COMPARISON OF THE PERFORMANCE OF THE LEARNING ALGORITHMS FOR VERIFICATION PHISHING UNIFORM RESOURCE LOCATOR (URL) USING MACHINE LEARNING

A MINI THESIS SUBMITTED IN PARTIAL FULFULMENT

OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF

SCIENCE IN INFORMATION TECHNOLOGY

OF

THE UMIVERSITY OF NAMIBIA

BY

KEPHAS LOIDE MWEDIFOLA

201401571

November 2018

MAIN SUPERVISOR: MR. GEBRIEL .T. NHINDA

CO-SUPERVISOR: DR. NALINA SURESH

## ABSTRACT

Phishing is an online criminal act that occurs when a malicious webpage mimics a legitimate webpage to acquire sensitive information from the user [1]. Detecting phishing websites is one of the crucial problems facing the internet community specifically via emails because of its high impact on the day-to-day online transactions performed. Moreover, there is no doubt that phishing, as a phenomenon, is both highly successful and generally difficult to detect and prevent in a reasonable amount of time [2]. Namibia has experienced its own share in cyber-attacks in the ream of electronic banking transactions which prompt the Namibia government to come up with a draft bill on electronic transactions and cybercrime [3]. Despite a number of solutions to mitigate phishing by previous researchers, there is still no conclusive solution to phishing attacks particularly in the universities environment, and university of Namibia (UNAM) is not an exception. Therefore, this study aims to compare the performance of learning algorithms (Naïve Bayesian, Decision tree, and Logistic regression) for verification of phishing URLs using machine learning techniques. Furthermore, the study provide a better understanding on two or more machine learning algorithms that could be used to verify and confirm compromised and phishing URLs in the cyberspace. The study focused mainly on experimental research approach and principle of Personal extreme programming (PXP) development methodology is used for this prototype. PXP is designed to be applied by individual software engineers and is iterative. Applying its practices allows developer to be more flexible and responsive to changes [2]. The experiment is performed using a 7030 URLs dataset, which were divided into two samples: training and testing, 80% for training and 20% for testing and the observed result showed that decision tree provided the good accuracy of 91% as compared to Naïve Bayesian 58% and Logistic regression 85% respectively.

**Keywords**- Phishing URL, Naïve Bayesian, Decision Tree, Logistic regression, Machine learning, lexical features

# List of Publication(s)/Conference(s) proceedings

Table of Contents

[ABSTRACT i](#_Toc530050181)

[List of Publication(s)/Conference(s) proceedings ii](#_Toc530050182)

[List of Tables iii](#_Toc530050183)

[List of Figures iii](#_Toc530050184)

[List of Abbreviations and/or Acronyms iii](#_Toc530050185)

[ACKNOWLWEDGEMENTS iv](#_Toc530050186)

[DECLARATIONS iv](#_Toc530050187)

[1. Introduction 1](#_Toc530050188)

[1.1 Background of the study 1](#_Toc530050189)

[1.2 Statement of the problem 2](#_Toc530050190)

[1.3 Objectives of the study 2](#_Toc530050191)

[1.4 Significance of the study 3](#_Toc530050192)

[1.5 Limitation of the study 3](#_Toc530050193)

[2. Literature Review 3](#_Toc530050194)

[3. Research Methods / Development Methodology 5](#_Toc530050195)

[3.1 Research Design 8](#_Toc530050196)

[3.2 Population and Sample 8](#_Toc530050197)

[3.3 Research Instruments 9](#_Toc530050198)

[3.4 Procedure 9](#_Toc530050199)

[3.5 Data analysis 9](#_Toc530050200)

[*Figure 1* 10](#_Toc530050201)

[4. Research Ethics 10](#_Toc530050202)

[5. Results: 10](#_Toc530050203)

[*Figure 2* 11](#_Toc530050204)

[6. Discussions: 11](#_Toc530050205)

[7. Conclusions: 12](#_Toc530050206)

[8. Recommendations 12](#_Toc530050207)

[9. References 12](#_Toc530050208)

[Appendices 14](#_Toc530050209)

# List of Tables

# List of Figures

# List of Abbreviations and/or Acronyms

|  |  |
| --- | --- |
| **URL** | Uniform Resource Locator |
| **PXP** | Personal extreme programming |
| **IT** | Information Technology |
| **NB** | Naïve Bayesian |
| **DT** | Decision Tree |
| **LR** | Logistic Regression |
| **UNAM** | University of Namibia |
| **FNR** | False Negative Rate |
| **FPR** | False Positive Rate |
| **TP** | True Positive |
| **TN** | True Negative |
| **FP** | False Positive |
| **FN** | False Negative |

## ACKNOWLWEDGEMENTS

I am thankful to my uncle Thomas H. Adila for loving and always being there for me and to my family members for their love, support and patience along the way. I thank Dr. N. Suresh of University of Namibia, School of computing, Information technology department for supervision and guidance through this process, her dedication and commitment was highly appreciated.

## DECLARATIONS

I, Kephas Loide Mwedifola, hereby declare that this study is my own work and is a true reflection of my research, and that this work, or any part thereof has not been submitted for a degree at any other institution.

No part of this thesis/dissertation may be reproduced, stored in any retrieval system, or transmitted in any form, or by means (e.g. electronic, mechanical, photocopying, recording or otherwise) without the prior permission of the author, or The University of Namibia in that behalf.

I, Kephas Loide Mwedifola, grant The University of Namibia the right to reproduce this thesis in whole or in part, in any manner or format, which The University of Namibia may deem fit.

……………….……………… ……………………………… ….………...  
**Name of Student Signature Dat**

# Chapter 1: Introduction

## 1.1 Background of the study

Phishing is an online criminal act that occurs when a malicious webpage mimics as a legitimate webpage so as to acquire sensitive information from the user [1]. Detecting phishing websites is one of the crucial problems facing the internet community because of its high impact on the day-to-day online transactions performed. There is no doubt that phishing, as a phenomenon, is both highly successful and generally difficult to detect and prevent in a reasonable amount of time [3]. Furthermore, detection of phishing URLs has become increasingly difficult due to the evolution of phishing operations and the efforts to avoid mitigation by blacklists. The current state of cybercrime has made it possible for a phisher to host operations with short lifecycles that diminish blacklist effectiveness [4]. Moreover, a phisher uses social engineering and technical deception to fetch private information from the web user. The phishing web pages generally have alike page layouts, blocks and fonts to mimic legitimate web pages in an endeavor to influence web users to obtain personal details such as username and password. Over the years, online baking has become very popular as more financial institutions have begun to offer free online services [5].

## Statement of the problem

The fast growth and progress of phishing techniques create several challenge in web security. Furthermore, security analysts throughout the world are constantly challenged by the phishing community as new and advanced methods are developed each day. In addition, Namibia is not an exeption as it has experienced its own share in cyber-attacks in the realm of electronic banking transactions which prompt the Namibian government to come up with a draft bill on electronic transactions and cybercrime [6].

Moreover, despite several solutions to mitigate phishing by previous researchers, there is still no conclusive solution to phishing attacks particularly in the Universities environment, and University of Namibia (UNAM) is not an exception. Currently, at UNAM, there is Cyberoam in place, but little study has been done on the performance study. Hence the mail server still receives phishing URLs and Information Technology (IT) infrastructure department staff, had to warn the users not to open such emails. Therefore, there is a needs to try and come up with a system that can tackle the situation to mitigate Phishing URLs.

## Objectives of the study

The main objective of this study was to evaluate the performance of learning algorithms for verification of phishing URLs using machine learning.

The following were the sub-objectives of the study:

1. To determine which of the three algorithms is suitable for detecting phishing URLs.
2. To use standard metrics for measuring the performance of the learning algorithms (Naïve Bayes classifier, Decision Tree and Logistic Regression) to benchmark the performance of the algorithms.
3. To evaluate each model from the perspective of accuracy, precision, recall and F-measure.

## Significance of the study

The study was significant as it provide a better understanding on two or more machine learning algorithms that could be employed to verify and confirm compromised and phishing URLs in the cyberspace.

## Limitation of the study

In this study, only three learning algorithms were trained and tested basically: Naïve Bayesian, Decision Tree and Logistic Regression and they are only be trained and tested based on lexical features.

# Chapter 2: Literature Review

This section provides an overview on some of the major studies conducted on phishing URL and the algorithms to detect phishing URLs:

Basnet and sung [7], proposed a novel approach for classifying legitimate malicious URLs using supervised learning across the features from various web services. They applied the web mining-based heuristics on logistic regression classifier and demonstrate that Logistic Regression can detect phishing URLs with an accuracy of 99%. However, the content-based approach requires access to the phishing site. Moreover, the heuristic can still be integrated with a keyword, lexical, host and content-based features to improve phishing URLs detection.

Azeez and Oluwatosin [8], explored how malicious link in emails can be detected from lexical and host-based features of their URLs to protect users from identity theft attacks using Naïve Bayesian classifier. However, even though Naïve Bayesian was the best for their approach, more classifier or algorithms could have been used to enhance their findings.

E and K [9], proposed a system that uses lexical features WHOIS features, PageRank and Alexa Rank and PhishTank-based features on random forest and content-based algorithms to classify phishing URLs. They demonstrated that by applying web mining heuristics on random forest algorithms, a precision of more than 90% was achieved and force negative rate (FNR) and force positive rate (FPR) rates of less than 1%. However, more improvement needs to be done on the content-based algorithm as only less than 65% precision was achieved. Moreover, there is a need to work on a selection of more features for the content-based algorithm to increase the precision and decrease the FNR and FPR.

Ma et al. [10] explore an online learning approach for classifying URLs automatically as either malicious or benign, based on supervised learning across both lexical and host-based features. However, their approach is complementary to blacklisting which cannot predict the status of a previously unseen URLs.

Blum et al.[4], explore the possibility of utilizing confidence weighted classification combined with content-based phishing URL detection to produce a dynamic and extensive system for detection of present and emerging type of phishing domains based on lexical features only. However, more features could be experimented using different learning algorithms to add to the value of the models and improve the accuracy.

Despite the amount of research done by the previous researcher, there still no definitive solution to phishing problem. Hence, there is a need to improve the suggested methods by the previous researchers to find a conclusive solution. In this study, the researcher studied the performance of learning algorithms for verification of phishing URLs, using Naïve Bayesian, Decision tree, and Logistic regression, to determine the suitable learning algorithm that can detect Phishing URLs.

# Chapter 3: Research Methods / Development Methodology

In this section, the researcher discussed the research design used in this study, the population and sampling process, the procedure on how data was collected and analysed and the software development methodology used (requirements, planning, iteration initialization, Design, Implementation, System testing and Retrospective).

3.1. Research Design

In order to meet the objective of this study, a quantitative research design with experimental as an approach used in this study. Experimental research approach is used during training and testing of algorithmic models based on lexical features using machine learning. According to [10], Lexical features are items of data selected from the URLs that allow us to capture this observable difference between the appearance of a legitimate URLs and that of phishing URLs. Machine learning is a set of techniques that allow implementing adaptive algorithms to make predictions and to auto-organize input data according to their common features [11].

3.2. Population and sampling

This study used Legitimate and Phishing that have been collected from University of Namibia, Computer centre, IT infrastructure department that were manually collected. The dataset was a collection of 7030 URLs used for experiments in machine learning such as detection URLs and URLs classification. The data is organised in two (2) groups, each corresponding to a different class. In order to evaluate the three algorithms in a scientific manner, the data is splitted into two section: training and testing data in order to have out of sample testing. This is important since Naïve Bayesian, Logistic Regression and Classification or regression (CART) are supervised learning algorithms, which means that one need to manually classify data into the correct classes then train a Naïve Bayesian, Logistic Regression or CART model with it and eventually use this model to predict unlabelled or testing data.  
3.3. Procedures

3.1.3. Procedures  
A standard dataset will be collected from the University of Namibia mainly Computer center,  
Information Technology (IT) Infrastructure department [14]. Furthermore, the data will be  
sorted into a standard dataset that will contain both legitimate and phishing URLs.  
Additionally, the dataset will be applied to the learning algorithms such as Naïve Bayesian,  
Decision Tree, and Logistic Regression, to evaluate the performance of the learning  
algorithms by comparing the result of the experiment, to determine whether this learning  
algorithm can classify the URLs as a legitimate or phishing URLs. Furthermore, a  
comparative analysis of the performance of learning algorithms will be made for verification  
of vulnerable and compromised URLs after the experiment.

The main objective of the study was to compare the performance of Naïve Bayesian, Logistic Regression and Decision Tree to detect phishing URLs.

This section describes the methodology that was used in the study namely the data used, requirement analysis, system framework and finally the system implementation.

The researcher proposed a lexical features approach to determine phishing URLs using information available only on URLs. The researcher extravagance the problem of detecting phishing URLs as a binary classification problem using supervised classification with legitimate URLs belonging to the negative class (0) and phishing URLs belonging to positive class (1). The research collected the phishing and legitimate URLs manually and create a dataset. The first script is to load the data set and extract feature by using various publically available resources in order to classify the instances into their corresponding classes.

The researcher used lexical features to predict phishing URLs and this were the X data and Y data which were URLs.

**Data set**

This study used Legitimate and Phishing that have been collected from University of Namibia, Computer centre, IT infrastructure department that were manually collected. The dataset was a collection of 7030 URLs used for experiments in machine learning such as detection URLs and URLs classification. The data is organised in two (2) groups, each corresponding to a different class. In order to evaluate the three algorithms in a scientific manner, the data is splitted into two section: training and testing data in order to have out of sample testing. This is important since Naïve Bayesian, Logistic Regression and Classification or regression (CART) are supervised learning algorithms, which means that one need to manually classify data into the correct classes then train a Naïve Bayesian, Logistic Regression or CART model with it and eventually use this model to predict unlabelled or testing data.

**Requirements analysis**

This section defined the requirements for this study:

1. Reading the dataset, which means loading the dataset into memory.
2. Preparing the data by labelling the data and extraction of features.
3. Creation and training of Naïve Bayesian, Logistic Regression and Decision Tree models, the three algorithms under supervised learning algorithms.
4. Testing of the models, in order to decide whether a model is accurately capturing a pattern, in order to evaluate the model. Using the testing subset of the dataset, the three models were tested and evaluated based on, precision, recall and F-measure metrics.

**System design**

This section show the framework of URLs verification and an overview of the tools used in this study. The following figure 3 shows the conceptual design.

Input

X

ytrain

ytest

Xtrain

xtest

Data

ML Algorithm

Generated Model

M

Output

Y?

Data

y



*Figure 3: URLs verification using learning algorithms*

**Detailed design description**

**System Implementation**

**Development tools**

The tools used in the study were:

1. Python programming language: version 3.6.6
2. Scikit-learn: An open source machine learning library for Python programming language.
3. Numpy: The fundamental package for the scientific computing with python.
4. Scipy: An open source library of scientific tools.
5. Matplotlib: a python 2D plotting library for producing publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

## Research Design

In order to meet the objective of the study, a quantitative research design with experimental as an approach was used. Experimental research approach was used during training and testing of data, on the performance of the three learning algorithms based on lexical features using machine learning techniques. According to [10], Lexical features are items of data selected from the URLs that allow us to capture this observable difference between the appearance of a legitimate URLs and that of phishing URLs. Machine learning is a set of techniques that allow implementing adaptive algorithms to make predictions and to auto organize input data according to their common features [11].

## Population and Sample

A standard dataset consisting of legitimate and phishing URLs was there targeted  
population for the study. Furthermore, a purposive sampling is used to select the sample of the study. Purposive sampling is a sampling technique that deliberately hand-picks the sample by choosing instances that are likely to produce variable data to meet the purpose of the research [12]. The study is interested in URLs both phishing and legitimate URLs. Therefore, a purposive sampling is used for this study to select the sample.

## Research Instruments

The tools used in the study were:

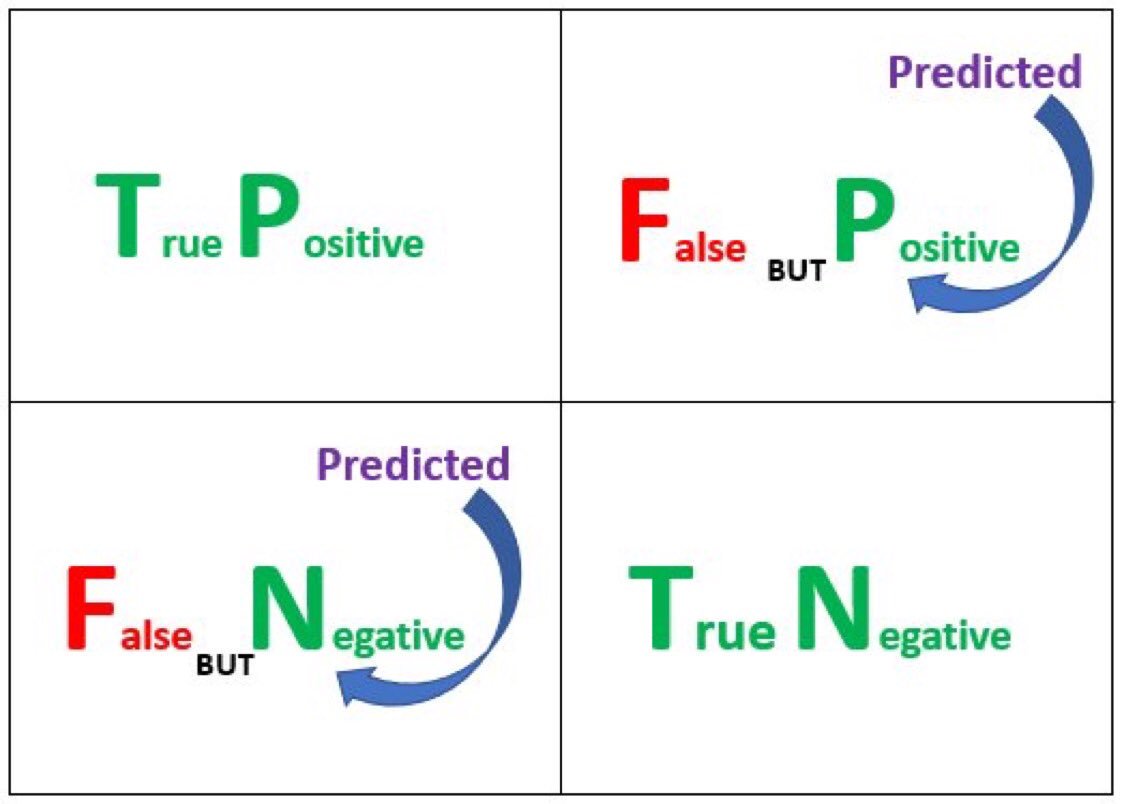
1. Python programming language: version 3.6.6
2. Scikit-learn: An open source machine learning library for Python programming language.
3. Numpy: The fundamental package for the scientific computing with python.
4. Scipy: An open source library of scientific tools.
5. Matplotlib: a python 2D plotting library for producing publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

## 3.4 Procedure

A standard dataset is collected from the University of Namibia mainly Computer centre, Information Technology (IT) Infrastructure department [14]. Furthermore, the data is sorted into a standard dataset that contain both legitimate and phishing URLs. Additionally, the dataset is applied to the models to evaluate the performance of the models by comparing the result of the experiment and determining at what precision can this models (learning algorithm) predict the URLs as a legitimate or phishing URLs. Moreover, a comparative analysis of the performance of the models were made for verification of legitimate and compromised URLs after the experiment.

## Data analysis

In this study, the researcher compared generated results of the three models (learning algorithms) and interpret the evaluation of the models using a confusion matrix to provide empirical evidence to support this study. Moreover, by evaluating the performance of each algorithm, their overall accuracy is compared in order to determine the suitable algorithm in detection phishing URLs. The following figure 1 by [11] illustrates the confusion matrix to benchmark three algorithms table.



*Figure 1: confusion matrix by*

In this section, we focus on the performance of the binary classification model. Based on confusion matrix, corresponding four categories of shortcoming prediction result are described as follows:   
• True Positive (TP, collect classified phishing URLs)   
• True Negative (TN, collect classified phishing URLs)   
• Force Positive (FP, non-phishing URLs wrongly classified as phishing), and  
• Force Negative (FN, phishing URLs wrongly classified as non-phishing).

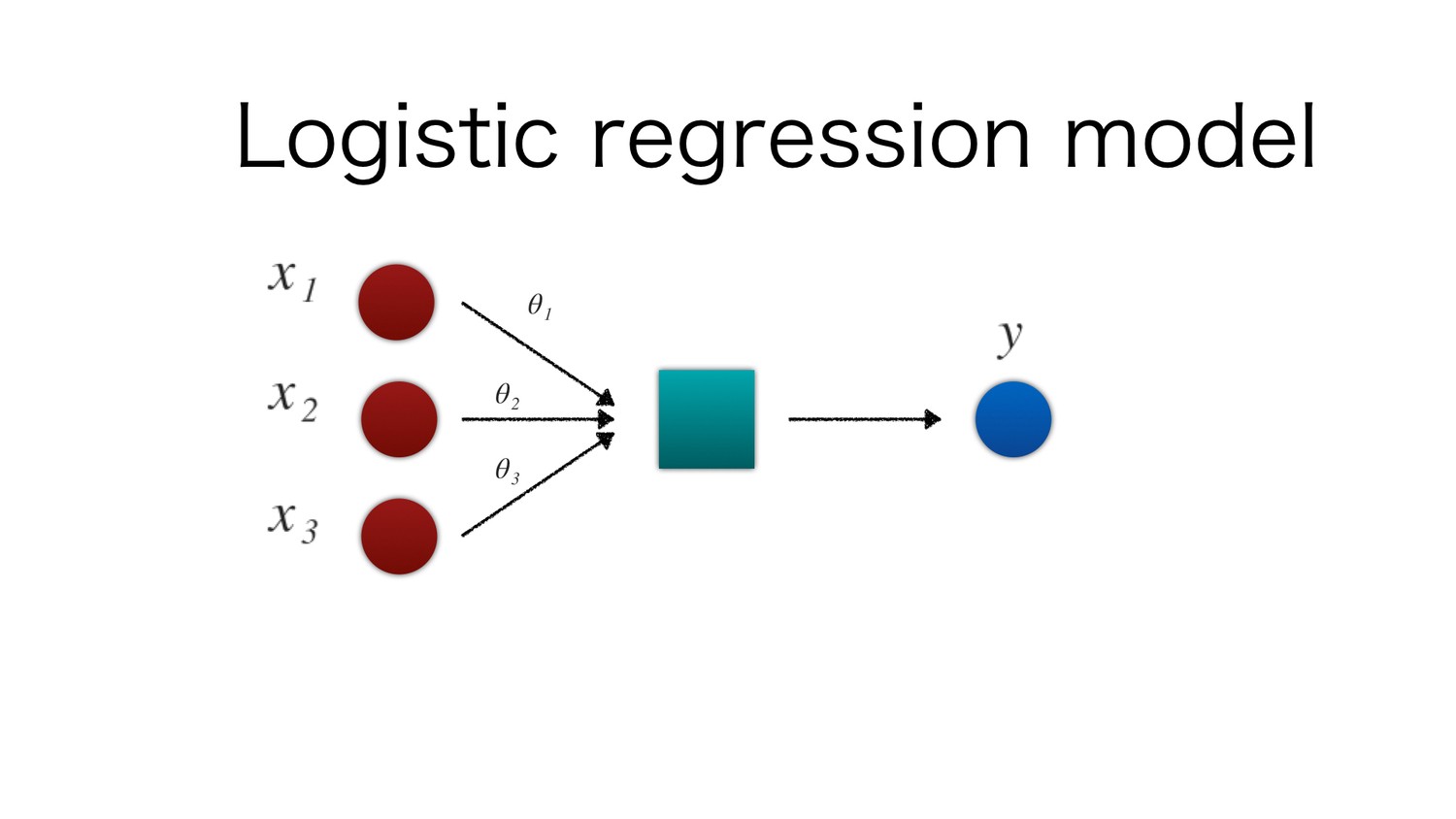
# Chapter 4: Research Ethics

The participants involved in the research, were asked first for their consent before any data was collected for the study. Moreover, the data was not shared with any third party and it’s only used for the purpose of this study.

# Chapter 5. Results:

In order to test whether a model is accurately grasping a pattern, the researcher have to evaluate the model. The result of this evaluation is used in deciding whether the model is trustworthy for the purpose it is used for and, this is an effective way for guiding the researcher in making future improvements to the model. The testing of data through prediction after extracting the relevant features was performed by using the following learning algorithms namely Naïve Bayesian, Logistic Regression and Decision Tree.

Logistic Regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome [12]. The outcome is measured with a dichotomous variable (in which there are only two possible outcome). It is used to predict a binary outcome (1 or 0, yes or No, True or False) given a set of independent variables. Logistic regression was developed by statistician David Cox in 1958. This binary model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). The following figure show a logistic regression model.



*Figure 2: Logistic regression*

# Chapter 6: Discussions:

The study aimed at comparing the performance of Naïve Bayesian, Logistic Regression and Decision Tree by verifying/classifying URLs. This was accomplished by using a dataset consisting of 7030 URLs. The models were evaluated by comparing their precision, recall and F-measure, Area under the curve (AUC), when the three models were run against the dataset.

# Chapter 7: Conclusions

Security analysts throughout the world are constantly challenged by the phishing community as new and advanced methods are developed each day. In this evolving environment, it’s every researcher’s main responsibility to deceive a system that can tackle the situation. In this study, when the researcher compared the different learning algorithms, the researcher identified Decision Tree as the suitable model for detecting phishing URLs. For future work the research intend to enhance the system by incorporating an online mode to improve the accuracy and help to achieve better performance as the system becomes dynamic.

# Chapter 8: Recommendations

In the future, it would be very interesting to investigate what makes each of the models perform the way they do, which can be from the score each model achieves in predicting the URLs, or the resources used, time and computational power and memory during training and testing. Furthermore, one can investigating the impact of feature extraction and since the algorithms were tested with a standard dataset it, it’s advanced to test them with newer dataset in the future. Spear phishing attack was also not attempted.

# Chapter 9: References

[1] S. Marchal, J. François, R. State, and T. Engel, “PhishStorm: Detecting Phishing With Streaming Analytics,” *IEEE Trans. Netw. Serv. Manag.*, vol. 11, no. 4, pp. 458–471, Dec. 2014.

[2] Y. Dzhurov, I. Krasteva, and S. Ilieva, “Personal Extreme Programming – An Agile Process for Autonomous Developers,” p. 8.

[3] P. D. Dudhe and P. L. Ramteke, “Detection of Websites Based on Phishing Websites Characteristics,” vol. 3, no. 4, p. 7, 2007.

[4] A. Blum, B. Wardman, T. Solorio, and G. Warner, “Lexical feature based phishing URL detection using online learning,” 2010, p. 54.

[5] S. C. Jeeva and E. B. Rajsingh, “Intelligent phishing url detection using association rule mining,” *Hum.-Centric Comput. Inf. Sci.*, vol. 6, p. 10, Jul. 2016.

[6] N. E. S. Reporter, “Cybercrime a threat to national security – Shanghala,” *New Era Newspaper Namibia*, 15-May-2018.

[7] R. B. Basnet and A. H. Sung, “Mining Web to Detect Phishing URLs,” in *2012 11th International Conference on Machine Learning and Applications*, 2012, vol. 1, pp. 568–573.

[8] N. A. Azeez and A. Oluwatosin, “CyberProtector: Identifying Compromised URLs in Electronic Mails with Bayesian Classification,” in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2016, pp. 959–965.

[9] B. E. and T. K., “Phishing URL Detection: A Machine Learning and Web Mining-based Approach,” *Int. J. Comput. Appl.*, vol. 123, no. 13, pp. 46–50, Aug. 2015.

[10] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists: learning to detect malicious web sites from suspicious URLs,” 2009, p. 1245.

[11] K. Borne, “Removing Confusion from the Confusion Matrix  —  False Negatives vs False Positives: http://bit.ly/2ni06pj  #abdsc #Statistics #DataScience #MachineLearning by @venksaiyanpic.twitter.com/vzRHY95XbA,” *@KirkDBorne*, 02-Feb-2018.

[12] P. Chandrayan, “Machine Learning Part 3 : Logistic Regression,” *Towards Data Science*, 26-Aug-2017. [Online]. Available: https://towardsdatascience.com/machine-learning-part-3-logistics-regression-9d890928680f. [Accessed: 19-Oct-2018].

# Appendices