A NEW APPROACH FOR THE DETECTION OF FAKE WEBSITE USING CLASSIFICATION METHOD

Mrs. S. DEEPIKA[1]
Assistant Professor, Department of B.Com Business Analytics
PSGR Krishnammal College for Women, Coimbatore, India.
deepika@psgrkcw.ac.in

R. ROSHNI[2]
UG Scholar, Department of Business Analytics
PSGR Krishnammal College for Women, Coimbatore, India.
roshrajan2601@gmail.com

ABSTRACT:
Fake websites are basically created by an individual or an organization with the intention of misleading the viewers and readers. But fake websites are very similar to original websites, where there's a bit difficulty in identifying them. Here, the theme is to identify such websites. The application used in here is Machine Learning, which refers to the study of algorithms using data, which supports its improvement. The algorithm used is Naïve Bayesian classifier algorithm, which eases the prediction by the way of classification. Thus the websites that are fake and original are detected on the basis of black list with the below processes.
KEYWORDS – Naïve Bayesian,age

I. INTRODUCTION

Websites play a very important and essential role in our day-to-day life. Millions of websites are located each day but not every website visited is original. Hence, the main theme of this paper is finding the websites which are original and fake in an accurate number by using AGE OF DOMAIN and GOOGLE INDEX, where AGE OF DOMAIN refers to the year of registration, year of updation, and year of expiry of the website. On the other hand, Google Index refers to the data that are stored as information in the database in the Google search engine.

Python is slowly gaining traction in data science community especially with machine learning (ML). The prime reason for this is the ease of the programming language over others in terms of incorporating ML concepts. When it comes to providing a development environment for Python, there is much notebook software which facilitates programming as well as help with data science in Python. One popular notebook is the Jupyter Notebook —sometimes referred to as ,Jupyter. Some of its features include statistical modeling, data visualization and support for machine learning.[8]

JupyterLab is an open-source web application primarily designed to provide a user interface based on Jupyter Notebook. The installation can be done using simple Python code for Anaconda and pip software package. The packages are available for Windows, Mac and Linux operating systems (OS), and are necessary to run JupyterLab. [10]

As this project is based on classification process, the algorithm used in here is Naïve Bayesian Classifier algorithm. Naïve Bayesian classifier algorithm is one of the best algorithms of Machine Learning. This paper is finely about the analysis of the websites that is done based on the black list to detect the fraud websites using the database provided. And to know the accuracy level of the result found. The tool that is used in here is "JUPYTER". The detailed processes involved in this project are seen below in the paper.
Naïve Bayesian is among the simplest probabilistic classifiers. In the learning process of this classifier with the known structure, class probabilities and conditional probabilities are calculated using training data, and then variables of these probabilities are used to classify new observations [9]. In this paper, we introduce Naïve Bayesian classifier where the probabilities are considered as variables. Numerical experiments are conducted on several real world binary classification datasets. The performances of these models are compared with the Naïve Bayesian classifier. The obtained results demonstrate that the proposed models can significantly improve the performance of the Naive Bayesian classifier. [10]

II.RELATED WORKS

Domain Age plays an important role in email deliverability. The term "domain age" shows how old your domain is. In other words, how long your domain has existed. Domain age is one of the first factors checked by anti-spam filters. The general rule is the older the domain, the more trustworthy it gets, provided it has been used with care and following good practices of email deliverability.[1]

DomainAge is the time of a domain's existence, which is counted from the moment of its first registration. This is one of the important Google ranking factors, which is why there is a common opinion that the older the site is, the higher the search results are expected to be. This is mainly because the long domain age translates into greater trust from both users and Google algorithms. [2]

Machine learning can be broadly classified into supervised and unsupervised learning. By definition, the term supervised means that the "machine" learns with the help of something—typically a labeled training data [3]. To make predictions or decisions, machine learning creates a mathematical model using a sample data which can also be called as training data. These algorithms do not explicitly perform a task or programmed to make predictions. Machine learning provides systems the capability to learn and improve from experience automatically without any explicit program execution because it is an application of artificial intelligence [4].The Naive Bayesian classifier is based on Bayesian's theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods [5].

Google index is another name for the database used by a search engine. Indexes contain the information on all the websites that Google (or any other search engine) was able to find. If a website is not in a search engine's index, users will not be able to find it. It is the source data in Google's index that ultimately determines the variable of different search terms and keywords. Search engines apply their algorithms to the available data, and measure the frequency of different factors under different conditions, which factors are related to one another and so on. The index includes not just the URLs, but all content, including texts, images, and videos and, in principle, everything within the HTML code of the URL.The information garnered from this analysis flows back into Google's algorithm to provide a new assessment of the index data, which attempts to understand which content best meets which user intent. The Google search results, or rankings, are then calculated on the basis of this content assessment [6].

Our approach is to classify phishing websites by incorporating key structural features in phishing websites and employing different machine learning algorithms to our dataset for the classification process. The use of machine learning from a given training set is to learn labels of instances. Our paper provides insights into the effectiveness of using different machine learning algorithms for the purpose of classification of phishing websites [7].

III. METHODOLOGY

METHODS USED:

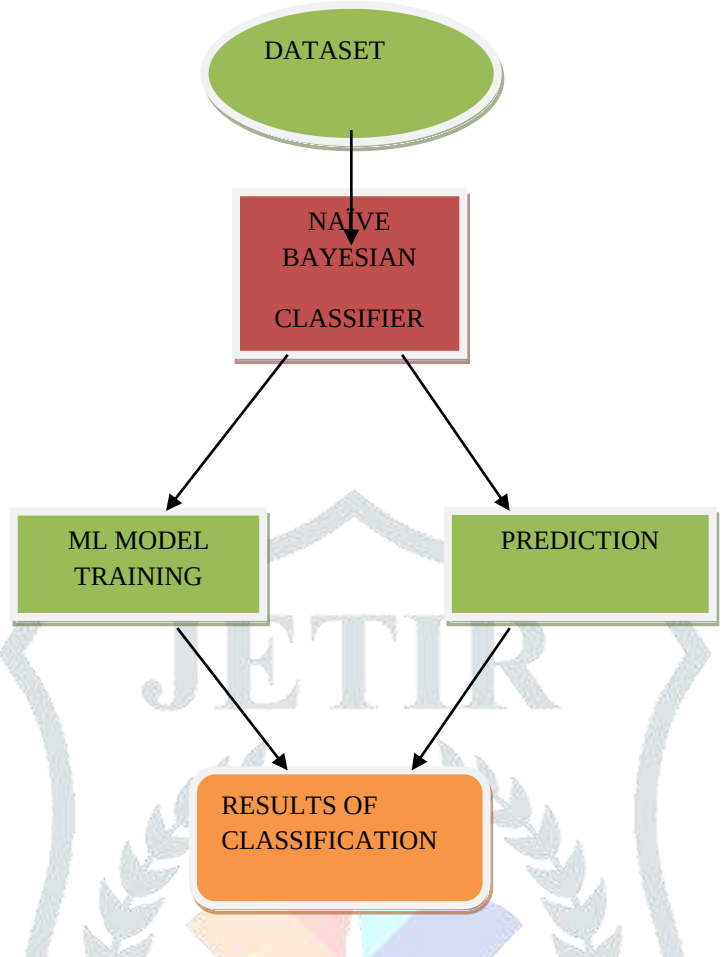
- Firstly, the dataset provided with the extension ".csv" is been imported as into the Jupyter Notebook.
- Then, the packages necessary for the upcoming processes using Naïve Bayesian Classifier algorithm are imported.
- Using the objective of "AGE OF DOMAIN" and "GOOGLE INDEX", the variables under these attributes are plotted using count plot, to segregate the number of fake websites and original websites in the way of bar diagram, where Age of Domain refers to the registered age of the websites and Google Index refers to the websites that are already saved in the Google search engine.

NAÏVE BAYESIAN CLASSIFIER ALGORITHM:
APPLICATION OF NAÏVE BAYESIAN CLASSIFIER ALGORITHM IN THIS PROJECT:

Once the segregation of the variables is done, the values are compared using Naïve Bayesian Classifier algorithm.

- ✓ The attributes that are to be compared, are taken for Machine Learning training and testing model using test and train dataset.
- ✓ Each attribute is separately compared as test and train with prediction variable.
- ✓ Then after applying the Naïve Bayesian Classifier algorithm, the results are found in the type of array.
- ✓ As an accurate result is needed for much clarity, the accuracy of the result is found as in the type of percentage.

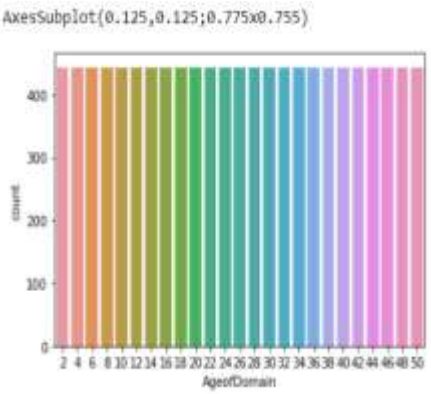
DATA FLOW DIAGRAM



In the above fig 4.1, the Data Flow Diagram of the whole process is shown.

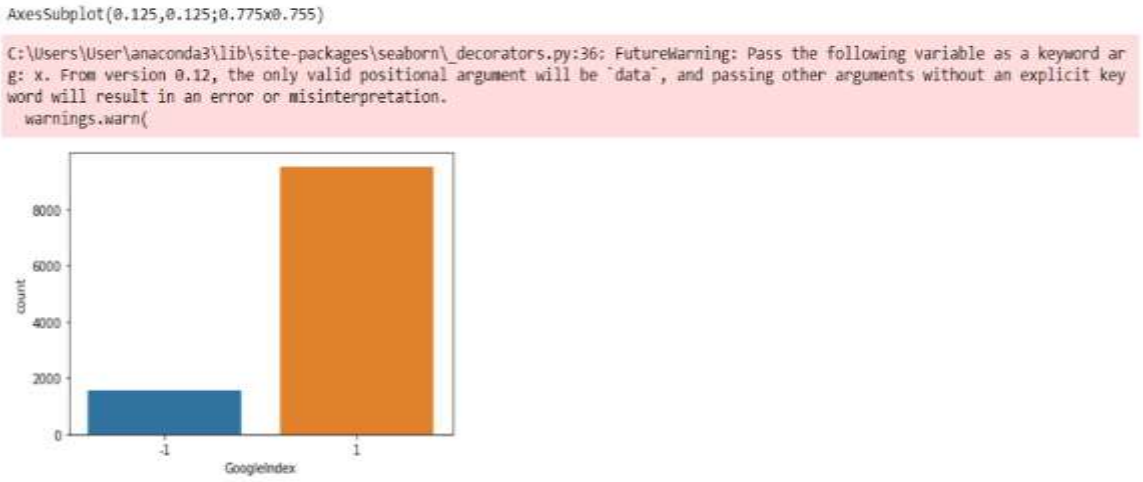
IV. RESULT

Fig4.2



In the above figure 4.2, x-axis depicts the Age of Domain and y-axis depicts the count of the variables in the attribute. Thus the attribute "Age of Domain" has the variables of 2,4,6……..50. Each variable has the count of 400 and above.

Fig 4.3



In the above fig 4.3, x-axis depicts the Age of Domain and y-axis depicts the count of the variables in the attribute. Thus the attribute "Google Index"has the variable of -1 that is less than 2000 which is fake and 1, that is above 8000 which is original.

ACCURACY:

Accuracy = (y_test,y_pred) *100 = 91.027496

In the above fig 4.4, the process of data preprocessing has taken place, the process of training the model and fitting the model as prediction variable and evaluation of model to get the accuracy score is done. Each attribute is compared with Naive Bayes Classifier algorithm with prediction model. Thus, the result gained from these processes is in array value, the accurate value is calculated and the result for that is 91.027496% (as percentage) as an accurate value as it is proposed and done according to the Naïve BayesianClassifier algorithm.

V. CONCLUSION AND FURTHER WORK

CONCLUSION

Thus, the above paper showed the output, whether the given websites are original or fakeon the basis of the objective (i.e) Age of Domain and Google Index, with Naïve Bayesian Classifier algorithm. The accurate value of the websites that are found in large number is also found in the form of percentage. By this paper, the conclusion is that, the original websites are high in number than of the fake websites.

FURTHER WORK

As it is finalized that categorical variables are best for when applying Naïve Bayesian Classifier algorithm, it can be enhanced where not only categorical variables can be performed well, but also the variables that are in numerical form..

REFERENCES:

[1].https://help.woodpecker.co/article/296-domain-age#howold

[2].https://delante.co/definitions/domain-age/#:~:text=Domain%20Age%20is%20the%20time,results%20are%20expected%20to%20be

[3]. Leigh Metcalf, Jonathan M.Sping- " Blacklist Ecosystem Analysis: Spanning Jan 2012 to Jun 2014; ACM workshop on Information Sharing and Collaborative Security; October 2015;pp 13-22; doi.org/10.1145/2808128.2808129.

[4]. Ms. Sophiya. Shikalgar, Dr. S. D. Sawarkar, Mrs. SwatiNarwane- International Journal of Engineering Development and Research- "Detection of URL based phishing attacks using machine learning": A Survey; Volume 7, Issue 2; ISSN: 2321-9939.

[5].https://www.saedsayad.com/naive_Bayesianian.htm

[6].https://www.searchmetrics.com/glossary/index/#:~:text=An%20index%20is%20another%20name, be%20able%20to%20find%20it.

[7]. Basnet R., Mukkamala S., Sung A.H. (2008) Detection of Phishing Attacks: A Machine Learning Approach. In: Prasad B. (eds) Soft Computing Applications in Industry. Studies in Fuzziness and Soft Computing, vol226. Springer, Berlin, Heidelberg.https://doi.org/10.1007/978-3-540-77465-5_19

[8].https://analyticsindiamag.com/jupyterlab-what-is-it/

[9]. Author- Taheri S., sonataheri@students.ballarat.edu.au; sona.taheri@unisa.edu.au-Centre for Informatics and Applied Optimization, School of Science, Information Technology and Engineering, University of Ballarat, Victoria 3353, Australia
[10]. Author- Mammadov M., m.mammadov@ballarat.edu.au; musa.mammadov@nicta.com.au-Centre for Informatics and Applied Optimization, School of Science, Information Technology and Engineering, University of Ballarat, Victoria 3353, Australia; Victoria Research Laboratory, National ICT Australia, Victoria 3010, Australia