Comparison of the performance of the learning algorithms for verification of phishing uniform resource locator (URLs) using machine learning

By

Kephas Loide M  
201401571

Main Supervisor:

Mr. Gabriel Tuhafeni Nhinda

Co-supervisor:

Dr. Nalina Suresh

A Research Proposal submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Information Technology  
 University of Namibia

Dated (11.June. 2018)

DECLARATION

I declare that this research paper is my original work and has not been presented to any other University for an academic degree.

Sign: …………………………………… Date: ………………………………………..

Kephas Loide Mwedifola

ACKNOWLEDGEMENTS

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.

Abstract

# Chapter 1

## Introduction

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 [2]. 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 [3]. 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 [4].

# Chapter 2

## Problem statement

Namibia has experienced its own share in cyber-attacks in the ream of electronic banking transactions which prompt the Namibian government to come up with a draft bill on electronic transactions and cybercrime [5]. However, due to rapidly evolving technologies, this regulation needs to be drafted with flexibility in mind, by taking into account the need for legal certainty and precision. Furthermore, 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 find a solution to mitigate Phishing URLs.

## Research objectives

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

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 apply an appropriate standard dataset to test the three learning algorithms on.
3. To evaluate the performance of these learning algorithms in terms of accuracy using confusion matrix.
4. To verify whether a URL is a legitimate or phishing URLs.
5. To evaluate each model from the perspective of precision, recall and F-measure.

# Chapter 3

## 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 [6], 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 [7], 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 [8], 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. [9] 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.[3], 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, we propose the performance of learning algorithms for verification of phishing URLs, using Naïve Bayesian, Decision tree, and Logistic regression, to determine the best learning algorithm that can detect Phishing

URLs.

# Chapter 4

## Research Methodology and Process

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 and phishing URLs belonging to positive class. 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 are the X data and Y data which are URLs. 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 Decision Tree 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 Decision Tree model with it and eventually use this model to predict unlabelled or testing data.

The researcher then apply various machine learning algorithms to build models from training data, which is comprised of pairs of feature values and class labels. A separate set of test data are then provided to the models, and the predicted class of the data instance is compared to the actual class of the data to compute the accuracy of the classification models. Figure 3 shows an overview of graphical representation of phishing detection framework.

Input

X

ytrain

ytest

Xtrain

Xtest

Data

 Set

ML Algorithm

Generated Model

M

Output

Y?

Data

y



*Figure 3: Shows a flow diagram for the system in the designing phase.*

Data set

A dataset of 7030 URLs was collected for the purpose of performing this comparison, of these, 3494 are legitimate URLs and 3536 are phishing URLs. The legitimate and Phishing have been collected from University of Namibia, Computer centre. The phishing URLs have been collected from UNAM computer centre, IT infrastructure department using Cyberoam.

Predicting and Result

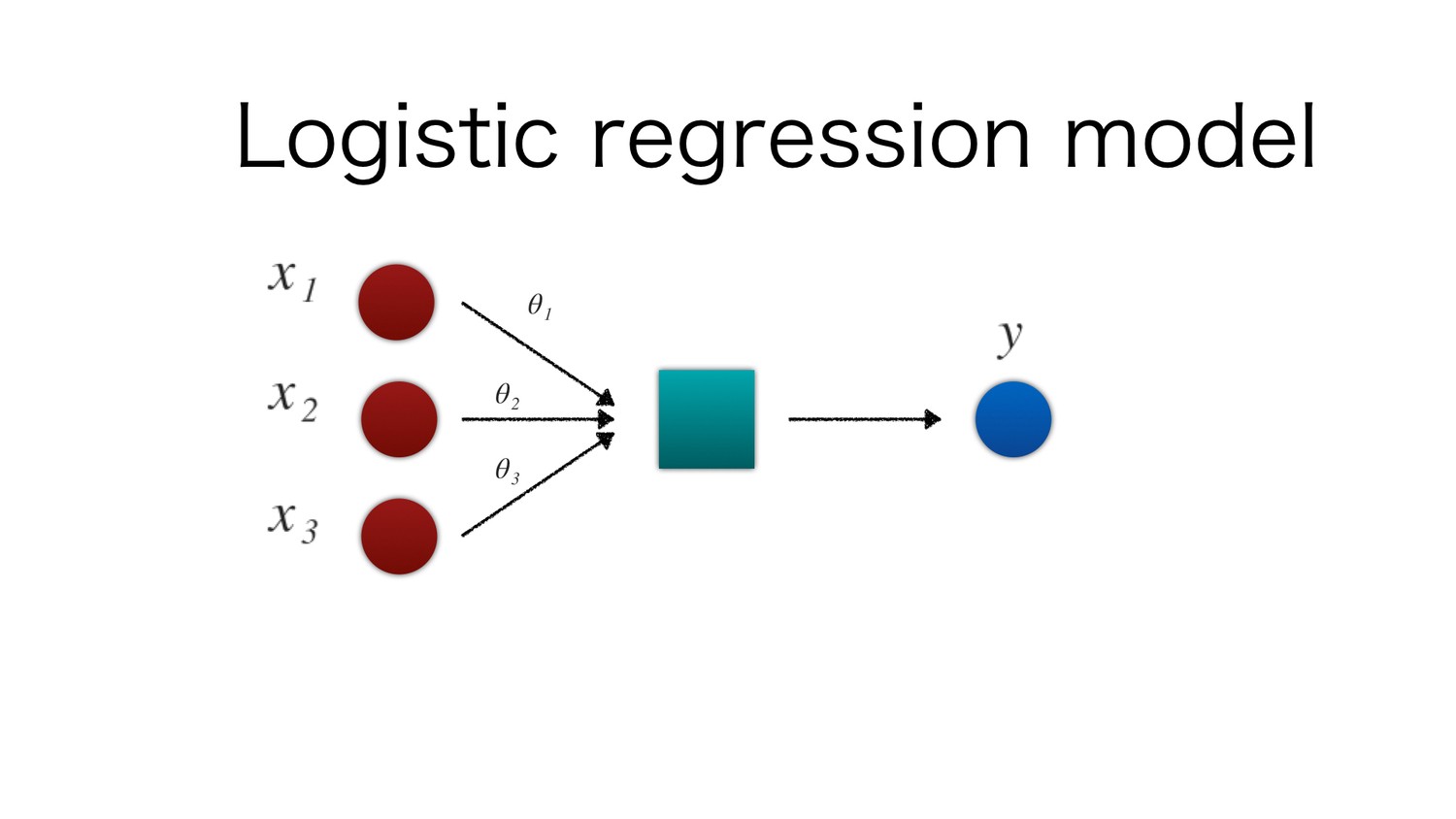
In order to decide whether a model is accurately capturing a pattern, the researcher have to evaluate that model. The result of this evaluation is important in deciding whether the model is trustworthy for the purpose it is used for. Furthermore, this is an effective way for guiding the researcher in 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.

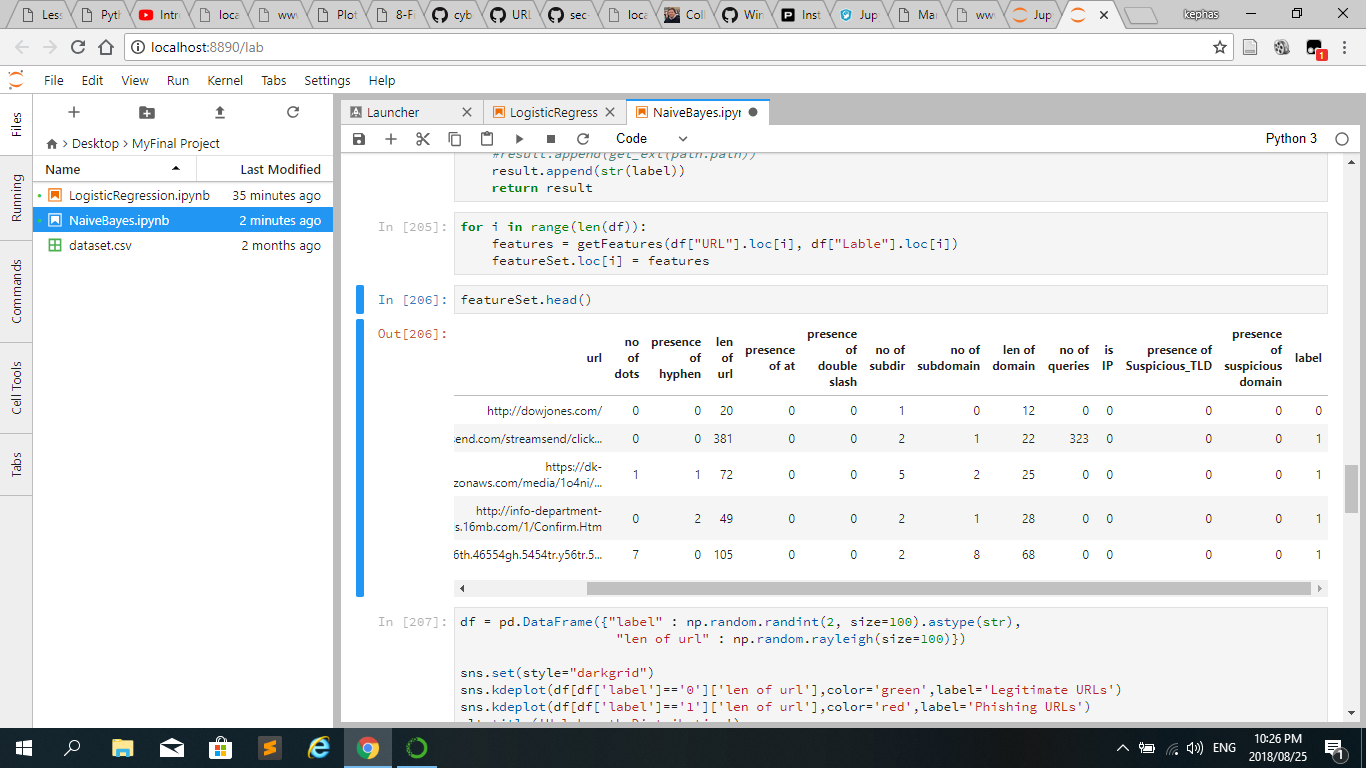
# Chapter 5

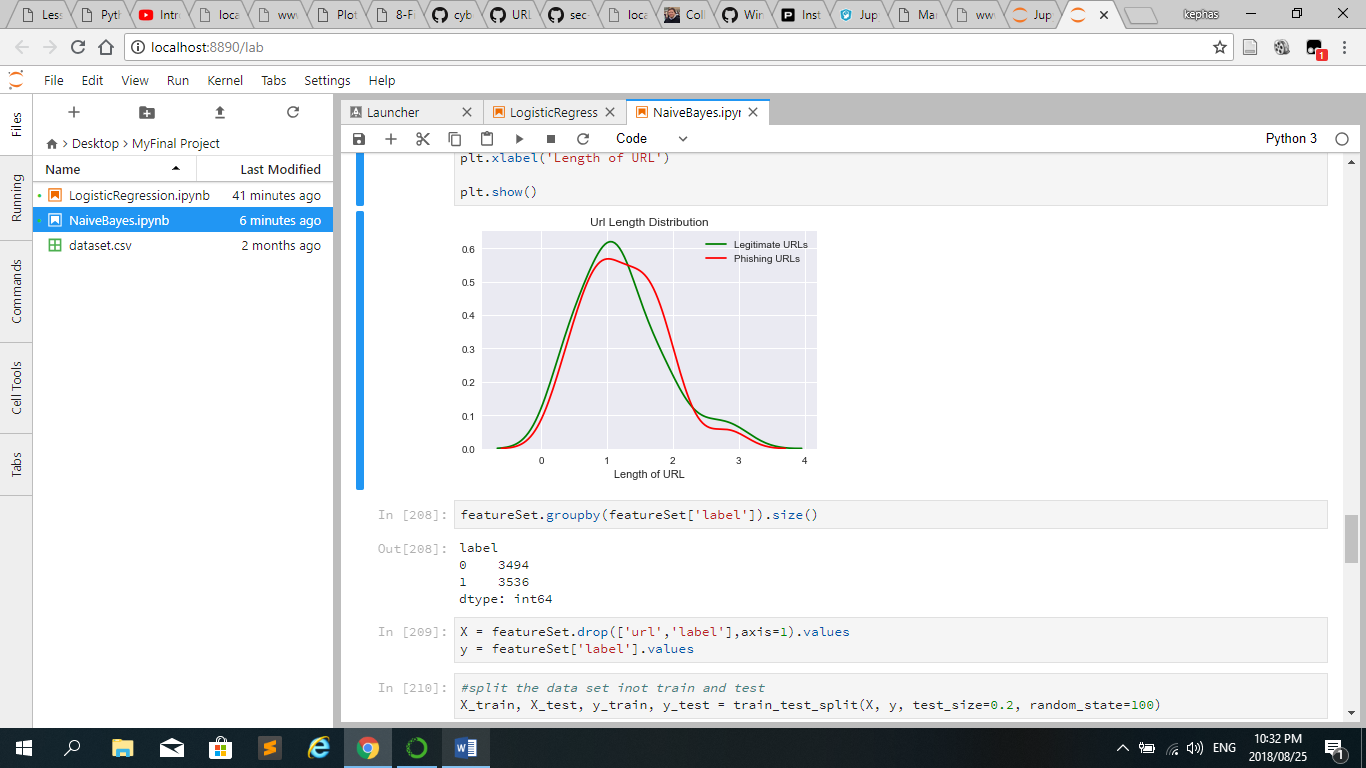
## Research Finding and Discussion

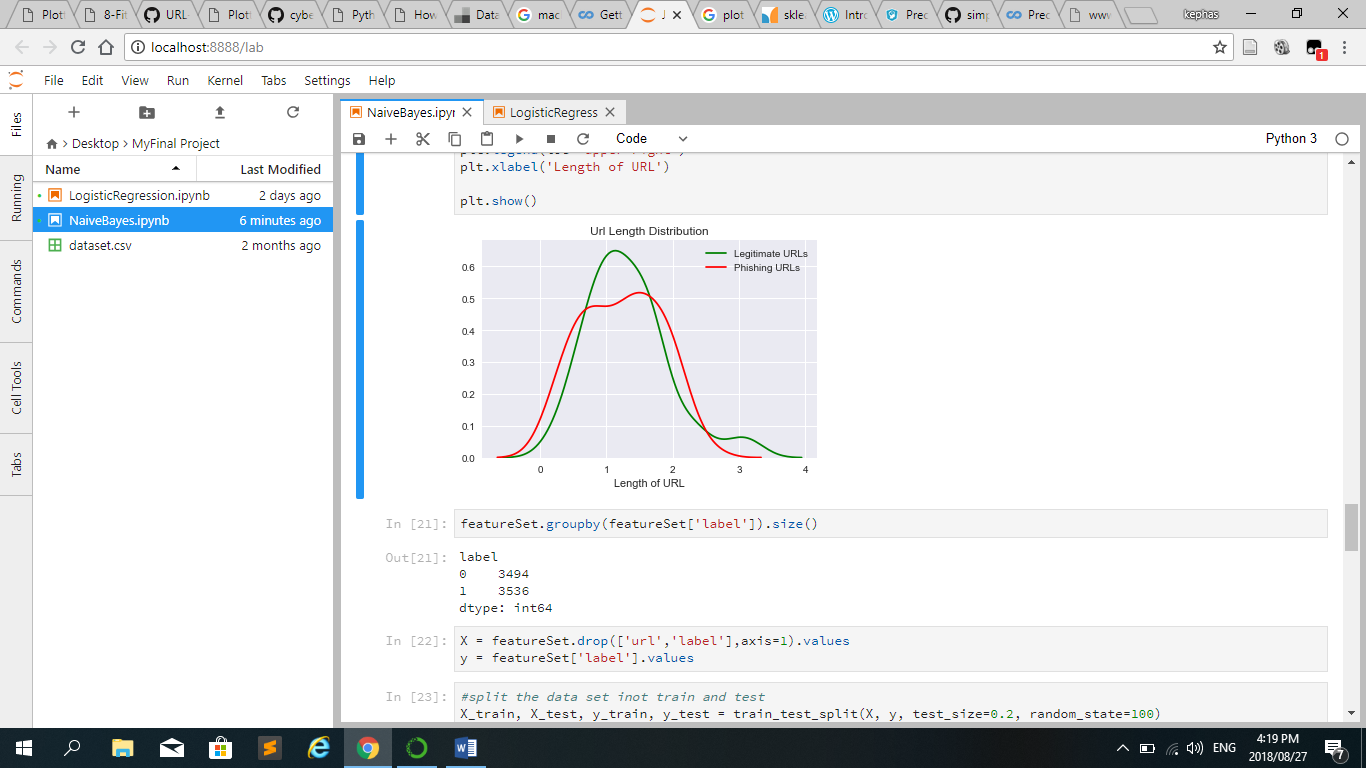
Logistic Regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome [10]. The outcome is measured with a dichotomous variable (in which there are only two possible outcome). It is used to predict a binary outcome (1 or 0, 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.

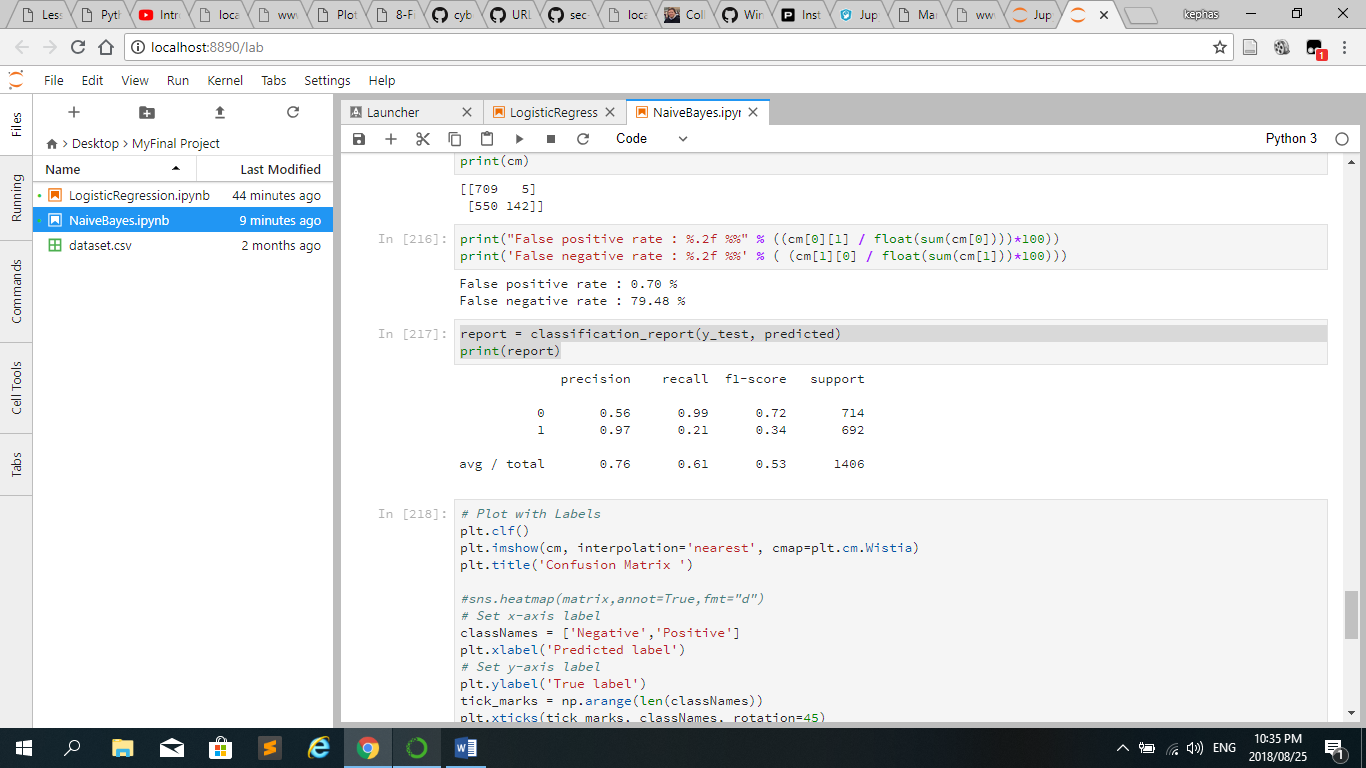


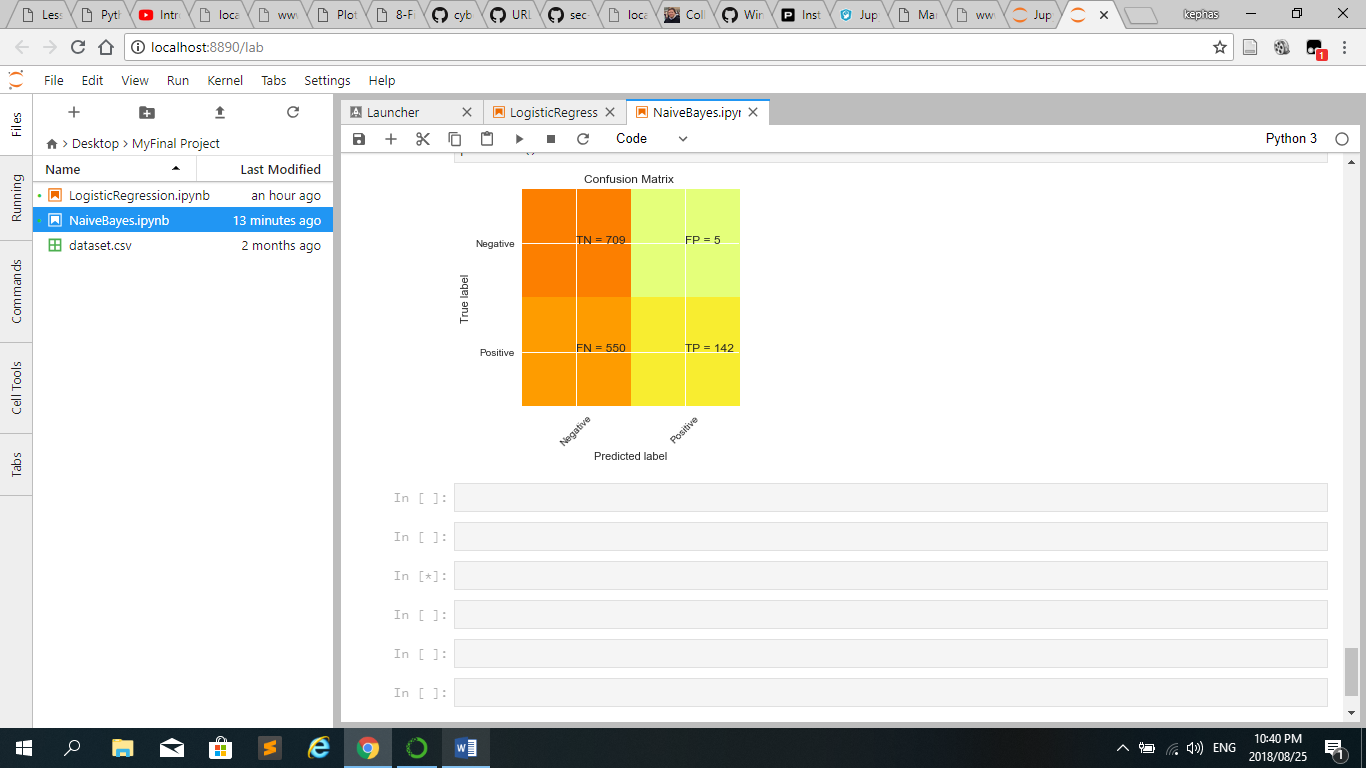
*Figure 1: Logistic regression*







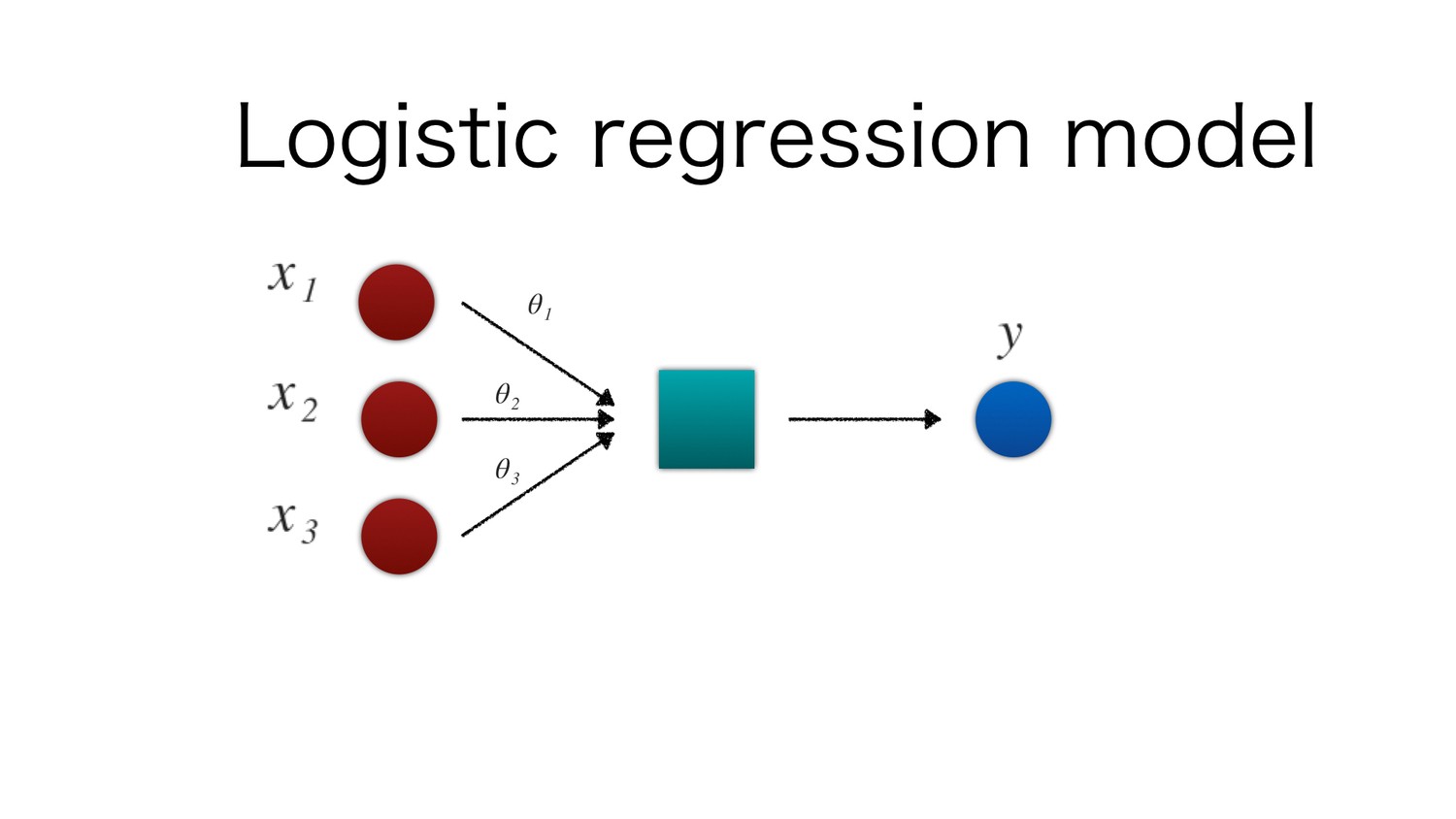




## Limitations of the study

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. Furthermore, all the data came from a single source which is UNAM computer center.

Logistic Regression is a machine learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logical regression, the depended variable is a binary variable that contains data coded as 1 or 0. In other words, the logistic regression model predicts P(Y = 1) as a function of X.



# Chapter 6

## Summary and Conclusion

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 the 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 7

## Future Work and Recommendations

Overfitting

Overfitting is when a machine learning model perform worse on new data than on their training data. Overfitting occurs when a model learns the details and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize [11].

Input

X

ytrain

ytest

Xtrain

xtest

Data

ML Algorithm

Generated Model

M

Output

Y?

Data

y



Model Reduction

Spear attacks