CONCISE COMPARISION OF MACHINE LEARNING MODELS TO DETECT THE PHISHED WEBSITES

MACHAVARAPU.LAKSHMI SAI PRASANNA[1], NARENDRA.ANUSHA[2], KOTHURU.CHRISTIANA HEDEN[3], MUTCHUPALLI.KIRAN [4], Student, Department of CSE, NRI INSTITUTE OF TECHNOLOGY,Vijayawada, A.P., India.

CH.SURYA KIRAN Professor, Department of CSE, NRI INSTITUTE OF TECHNOLOGY, Vijayawada, A.P., India.

ABSTRACT

Information is sensitive and this can beused for any purposes.That's the reason we are focusing much on security measures. Phishing is the malicious attack which we can see in general,while we are working with websites. Because some of the websites can be accessed in an unauthorized manner. We need to define a latest trend in technology to detect which web site is phished ,which are the relevant websites we need to check for the work. Machine learning is the process of understanding the websites which are phished and which are at most health check. We are providing the comparison of Machine lerning classification methods by providing the common dataset and checking the accuracy by using confusion matrix as the performance metrics. Machine Learning algorithms like SVM, RF etc were implemented.

INTRODUCTION

Machine learning is the prominent method which is being used in various sectors. In security and compliance we need to implement a best method and make it understand the need of security in different applications.Phishing is the mechanism used by the intruder to know the users data using a false pattern in webpage URL. Websites are the means of data transfer and there is a large security breech in those areas. These are the means of data leakage and we need to provide a solution using machine learning methods by predicting which is the site effected with phishing method. The methods which are being implemented in this research implementation are comparison methodof different machine learning models like random forest, support vector machine and so on.

- Data preprocessing:

In this, the data set iscollected from UCI data repositoryconsists of features like having'IP_Address','URL_Length','Shortining_Service','having_At_Symbol','double_slash_redirecting', 'Prefix_Suffixetc. This data is represented in .csv format By using Jupyter

Notebook tool which is present in Anaconda Navigator.In that software we can getaccess to the dataset and canperform certain operations byimporting libraries.

- Data splitting:

In this dataset, the data will seperated as Independent and Dependent variables .The independent varaibles can present on or many in the data set.The splitted data is also considered as training data set and the other is considered as testing data set. In this we took training data set as 80% then testing data set as 20%.

- Data Evaluation:

In this the training dataset will undergoes different Machine Learning algorithms namely Logistic Regression, Random Forest Classifier, Decision Tree, KNN, Naïve Bayes, SVM and Neural networks etc and it accuracy is calculated. Out of all algorithms Random Forest Classifier gives more accurate result.

EXISTING SYSTEM

It only describes about the detection of phished websites using any one of the classification techniques.The existing models are not cost effective and require the good configuration of the device to run the model.The previous projects used the different dataset in order to get the highest accuracy which made their project insufficient due to the lack of detailed as well as valid contents in the data.

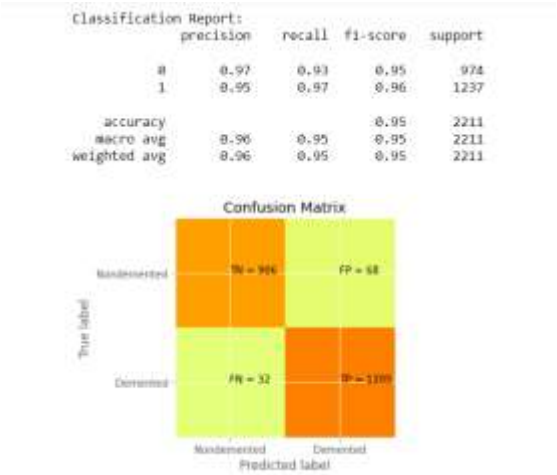
PROPOSED SYSTEM

In this, it mainly focuses on comparing the different Machine Learning Classification models which plays a major role in detection of phished websites.By using this we can identify the most accurate model in detection websites whether they are phished or not.We are taking our dataset which consists of -1,0,1 values.Here -1 indicates phished website state 1 indicates normal website state and 0 indicates the null value, which is helpful in detection of phished websites.The 0 value will be replaced as -1 by using the rename method.The dataset will undergoes the process of finding the accuracy of the classification models.As the Random Forest model is most accurate model, which will be suitable for the best classification model in order to detect the phished websites.

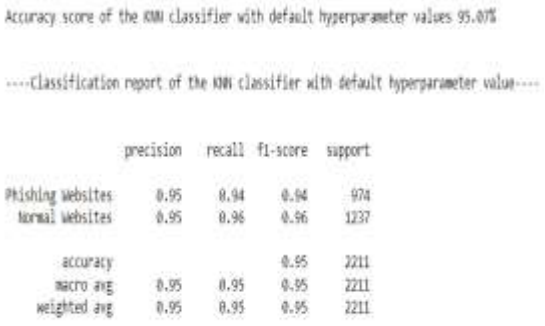
IMPLEMENTATION

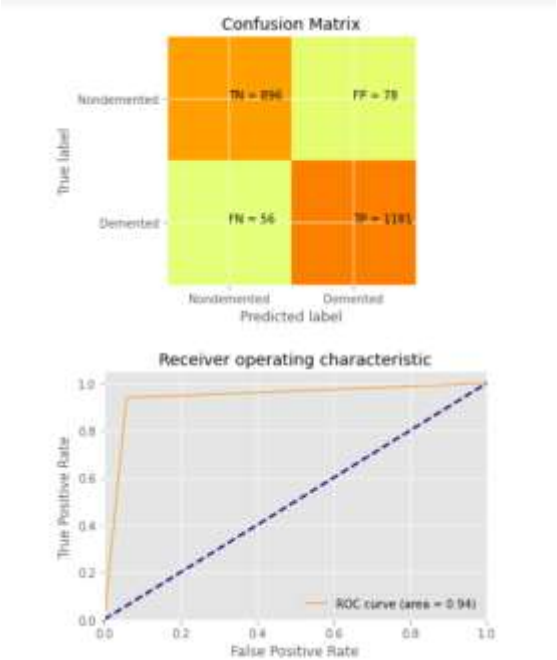
Data preprocessing techniques like collection of the dataset from the UCI data repository.The dataset can accessed with the help of Jupyter Notebook which is situated in the Anaconda Navigator.In this navigator we created the virtual environment for the better result.The libraries such as Numpy ,Pandas are imported in order to certain operations. Matplotlib gives the graphical represention of the data.Confusion matrix and the ROC graph,which plays a key role in accuracy representation of Decision Tree,SVM and Random Forest models.The Keras library is

imported for the implementstion of Neural Networks.In SVM model,it consists of node modules with and without kernal method.The Kernal SVM model decides whether the data is linear or not.Decision Tree also consits of two modules such as with and without entropy.The without entropy module represents the default mode of Decision tree.Random forest consists of two modules named as without and without estimators.The estimators indicates the desired partition of sub classes.By importing the required libraries will find out the accuracies of different classificaiton models.



SAMPLE SCREEN SHOTS





CONCLUSION

This project involves in comparing the accuracy of detecting the phished websites by considering the feature like having _IP_Address,URL_Length,Shortining_Service etc.The six different classification models is applied on the dataset.Out of six different models, Random Forest algorithm gives the highest accurate result.

FUTURE SCOPE FOR FURTHER DEVELOPMENT

On the same dataset,we would like to add more implementation of all classification models as well as implementation of various Neural networks and Natural Language Processing,which will be an advantage of detecting the phished webistes.

REFERENCES

[1] Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning algorithms for phishing website detection. In Proceedings of the antiphishing working groups 2nd annual ecrime researchers summit, eCrime '07, ACM, New York, NY, USA (pp. 60– 69). [2] Abdelhamid N, Ayesh A, Thabtah F (2014) Phishing detection based associative classification data mining. Expert Syst Appl 41:5948- 5959.

[3] Buber, E., Diri, B., & Sahingoz, O. K. (2017). NLP based phishing attack detection from URLs. In International Conference on Intelligent Systems Design and Applications (pp. 608-618).

[4] Babagoli, M., Aghababa, M. P., & Solouk, V. (2018). Heuristic nonlinear regression strategy for detecting phishing websites Soft Computing(pp.1-13).

[5] Buber, E., Diri, B., & Sahingoz, O. K. (2017). Detecting phishing attacks from URL by using NLP techniques. In 2017 International conference on computer science and Engineering (UBMK) (pp. 337– 342). 28.

[6] In A. Abraham, P. K. Muhuri, A. K. Muda, & N. Gandhi (2019), Intelligent systems design and Applications, springer.

[7] Zhang, D., Yan, Z., Jiang, H., & Kim,T. (2014). A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites. Information & Management, 51(7), 845-853.

[8] Feroz, M. N., & Mengel, S. (2015). Phishing URL detection using URL ranking. In 2015 ieee international congress on big data (pp. 635-638). IEEE.

[9] Chen, J., & Guo, C. (2006). Online detection and prevention of phishing attacks. In 2006 First International Conference on Communications and Networking in China (pp. 1-7).

[10] Olivo, C. K., Santin, A. O., & Oliveira, L. S. (2013). Obtaining the threat model for e-mail phishing. Applied soft computing, 13(12), 4841-4848.

[11] Khonji, M., Jones, A., & Iraqi, Y.(2011). A study of feature subset evaluators and feature subset searching .

[12] Afroz, S., & Greenstadt, R. (2011). Phishzoo: Detecting phishing websites by looking at them. In 2011 IEEE fifth international conference on semantic computing (pp. 368- 375).

[13] Blum, A., Wardman, B., Solorio, T., & Warner, G. (2010). Lexical feature based phishing URL detection using online learning. In Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security (pp. 54-60).

[14] Medvet, E., Kirda, E., & Kruegel, C. (2008). Visual similarity-based phishing detection. In Proceedings of the 4th international conference on Security and privacy in communication netowrks(p 1-6).

[15] Miyamoto, D., Hazeyama, H., & Kadobayashi, Y. (2007). A proposal of the AdaBoost-based detection of phishing sites. In Proceedings of the joint workshop on information security.

[16] Prakash, P., Kumar, M., Kompella, R. R., & Gupta, M. (2010, March). Phishnet: predictive blacklisting to detect phishing attacks. In 2010 Proceedings IEEE INFOCOM (pp. 1-5).

[17] Wardman, B., & Warner, G. (2008). Automating phishing website identification through deep MD5 matching. In 2008 eCrime Researchers Summit (pp. 1-7).