# CLASSIFICATION OF ORIGINAL AND FAKE WEBSITE USING URL

**Mrs. S.DEEPIKA**[1]
Assistant Professor, Department of ( A&F and BA)
PSGR Krishnammal College For Women, Coimbatore, India.
deepika@psgrkcw.ac.in
**P.ABARNA**[2]
UG Scholar, Department of Business Analytics
PSGR Krishnammal College For Women, Coimbatore, India.
abarnaprakash01@gmail.com

**ABSTRACT**
Websites have the ability to track what peoples are searching and it will also show related searches. The main thing in using website is peoples are unaware of realizing the difference between original and fake websites. Naïve Bayes Classifier Algorithm is a Machine Learning Algorithm that have used to classify the difference between the original and fake websites using selective attributes. Naives Bayes Algorithm is a supervised learning in Machine Learning where all algorithm have one basic common principle.

**Keywords**: Naïve Bayes Classifier, Classification, Machine Learning.

## I.INTRODUCTION

Websites are now a days used by all around the world. From that all the websites are comes with some special characters, integer, alphabets, symbols. Analysis of black list in website fraud detection is a review to show how websites were hacked by fraud. This paper determines and analyse the modules in accordance with the attributes introduced in dataset. The tool we used for this project is “JUPYTER” which makes the coding easier. This project determines the count of original websites and fake websites. Naïve Bayes one of the most popular classification algorithm which has been used to do simplest predictions. This helps in building the fast machine learning models that makes quick predictions as much as possible. Machine learning is a subset of artificial intelligence. This presents first a machine learning tree, and then focuses on the matrix algebra methods in machine learning including single-objective optimization, feature selection, principal component analysis, and canonical correlation analysis together with supervised, unsupervised, and semi-supervised learning and active learning. More importantly, this chapter highlights selected topics and advances in machine learning: graph machine learning, reinforcement learning, Q-learning, and transfer learning.[10]

## II.RELATED WORKS

Machine learning needs two things to work, data (lots of it) and models. When acquiring the data, be sure to have enough features populated to train correctly our learning model. The primary data collected from the online sources remains in the raw form of statements, digits and qualitative terms. The raw data contains error, omissions and inconsistencies. It requires corrections after careful scrutinizing the completed questionnaires. Types of approaches-Supervised learning algorithms and Semi-supervised algorithms, Unsupervised learning algorithms, Reinforcement Learning. [6]

The phishing detection technique is divided into blacklist based and heuristic based approaches. In blacklist based approaches, it maintains a database which has the list of the URL address and they are classified as malicious. The advantages of blacklist based approaches are easy implementation and low falls positive rate. But it can’t detect the phishing site that is not in a database. Pattern recognition was another method which is used for detecting the phishing websites.[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. Classification algorithms- Naive Bayes Classifier, Logistic Regression, Decision Trees, Random Forest, Neural Network, Nearest Neighbor, Support vector machine (SVM).[7]

Presence of IP address in URL, if IP address present in URL then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.[9]

Short URLs have become ubiquitous. Especially popular within social networking services, short URLs have seen a significant increase in their usage over the past years, mostly due to Twitter's restriction of message length to 140 characters.[2]
One solution approach is to use a blacklist of malicious URLs developed by anti-virus groups. The problem with this approach is that the blacklist cannot be exhaustive because new malicious URLs keep cropping up continuously. Thus, approaches are needed that can automatically classify a new, previously unseen URL as either a phishing site or a legitimate one. Such solutions are typically machine-learning based approaches where a system can categorize new phishing sites through a model developed using training sets of known attacks. [1]

## III.METHODOLOGY

### SUPERVISED LEARNING

A supervised learning algorithm is adopted here to build model is Naïve Bayes. This section gives a brief overview of this algorithm. This classifier is elementary Bayes theorem. It can achieve relatively good performance on classification tasks [8].

### NAÏVE BAYES CLASSIFIER ALGORITHM

Naive Bayes classifier greatly simplifies learning by assuming that features are independent given the class variable. In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. In spite of their naive design and apparently over-simplified assumptions, Naive Bayes classifiers have worked quite well in many complex real-world situations. An advantage of the Naïve Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. [5]

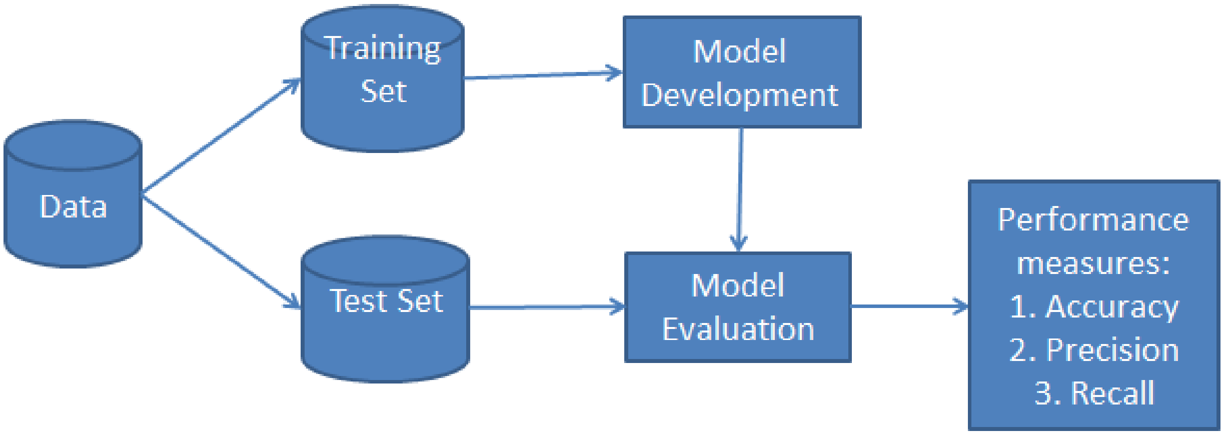
The class with the highest posterior probability is the outcome of prediction. ... **Naive Bayes** uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly **used** in text classification and with problems having multiple classes.

### BAYESIAN THEOREM

A Bayesian network consists of a structural model and a set of conditional probabilities. The structural model is a directed graph in which nodes represent attributes and arcs represent attribute dependencies. Attribute dependencies are quantified by conditional probabilities for each node given its parents. Bayesian networks are often used for classification problems, in which a learner attempts to construct a classifier from a given set of training examples with class labels.[4]

**FLOWCHART OF NAÏVE B AYES CLASSIFIER ALGORITHM**

**FIG- 3.1**

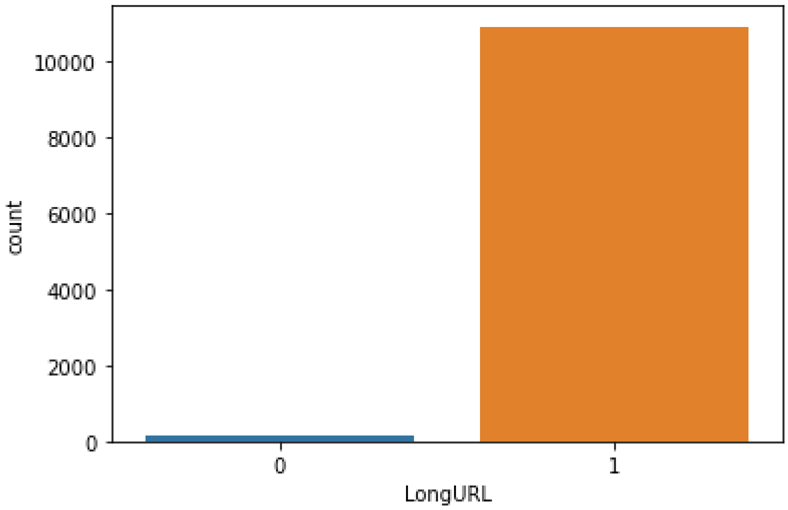


This fig 3.1 is a data flow diagram that determines entire process of the project. The Dataflow Diagram provides information about the users input and output and defined with rectangle, oval and so on. Here this describes the data by training and testing the dataset by model development and model evaluation process. And the output will be predicted by accuracy, precision and recall.

## IV. RESULT

**FIG- 4.1**

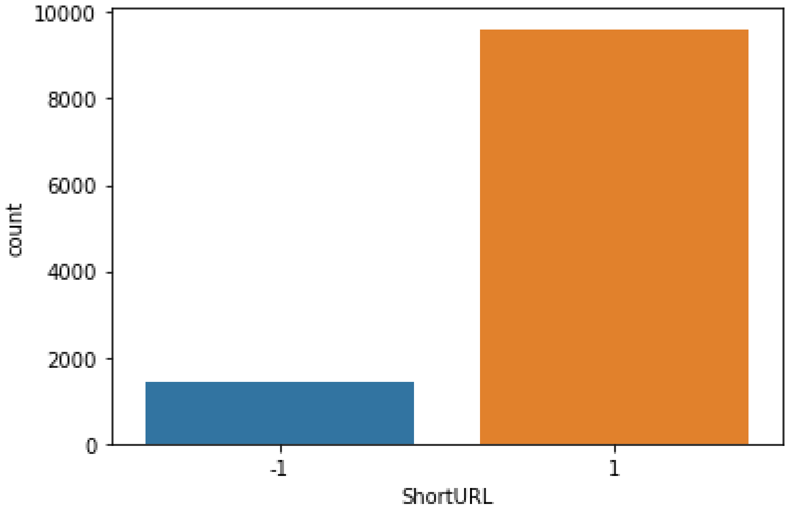
```
<AxesSubplot:xlabel='LongURL', ylabel='count'>
```



In the fig 4.1, it determines the count of original websites and fake websites in Long URL. It satisfies that the original website count is higher than the fake website count.The original website count is above 10000 and the fake website count lies between 0 and 2000.

**FIG- 4.2**

```
<AxesSubplot:xlabel='ShortURL', ylabel='count'>
```



In the fig 4.2, it determines the count of original websites and fake websites in Short URL. It also satifies the original website count is higher than the fake website count. This diagram plot shows that 1 have the higher count on original website and lies between 8000 and 10000 and -1 lies between 0 and 2000.

**USING NAÏVE BAYES (FINDINGS)**

Gaussian NB() is a Naïve Bayes Classifier used to predict the accuracy of the selective attributes i.e Long URL, Short URL. This Long and short URL are test as considering Long URL as predict_ X, Short URL as predict_ Y1. Then, these two attributes are train by train predict_X, train predict_Y. After test train functions it is important to apply Naïve Bayes Classifier by predicting array values. Then output will be analysed by testing, predicting and multiplying.The accuracy score value is 87. 626 that accuracy implies that there is 87% difference between the original and fake websites. In other words, 87% of original websites have found in accordance to the selective attributes.

## V. CONCLUSION & FURTHER WORK

In this paper, Jupyter Notebook tool is used to analyse the accuracy of original website using Naïve Bayes Classifier, a supervised machine learning algorithm. The performance of this classification is easy to predict different numerical and categorical variables and it is adoptable for solving different or multi-class prediction problems.

### FURTHER WORK

It is suggested to enhance the Naïve Bayes Classification on predicting numerical variables. Sentiment Analysis, spam filtering, recommendation systems etc.., will be made easy if analysing by Naïve Bayes Classification.

### REFERENCES

1. Arun Kulkarni , Leonard L. Brown, (IJACSA) International Journal of Advanced Computer Science and Applications- "Phishing Websites Detection using Machine Learning";Volume- 10,Issue No- 7; 2019.
2. Demetris Antoniades, Iasonas Polakis, Georgios Kontaxis, Elias Athanasopoulos, March 2011, DOI:10.1145/1963405.1963505, Source- DBLP, Conference: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011
3. Dr. G.Ravi Kumar, Dr. S. Gunasekaran, Nivetha.R, Sangeetha Prabha.K, Shanthini.G, Vignesh. A.S.- International Journal of Engineering Applied Sciences and Technology- "URL Phising data analysis and detecting phising attacks using Maching Learning in NLP"; Vol-3, Issue10, ISSN No: 2455-2143, February- 2019.
4. Harry Zhang, Liangxiao Jiang, Jiang Su, Copyright c 2005, American Association for Artificial Intelligence (www.aaai.org).
5. Jasmina Novakovic, Faculty of Public Administration, Megatrend University, Bulevar umetnosti 29, 11070 Novi Beograd, Srbija (telefon: 381-11-2092111, e-mail: jnovakovic@megatrend.edu,rs )
6. Mrs. Vaneeta M, Pratik N N, Prajwal D, Pradeep K S, Suhas Kakade- Journal of Emerging Technologies and Innovative Research (JETIR)-" Detection of Phising Websites using Machine Learning techniques"; Volume 7,Issue 6; June 2020.
7. Ms. Sophiya. Shikalgar, Dr. S. D. Sawarkar, Mrs. Swati Narwane- 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.
8. P. Domingos, and M. Pazzani, ³Feature selection and transduction for prediction of molecular bioactivity for drug design´, Machine Learning, 29:103-130, 1997
9. Rishikesh Mahajan, Irfan Siddavatam- International Journal of Computer Applications 181(23):45-47- "Phishing Website Detection Using Machine Learning Algorithms"; Volume 181, Issue 23; October 2018.
10. Xian- Da Zhang, A Matrix Algebra Approach to Artificial Intelligence, pp 223-440| cite as Machine Learning, Department of Automation, Tsinqhua University, First Online: 23 May 2020