INTRO TO CYBERSECURITY

PHISHING DETECTION

LAB 1: WRITING A CLASSIFIER FOR PHISHING DATASET

Lab Description: This lab is to write the python script as well as use WEKA to implement a binary classifier to estimate whether a website is a phishing website. The dataset contains 102816 web hits and 30 features were recorded for each of the hit. Also, a class value has been given for each of the record.

Example of phishing dataset:

```
having_IP_Address,URL_Length,Shortining_Service,having_At_Symbol,double_slash_redirecting,Prefix_Suffix
-1,1,1,1,-1,-1,-1,-1,-1,1,1,-1,1,-1,1,-1,-1,-1,0,1,1,1,1,-1,-1,-1,-1,1,1,-1,-1
1,1,1,1,1,-1,0,1,-1,1,1,-1,1,0,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,1,1,-1
1,0,1,1,1,-1,-1,-1,-1,1,1,-1,1,0,-1,-1,-1,-1,-1,0,1,1,1,1,1,-1,1,-1,1,0,-1,-1
1,0,1,1,1,-1,-1,-1,1,1,1,-1,-1,0,0,-1,1,1,0,1,1,1,1,-1,-1,1,-1,1,-1,1,-1
1,0,-1,1,1,-1,1,1,-1,1,1,1,1,0,0,-1,1,1,0,-1,1,-1,1,-1,-1,0,-1,1,1,1,1
-1,0,-1,1,-1,-1,-1,1,1,-1,1,1,-1,1,0,0,-1,-1,-1,0,1,1,1,1,1,1,1,-1,1,-1,-1,1
1,0,-1,1,1,-1,-1,-1,1,1,1,-1,-1,0,-1,-1,-1,0,1,1,1,1,1,-1,-1,-1,1,0,-1,-1
1,0,1,1,1,-1,-1,-1,1,1,1,-1,-1,0,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,0,1,-1
1,0,-1,1,1,-1,1,1,-1,1,1,0,1,-1,1,1,0,1,1,1,1,1,-1,1,1,0,1,1
1,1,-1,1,1,-1,-1,1,-1,1,1,1,1,0,1,-1,1,1,0,1,1,1,1,1,-1,0,-1,1,0,1,-1
1,1,1,1,1,-1,0,1,1,1,1,1,-1,0,0,-1,-1,-1,0,1,1,1,1,-1,1,1,1,1,-1,-1,1
1,1,-1,1,1,-1,1,-1,-1,1,1,1,1,-1,-1,-1,-1,-1,0,1,1,1,1,-1,-1,-1,-1,1,0,-1,-1
-1,1,-1,1,-1,-1,0,0,1,1,1,-1,-1,-1,1,-1,1,1,0,-1,1,-1,1,1,-1,-1,-1,1,0,1,-1
1,1,-1,1,1,-1,0,-1,1,1,1,1,-1,-1,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,1,1,-1
1,1,-1,1,1,1,-1,1,-1,1,1,-1,1,0,1,1,1,1,0,1,1,1,1,-1,1,-1,1,-1,1,1
1,-1,-1,-1,1,-1,0,0,1,1,1,1,-1,-1,0,-1,1,1,0,1,1,1,1,1,-1,-1,-1,1,0,1,-1
1,-1,-1,1,1,-1,1,1,-1,1,1,-1,1,0,-1,-1,-1,-1,0,1,1,1,1,1,-1,0,-1,1,1,-1,-1
1,-1,1,1,1,-1,-1,0,1,1,-1,1,1,0,-1,-1,-1,-1,0,1,1,1,1,-1,1,1,-1,1,1,-1,-1
```

Features Description:

	Description		Values
having_IP_Address	If The Domain Part has an IP Address	Phishing	1
	Otherwise	Legitimate	-1
URL_Length	URL length<54	Legitimate	-1
	URL length≥54 and ≤75	Suspicious	0
	otherwise	Phishing	1
Shortining_Service	TinyURL	Phishing	1
	Otherwise	Legitimate	-1
having_At_Symbol	Url Having @ Symbol	Phishing	1
	Otherwise	Legitimate	-1
double_slash_redirecting	ThePosition of the Last Occurrence of "// " in the URL > 7	Phishing	1
	Otherwise	Legitimate	-1
Prefix_Suffix	Domain Name Part Includes (-) Symbol	Phishing	1
	Otherwise	Legitimate	-1
having_Sub_Domain	Dots In Domain Part=1	Legitimate	-1
	Dots In Domain Part=2	Suspicious	0
	Otherwise	Phishing	1
SSLfinal_State	Use https and Issuer Is Trusted &and Age of Certificate≥ 1 Years	Legitimate	-1
	Using https and Issuer Is Not Trusted	Suspicious	0
	Otherwise	Phishing	1
Domain_registeration_length	Domains Expires on≤ 1 years	Phishing	1
	Otherwise	Legitimate	-1
Favicon	Favicon Loaded From External Domain	Phishing	1
	Otherwise	Legitimate	-1
port	"Port # is of the " Preffered Status	Phishing	1
	Otherwise	Legitimate	-1
HTTPS_token	"Using " HTTP Token in Domain Part of The URL	Phishing	1
	Otherwise	Legitimate	-1
Request_URL	% of Request URL <22%	Legitimate	-1
	%of Request URL≥22% and 61%	Suspicious	0
	Otherwise	Phishing	1
URL_of_Anchor	% of URL Of Anchor <31%	Legitimate	-1
	% of URL Of Anchor ≥31% And≤67%	Suspicious	0
	Otherwise	Phishing	1
Links_in_tags	% of Links in "<Meta>","<Script>" and "<"Link>\""<17%	Legitimate	-1
	% of Links in <Meta>","<Script>" and "<"Link>\"" ≥17% And≤81%	Suspicious	0
	Otherwise	Phishing	1
SFH	SFH is "about: blank\"" Or Is Empty	Legitimate	-1
	SFH "Refers To " A Different Domain	Suspicious	0
	Otherwise	Phishing	1
Submitting_to_email	Using "mail()\" or \"mailto:\" Function to Submit User Information"	Phishing	1
	Otherwise	Legitimate	-1
Abnormal_URL	The Host Name Is Not Included In URL	Phishing	1
	Otherwise	Legitimate	-1
Redirect	times of Redirect Page≤1	Legitimate	-1
	times of Redirect Page≥2 &And<4	Suspicious	0
	Otherwise	Phishing	1
on_mouseover	onMouseOver Changes Status Bar	Phishing	1
	It Does't Change Status Bar	Legitimate	-1
RightClick	Right Click Disabled	Phishing	1
	Otherwise	Legitimate	-1
popUpWidnow	Popoup Window Contains Text Fields	Phishing	1
	Otherwise	Legitimate	-1
Iframe	Using iframe	Phishing	1
	Otherwise	Legitimate	-1
age_of_domain	Age Of Domain≥6 months	Phishing	1
	Otherwise	Legitimate	-1
DNSRecord	no DNS Record For The Domain	Phishing	1
	Otherwise	Legitimate	-1
web_traffic	Website Rank<100,000	Legitimate	-1
	Website Rank>100,000	Suspicious	0
	Otherwise	Phishing	1
Page_Rank	PageRank<0.2	Phishing	1
	Otherwise	Legitimate	-1
Google_Index	Webpage Indexed by Google	Legitimate	-1
	Otherwise	Phishing	1
Links_pointing_to_page	number of Link Pointing to The Webpage=0	Phishing	1
	number of Link Pointing to The Webpage>0 and≤2	Suspicious	0
	Otherwise	Legitimate	-1
Statistical_report	Host Belongs to Top Phishing IPs or Top Phishing Domains	Phishing	1
	Otherwise	Legitimate	-1
Result		Phishing	1
		Legitimate	-1

You are required to implement it in three ways:

- Using the machine learning software WEKA. (https://waikato.github.io/weka-wiki/downloading_weka/)
- Writing a python script with the use of the package sklearn
- Writing a python script with the use of the package tensorflow and deep learning techniques.

Lab Environment: The student should have access to no matter a machine with Linux system or Windows system, but the environment for python is required as well as some packages such as numpy, tensorflow, pandas, matplotlib, and sklearn.

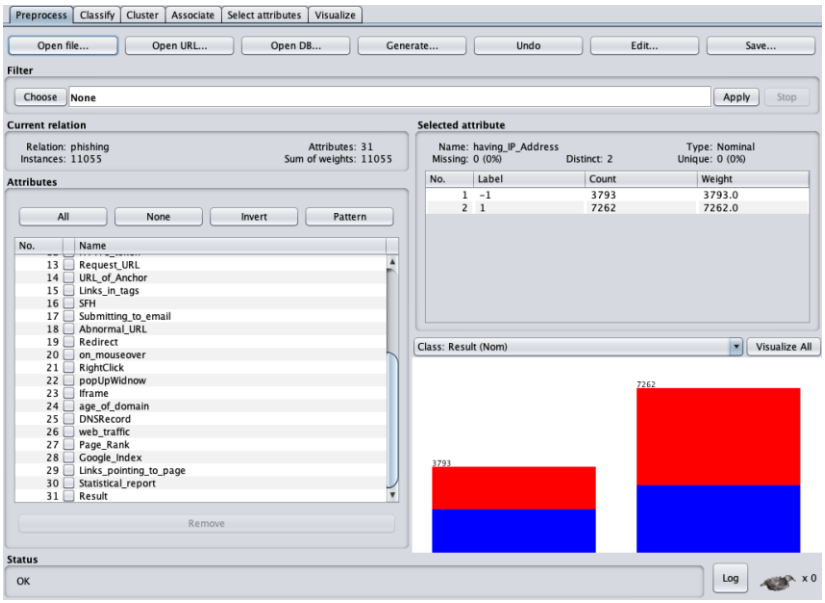
How to setup Anaconda environment and install packages:

1. Install Anaconda: http://docs.anaconda.com/anaconda/install.html
2. Create myidsenv environment (conda create --name myidsenv)
3. Activate myidsenv environment (conda activate myidsenv)
4. Install SkLearn package (pip install sklearn)
5. Similarly install "pandas"

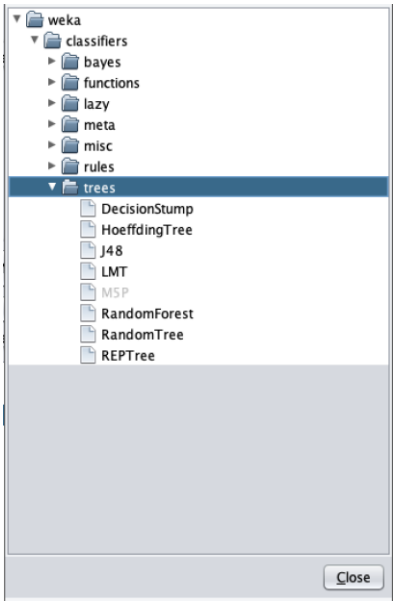
Lab Files that are Needed: For this lab you will need two files (phishing_l.csv and phishing.csv) the last column is the class value, others are the features.

LAB EXERCISE 1

- Import data into WEKA (explorer), the files of type should be specified (csv).



- Choose a proper classifier, such as RandomForest



- Specify the test option and the column of class
- Try different classifiers (at least 5) of different types (e.g., trees, functions, bayes, etc.) and log their performance (time to build model, performance metrics, confusion matrix).

LAB EXERCISE 2

- In this exercise, you need to implement several classifiers with the use of sklearn.
- You are provided with the code which you need to modify and run ("phishing_sklearn.py").

- Change the ratio between train/test datasets and analyze how it influences the performance of the phishing detectors.
- Add the code that calculates and prints the statistics metrics such as accuracy, recall, precision and f1 score.

LAB EXERCISE 3

- Use the same data you use in the exercise 1 and 2.
- To install tensorflow – "pip install tensorflow --user"
- Similarly install "matplotlib"
- In this exercise, you will implement an artificial neural network classifier based on Tensorflow
- The code is provided "phishing_tf.py"
- Define the learning rate, number of epochs, and the batch size for the artificial neural network
- Please print the statistics metrics such as accuracy, recall, precision and f1 score.
- Try various NN training parameters such as number of epochs, learning rate, and batch size. Document your observations.

WHAT TO SUBMIT

You should submit a lab report file which include the steps you preprocessed data, the necessary code snippet of your classifier and architecture. Also, the screenshot for both your code snippet and the result are needed.

Analyze your results: differences in performance and you thoughts on why they are different, which phishing detector is better and why (remember there may be many "correct" answers that is why it is important that you elaborate your thoughts).

You can call your file "Lab1_phishing_yourname.pdf".