic regression algorithm 96% has better accuracy than Decision Tree 93% in detecting malicious websites. It would be feasible to work on a Logistic regression algorithm than Decision Tree to detect malicious websites Attacks.

DECLARATIONS

Conflict of interests

No irreconcilable situation in this original copy

Author Contribution

Writer SV was engaged with information assortment, information investigation, original copy composing. Creator LRP was associated with conceptualization, information approval and critical review of manuscript.

.

Acknowledgements

The creators might want to thank the board, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences (Formerly known as Saveetha University) for giving the open doors and offices to explore review.

Funding:

The creators thank the accompanying associations for offering monetary help that empowered us to finish the review.

1.Saveetha University
2.Saveetha Institute of Medical And Technical Sciences
3.Saveetha School of Engineering
4.VRT Techno Solutions Pvt.Ltd

REFERENCES

Benin, S. R., S. Kannan, Renjin J. Bright, and A. Jacob Moses. 2020. "A Review on Mechanical Characterization of Polymer Matrix Composites & Its Effects Reinforced with Various Natural Fibres." *Materials Today: Proceedings* 33 (January): 798–805.

Brintha, N. C., C. Preethi, and J. T. Winowlin Jappes. 2021. "Exploring Malicious Webpages

Using Machine Learning Concept." *2021 2nd International Conference for Emerging Technology (INCET)*. https://doi.org/10.1109/incet51464.2021.9456222.

Chiramdasu, Rupa, Gautam Srivastava, Sweta Bhattacharya, Praveen Kumar Reddy, and Thippa Reddy Gadekallu. 2021. "Malicious URL Detection Using Logistic Regression." *2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS)*. https://doi.org/10.1109/coins51742.2021.9524269.

El-Din, Aml Emad, Ezz El-Din Hemdan, and Ayman El-Sayed. 2021. "Malweb: An Efficient Malicious Websites Detection System Using Machine Learning Algorithms." *2021 International Conference on Electronic Engineering (ICEEM)*. https://doi.org/10.1109/iceem52022.2021.9480648.

Gudipaneni, Ravi Kumar, Mohammad Khursheed Alam, Santosh R. Patil, and Mohmed Isaqali Karobari. 2020. "Measurement of the Maximum Occlusal Bite Force and Its Relation to the Caries Spectrum of First Permanent Molars in Early Permanent Dentition." *The Journal of Clinical Pediatric Dentistry* 44 (6): 423–28.

Lavreniuk, M., and O. Novikov. 2020. "Malicious and Benign Websites Classification Using Machine Learning Methods." *Theoretical and Applied Cybersecurity*. https://doi.org/10.20535/tacs.2664-29132020.1.209434.

Manjeri, Akshay Sushena, R. Kaushik, M. N. V. Ajay, and Priyanka C. Nair. 2019. "A Machine Learning Approach for Detecting Malicious Websites Using URL Features." *2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. https://doi.org/10.1109/iceca.2019.8821879.

McGahagan, John, Darshan Bhansali, Ciro Pinto-Coelho, and Michel Cukier. 2021. "Discovering Features for Detecting Malicious Websites: An Empirical Study." *Computers & Security*. https://doi.org/10.1016/j.cose.2021.102374.

M, Gowtham, M. Gowtham, Kolluru Sri Harsha, Jami Nikhil, Maturi Sai Eswar, and S. R. Ramesh. 2021. "Hardware Trojan Detection Using Supervised Machine Learning." *2021 6th International Conference on Communication and Electronics Systems (ICCES)*. https://doi.org/10.1109/icces51350.2021.9489081.

Mondal, Dipankar Kumar, Bikash Chandra Singh, Haibo Hu, Shivazi Biswas, Zulfikar Alom, and Mohammad Abdul Azim. 2021. "SeizeMaliciousURL: A Novel Learning Approach to Detect Malicious URLs." *Journal of Information Security and Applications*. https://doi.org/10.1016/j.jisa.2021.102967.

Nalini, Devarajan, Jayaraman Selvaraj, and Ganesan Senthil Kumar. 2020. "Herbal Nutraceuticals: Safe and Potent Therapeutics to Battle Tumor Hypoxia." *Journal of Cancer Research and Clinical Oncology* 146 (1): 1–18.

Prabakaran, Senthil, Ramalakshmi Ramar, Irshad Hussain, Balasubramanian Prabhu Kavin, Sultan S. Alshamrani, Ahmed Saeed AlGhamdi, and Abdullah Alshehri. 2022. "Predicting Attack Pattern via Machine Learning by Exploiting Stateful Firewall as Virtual Network Function in an SDN Network." *Sensors* 22 (3). https://doi.org/10.3390/s22030709.

Raja, A. Saleem, A. Saleem Raja, R. Vinodini, and A. Kavitha. 2021. "Lexical Features Based Malicious URL Detection Using Machine Learning Techniques." *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.04.041.

Rani, G. Elizabeth, G. Elizabeth Rani, A. Tirumala Vikas Reddy, V. Keerthi Vardhan, A. Sai Sri Harsha, and M. Sakthimohan. 2020. "Machine Learning Based Cibil Verification System." *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. https://doi.org/10.1109/icssit48917.2020.9214195.

Rayala, Rohit, Rohith Kuppa, Sashank Pasumarthi, and S. R. Karthik. n.d. "Malicious URL Detection Using Logistic Regression." https://doi.org/10.36227/techrxiv.14725539.v1.

Reddy, Poornima, Jogikalmat Krithikadatta, Valarmathi Srinivasan, Sandhya Raghu, and Natanasabapathy Velumurugan. 2020. "Dental Caries Profile and Associated Risk Factors Among Adolescent School Children in an Urban South-Indian City." *Oral Health & Preventive Dentistry* 18 (1): 379–86.

Sathish, T., and S. Karthick. 2020. "Gravity Die Casting Based Analysis of Aluminum Alloy with AC4B Nano-Composite." *Materials Today: Proceedings* 33 (January): 2555–58.

Sathish, T., D. Bala Subramanian, R. Saravanan, and V. Dhinakaran. 2020. "Experimental Investigation of Temperature Variation on Flat Plate Collector by Using Silicon Carbide as a Nanofluid." In *PROCEEDINGS OF INTERNATIONAL CONFERENCE ON RECENT TRENDS IN MECHANICAL AND MATERIALS ENGINEERING: ICRTMME 2019*. AIP Publishing. https://doi.org/10.1063/5.0024965.

Singh, A. K., and Navneet Goyal. 2019. "A Comparison of Machine Learning Attributes for Detecting Malicious Websites." *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*. https://doi.org/10.1109/comsnets.2019.8711133.

Sivasamy, Ramesh, Potu Venugopal, and Rodrigo Espinoza-González. 2020. "Structure, Electronic Structure, Optical and Magnetic Studies of Double Perovskite Gd2MnFeO6 Nanoparticles: First Principle and Experimental Studies." *Materials Today Communications* 25 (December): 101603.

Venu, Harish, and Prabhu Appavu. 2021. "Experimental Studies on the Influence of Zirconium Nanoparticle on Biodiesel–diesel Fuel Blend in CI Engine." *International Journal of Ambient Energy* 42 (14): 1588–94.

Vundavalli, Vara, Farhat Barsha, Mohammad Masum, Hossain Shahriar, and Hisham Haddad. 2020. "Malicious URL Detection Using Supervised Machine Learning Techniques." *13th International Conference on Security of Information and Networks*. https://doi.org/10.1145/3433174.3433592.

Wang, Xusheng, Linlin Zhang, Kai Zhao, Xuhui Ding, and Mingming Yu. 2022. "MFDroid: A Stacking Ensemble Learning Framework for Android Malware Detection." *Sensors* 22 (7). https://doi.org/10.3390/s22072597.

Wu, Min, and Mingsong Yang. 2011. "Privacy Preservation for Detecting Malicious Web Sites from Suspicious URLs." *2011 International Conference on Business Computing and Global Informatization*. https://doi.org/10.1109/bcgin.2011.106.

Yan, Xiaodan, Yang Xu, Baojiang Cui, Shuhan Zhang, Taibiao Guo, and Chaoliang Li. 2020. "Learning URL Embedding for Malicious Website Detection." *IEEE Transactions on Industrial Informatics*. https://doi.org/10.1109/tii.2020.2977886.

TABLES AND FIGURES

Table 1. Accuracy Table (LRA, DTA), the accuracy of the Logistic Regression algorithm is approximately 96.08 and Decision Tree algorithm is approximately 93.46.

Test Size	0.2	0.21	0.22	0.23
Logistic Regression Algorithm	96.63	96.92	96.33	97.11
Decision Tree Algorithm	85.01	93.98	94.33	94.23

Table 2. Group Statistics, that the mean accuracy and standard deviation for Logistic Regression algorithms is 96.08 and 2.47488. Decision Tree algorithm is 93.46 and 0.97673

	LRA,DTA	**N**	**Mean**	**Std. Deviation**	**Std. Mean Error**
Accuracy	LRA	10	96.0850	2.47488	.78263
	DTA	10	93.4630	0.97673	.30887

Table 3.Independent Samples Test, the comparison of accuracy for Innovative malicious website prediction classification using Logistic Regression algorithm and Decision Tree algorithm with significance rate 0.001 and standard error difference 0.

Independent Samples Test							
		Levene's Test for Equality of Variances		t-test for Equality of Means			
		F	Sig.	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
Accuracy	Equal variances assumed	18.455	.001	2.62200	.84137	.85435	4.38965
	Equal variances not assumed			2.62200	.84137	.78425	4.45975

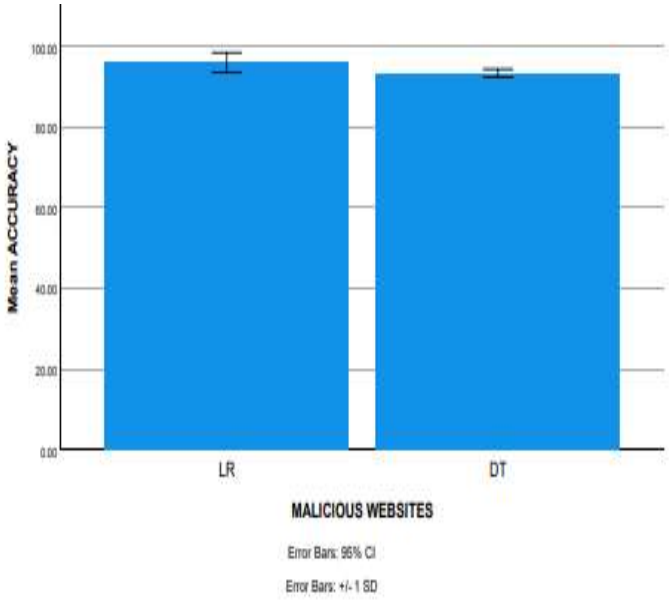

Fig. 1 Simple Bar Mean of Accuracy by LRA, DTA, the bar chart representing the comparison of mean accuracy of the Logistic Regression algorithm is 97% and Decision Tree algorithm is 93%. X-Axis: Logistic Regression algorithm vs Decision Tree algorithm. Y-Axis: Mean accuracy of detection ± SD.