Configuration Manual

MSc Research Project
Cybersecurity

Reuel Mushayakarara
Student ID: 22244611

School of Computing
National College of Ireland

Supervisor: Michael Prior

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Reuel Tafara Mushayakarara		
Student ID:	22244611		
Programme:	MSc Cybersecurity	**Year:**	Sep 2023
Module:	Practicum 2		
Lecturer:	Michael Prior		
Submission Due Date:	12 Aug 2024		
Project Title:	Evaluating Machine Learning Models for Effective Phishing URL Detection (Configuration Manual)		

Word Count: ..... **Page Count:** .....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Reuel Tafara Mushayakarara

Date: 12 Aug 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Reuel Mushayakarara
Student ID: 22244611

1 Research Lab

The research was conducted on a personal laptop computer with the below specification details

- System Manufacture and Model - Dell Precision 5540
- Operating System - Microsoft Windows 11 Pro
- Processor - Intel(R) Core (TM) i7-9850H CPU @ 2.60GHz, 2592 Mhz, 6 Cores, 12 Logical Processors
- RAM – 32GB
- Storage – 500GB SSD

2 Application

Anaconda Distribution Software was installed on the laptop. It is an open-source distribution of the Python programming language for data science (Anaconda 2024).

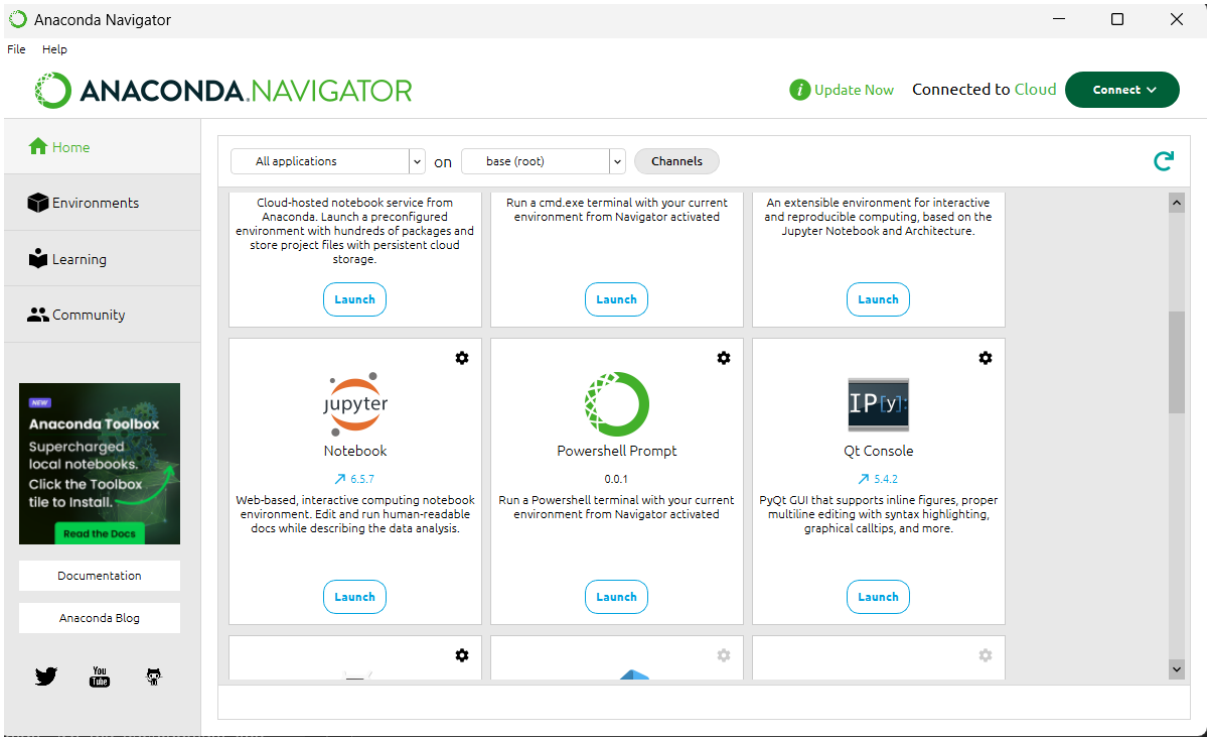

Fig 1

It includes

- Conda – a package manager for installing, updating and managing Python libraries and dependencies.
- Anaconda Navigator – a GUI desktop application that launches other development applications from the managed environment.
- Jupyter Notebook – it's a web-based, interactive computing notebook environment where one can edit and execute code written in python (Jupyter 2024). All the development and execution of the source code of this research was carried out on this application.

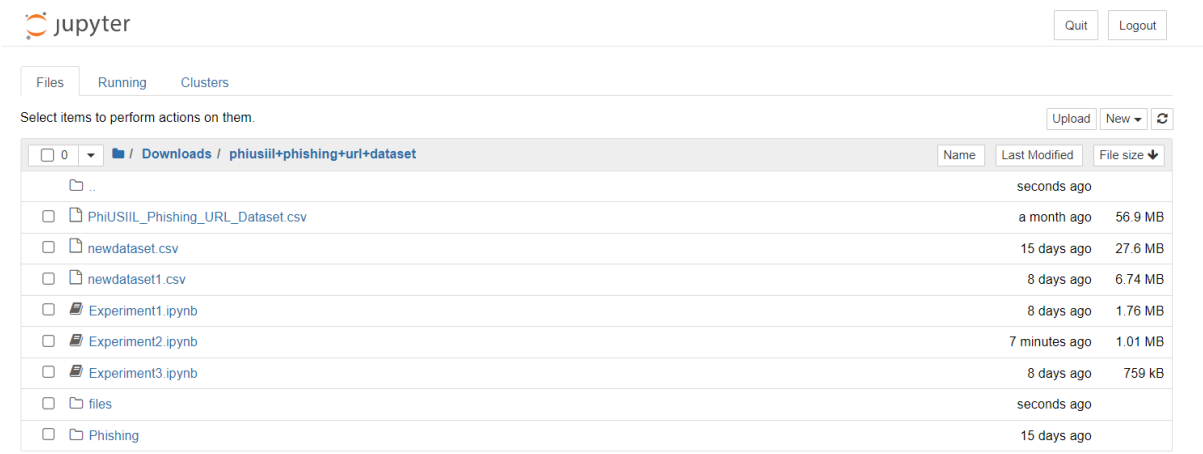

Fig 2

3 Install Guide

Anaconda Installation (Anaconda 2 2024)

1. Download the Anaconda installer.
2. Go to your Downloads folder and double-click the installer to launch. Note
3. Click **Next**.
4. Read the licensing terms and click **I Agree**.
5. It is recommended that you install for **Just Me**, which will install Anaconda Distribution to just the current user account. Only select an install for **All Users** if you need to install for all users' accounts on the computer (which requires Windows Administrator privileges).
6. Click **Next**.
7. Select a destination folder to install Anaconda and click **Next**
8. Choose whether to add Anaconda to your PATH environment variable or register Anaconda as your default Python. We **don't recommend** adding Anaconda to your PATH environment variable, since this can interfere with other software. Unless you plan on installing and running multiple versions of Anaconda or multiple versions of Python, accept the default and leave this box checked. Instead, use Anaconda software by opening Anaconda Navigator or the Anaconda Prompt from the Start Menu.
9. Click **Install**. If you want to watch the packages Anaconda is installing, click Show Details.
10. Click **Next**.
11. Optional: To learn more about Anaconda's cloud notebook service, go to https://www.anaconda.com/code-in-the-cloud. or click **Continue** to proceed.

12. After a successful installation you will see the “Thanks for installing Anaconda” dialog box:
13. If you wish to read more about Anaconda.org and how to get started with Anaconda, check the boxes “Anaconda Distribution Tutorial” and “Learn more about Anaconda”. Click the **Finish** button.

Web URL Detector

Navigate to the respective experiment folder and first run the retrain_model.py and thereafter run the app.py (Bichave 2022)

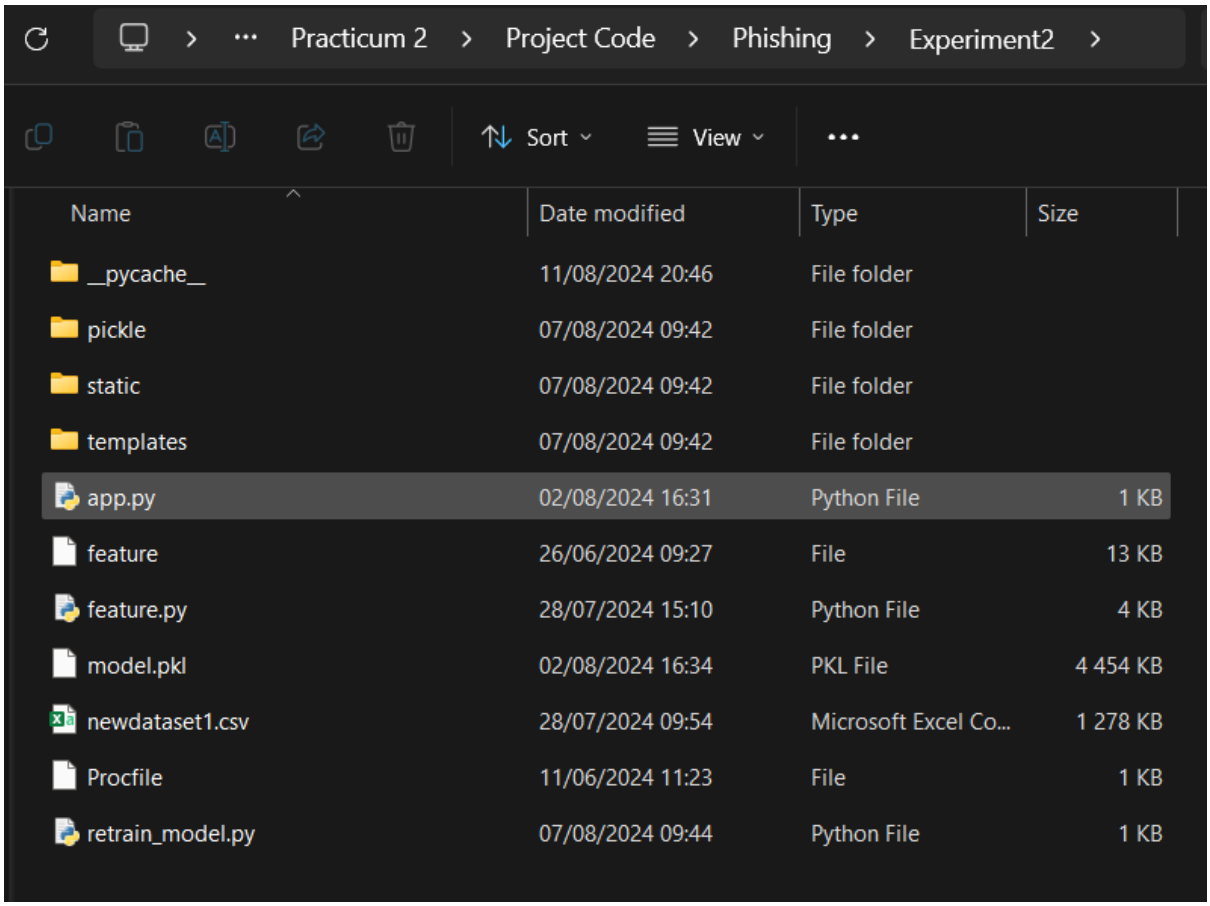

Fig 3

4 Execution of Code

Open Anaconda Navigator

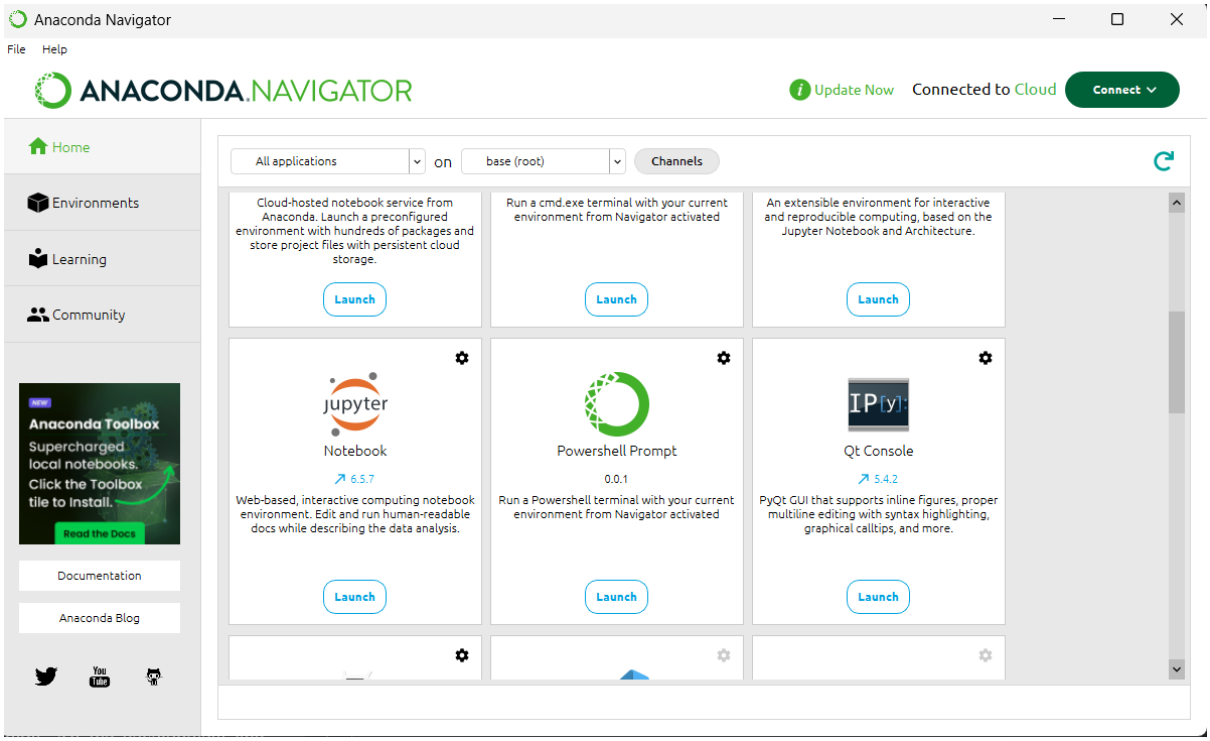

Fig 4

Open Jupyter Notebook and navigate to the location path containing the project files

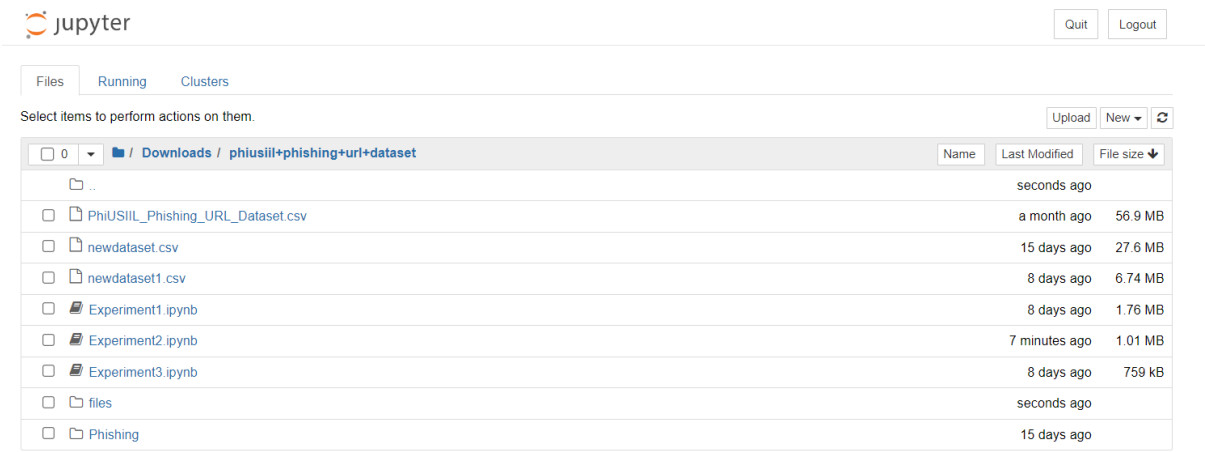

Fig5

Open the respective files for each experiment
Experiment 1

jupyter Experiment1 Last Checkpoint: 08/03/2024 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

```
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor, AdaBoostRegressor, ExtraTreesRegressor
from xgboost import XGBRegressor
from sklearn.ensemble import VotingRegressor, StackingRegressor
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, InputLayer, Input
from tensorflow.keras.utils import plot_model
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.metrics import accuracy_score, classification_report
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
```

In [2]:

```
# Load dataset
data = pd.read_csv('PhiUSIIL_Phishing_URL_Dataset.csv')
```

In [3]:

```
data.head()
```

Out[3]:

	FILENAME	URL	URLLength	Domain	DomainLength	IsDomainIP	TLD	URLSimilarityIndex	CharContinuationRa
0	521848.txt	https://www.southbankmosaics.com	31	www.southbankmosaics.com	24	0	com	100.0	1.0000
1	31372.txt	https://www.uni-mainz.de	23	www.uni-mainz.de	16	0	de	100.0	0.6666
2	597387.txt	https://www.voicefmradio.co.uk	29	www.voicefmradio.co.uk	22	0	uk	100.0	0.8666
3	554095.txt	https://www.sfnmjournal.com	26	www.sfnmjournal.com	19	0	com	100.0	1.0000
4	151578.txt	https://www.rewildingargentina.org	33	www.rewildingargentina.org	26	0	org	100.0	1.0000

5 rows × 56 columns

In [4]:

```
data.shape
```

Out[4]: (235795, 56)

In [5]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 235795 entries, 0 to 235794
```

Fig 6

Experiment 2

jupyter Experiment2 Last Checkpoint: 08/03/2024 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

```
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
```

In [2]:
```
newdf=pd.read_csv('PhiUSIIL_Phishing_URL_Dataset.csv')
```

In [5]:
```
df.head()
```

Out[5]:

	FILENAME	URL	URLLength	Domain	DomainLength	IsDomainIP	TLD	URLSimilarityIndex	CharContinuationRa
0	521848.txt	https://www.southbankmosaics.com	31	www.southbankmosaics.com	24	0	com	100.0	1.0000
1	31372.txt	https://www.uni-mainz.de	23	www.uni-mainz.de	16	0	de	100.0	0.6666
2	597387.txt	https://www.voicefmradio.co.uk	29	www.voicefmradio.co.uk	22	0	uk	100.0	0.8666
3	554095.txt	https://www.sfnmjournal.com	26	www.sfnmjournal.com	19	0	com	100.0	1.0000
4	151578.txt	https://www.rewildingargentina.org	33	www.rewildingargentina.org	26	0	org	100.0	1.0000

5 rows × 56 columns

In [6]:
```
df.columns
```

Out[6]:
```
Index(['FILENAME', 'URL', 'URLLength', 'Domain', 'DomainLength', 'IsDomainIP',
       'TLD', 'URLSimilarityIndex', 'CharContinuationRate',
       'TLDLegitimateProb', 'URLCharProb', 'TLDLength', 'NoOfSubDomain',
       'HasObfuscation', 'NoOfObfuscatedChar', 'ObfuscationRatio',
       'NoOfLettersInURL', 'LetterRatioInURL', 'NoOfDegitsInURL',
       'DegitRatioInURL', 'NoOfEqualsInURL', 'NoOfQMarkInURL',
       'NoOfAmpersandInURL', 'NoOfOtherSpecialCharsInURL',
       'SpacialCharRatioInURL', 'IsHTTPS', 'LineOfCode', 'LargestLineLength',
       'HasTitle', 'Title', 'DomainTitleMatchScore', 'URLTitleMatchScore',
       'HasFavicon', 'Robots', 'IsResponsive', 'NoOfURLRedirect',
       'NoOfSelfRedirect', 'HasDescription', 'NoOfPopup', 'NoOfiFrame',
       'HasExternalFormSubmit', 'HasSocialNet', 'HasSubmitButton',
       'HasHiddenFields', 'HasPasswordField', 'Bank', 'Pay', 'Crypto',
       'HasCopyrightInfo', 'NoOfImage', 'NoOfCSS', 'NoOfJS', 'NoOfSelfRef',
       'NoOfEmptyRef', 'NoOfExternalRef', 'label'],
      dtype='object')
```

In [3]:
```
df= newdf.drop(columns = ['FILENAME','URLLength', 'Domain', 'DomainLength', 'IsDomainIP',
       'TLD', 'URLSimilarityIndex', 'CharContinuationRate',
       'TLDLegitimateProb', 'URLCharProb', 'TLDLength', 'NoOfSubDomain',
       'HasObfuscation', 'NoOfObfuscatedChar', 'ObfuscationRatio',
       'NoOfLettersInURL', 'LetterRatioInURL', 'NoOfDegitsInURL',
       'DegitRatioInURL', 'NoOfEqualsInURL', 'NoOfQMarkInURL',
       'NoOfAmpersandInURL', 'NoOfOtherSpecialCharsInURL',
       'SpacialCharRatioInURL', 'IsHTTPS', 'LineOfCode', 'LargestLineLength',
       'HasTitle', 'Title', 'DomainTitleMatchScore', 'URLTitleMatchScore',
       'HasFavicon', 'Robots', 'IsResponsive', 'NoOfURLRedirect',
       'NoOfSelfRedirect', 'HasDescription', 'NoOfPopup', 'NoOfiFrame',
       'HasExternalFormSubmit', 'HasSocialNet', 'HasSubmitButton',
       'HasHiddenFields', 'HasPasswordField', 'Bank', 'Pay', 'Crypto',
       'HasCopyrightInfo', 'NoOfImage', 'NoOfCSS', 'NoOfJS', 'NoOfSelfRef',
       'NoOfEmptyRef', 'NoOfExternalRef'])
```

In [4]:
```
df.head()
```

Out[4]:

	URL	label
0	https://www.southbankmosaics.com	1
1	https://www.uni-mainz.de	1
2	https://www.voicefmradio.co.uk	1
3	https://www.sfnmjournal.com	1

Fig 7

Experiment 3

jupyter Experiment3 Last Checkpoint: 08/03/2024 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

```
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
```

In [2]:
```
newdf=pd.read_csv('PhiUSIIL_Phishing_URL_Dataset.csv')
```

In [3]:
```
df= newdf.drop(columns = ['FILENAME','URLLength', 'Domain', 'DomainLength', 'IsDomainIP',
        'TLD', 'URLSimilarityIndex', 'CharContinuationRate',
        'TLDLegitimateProb', 'URLCharProb', 'TLDLength', 'NoOfSubDomain',
        'HasObfuscation', 'NoOfObfuscatedChar', 'ObfuscationRatio',
        'NoOfLettersInURL', 'LetterRatioInURL', 'NoOfDegitsInURL',
        'DegitRatioInURL', 'NoOfEqualsInURL', 'NoOfQMarkInURL',
        'NoOfAmpersandInURL', 'NoOfOtherSpecialCharsInURL',
        'SpacialCharRatioInURL', 'IsHTTPS', 'LineOfCode', 'LargestLineLength',
        'HasTitle', 'Title', 'DomainTitleMatchScore', 'URLTitleMatchScore',
        'HasFavicon', 'Robots', 'IsResponsive', 'NoOfURLRedirect',
        'NoOfSelfRedirect', 'HasDescription', 'NoOfPopup', 'NoOfiFrame',
        'HasExternalFormSubmit', 'HasSocialNet', 'HasSubmitButton',
        'HasHiddenFields', 'HasPasswordField', 'Bank', 'Pay', 'Crypto',
        'HasCopyrightInfo', 'NoOfImage', 'NoOfCSS', 'NoOfJS', 'NoOfSelfRef',
        'NoOfEmptyRef', 'NoOfExternalRef'])
```

In [4]:
```
df.rename(columns= {'URL': 'url'}, inplace=True)
```

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 235795 entries, 0 to 235794
Data columns (total 2 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   url     235795 non-null  object
 1   label   235795 non-null  int64
dtypes: int64(1), object(1)
memory usage: 3.6+ MB
```

In [6]:
```
import pandas as pd
import ipaddress
import re
import urllib.request
from bs4 import BeautifulSoup
import socket
import requests
#from googlesearch import search
import whois
from datetime import date, datetime
import time
from dateutil.parser import parse as date_parse
from urllib.parse import urlparse
from bs4 import BeautifulSoup, XMLParsedAsHTMLWarning
import warnings

# Load the uploaded dataset
#df = newdf
warnings.filterwarnings("ignore", category=XMLParsedAsHTMLWarning)


class FeatureExtraction:
    def __init__(self, url):
        self.features = []
        self.url = url
        self.domain = ""
```

Fig 8

Web URL Detector

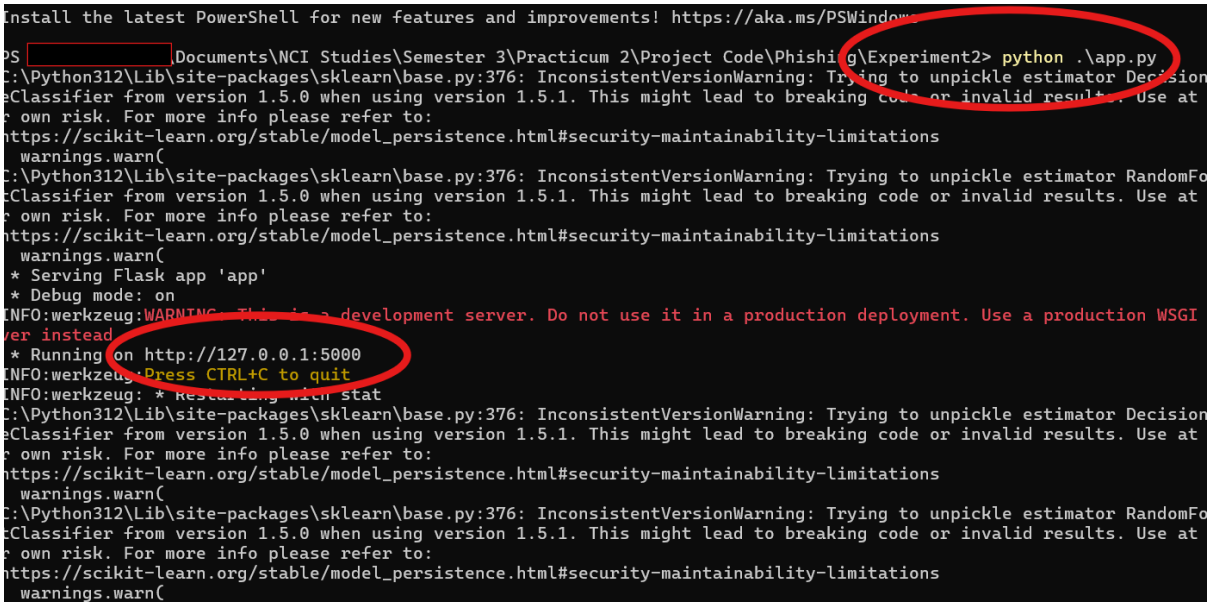
Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindow
PS [redacted]\Documents\NCI Studies\Semester 3\Practicum 2\Project Code\Phishi g\Experiment2> python .\app.py
:\Python312\Lib\site-packages\sklearn\base.py:376: InconsistentVersionWarning: Tr ing to unpickle estimator Dec sion
eClassifier from version 1.5.0 when using version 1.5.1. This might lead to breaking c or invalid resul . Use at
own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(
:\Python312\Lib\site-packages\sklearn\base.py:376: InconsistentVersionWarning: Trying to unpickle estimator RandomFo
Classifier from version 1.5.0 when using version 1.5.1. This might lead to breaking code or invalid results. Use at
own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(
* Serving Flask app 'app'
* Debug mode: on
INFO:werkzeug:WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI
ver instead.
* Running on http://127.0.0.1:5000
INFO:werkzeug:Press CTRL+C to quit
INFO:werkzeug: * Restarting with stat
:\Python312\Lib\site-packages\sklearn\base.py:376: InconsistentVersionWarning: Trying to unpickle estimator Decision
eClassifier from version 1.5.0 when using version 1.5.1. This might lead to breaking code or invalid results. Use at
own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(
:\Python312\Lib\site-packages\sklearn\base.py:376: InconsistentVersionWarning: Trying to unpickle estimator RandomFo
Classifier from version 1.5.0 when using version 1.5.1. This might lead to breaking code or invalid results. Use at
own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(

Fig 9

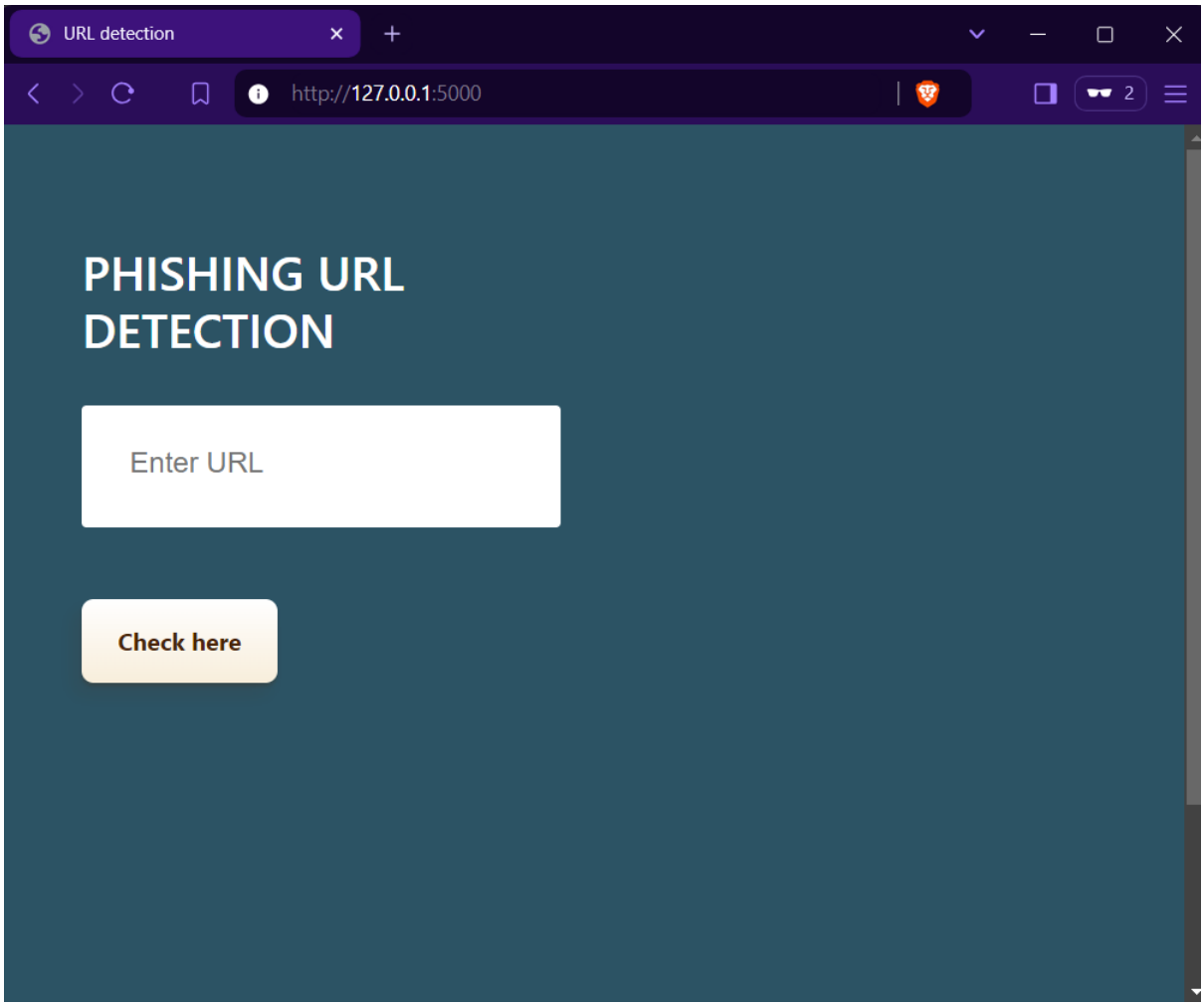
URL detection
http://127.0.0.1:5000
PHISHING URL DETECTION
Enter URL
Check here

Fig 10

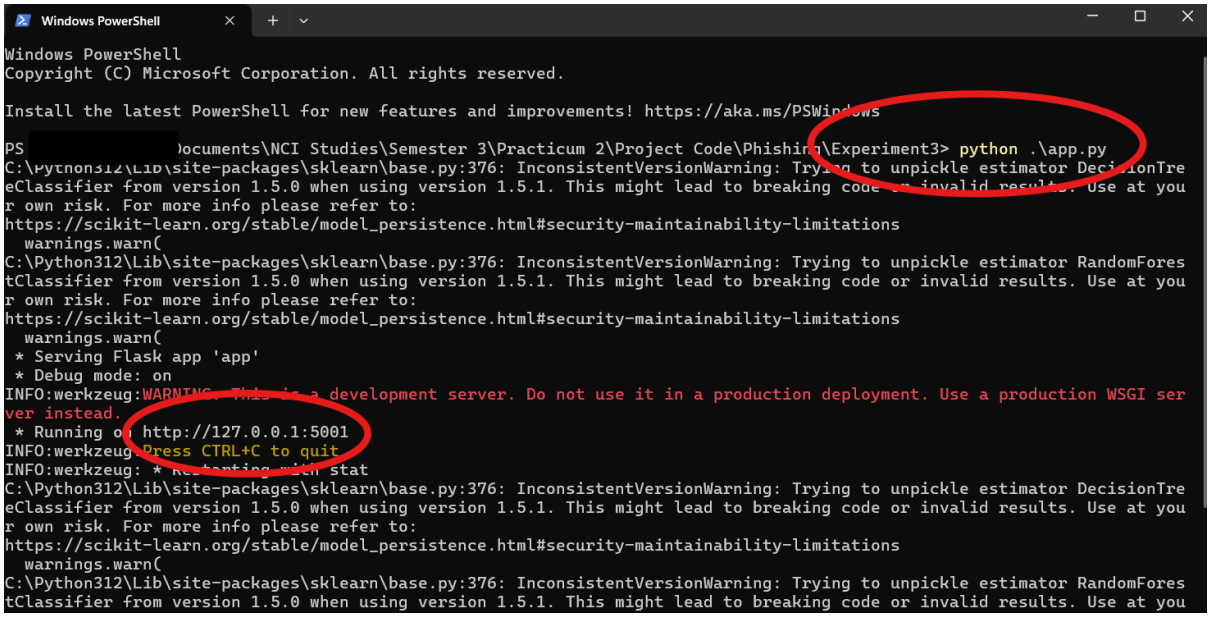
Windows PowerShell
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.
Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows
PS [redacted]ocuments\NCI Studies\Semester 3\Practicum 2\Project Code\Phishing\Experiment3> python .\app.py
C:\Python312\Lib\site-packages\sklearn\base.py:376: InconsistentVersionWarning: Trying to unpickle estimator DecisionTreeClassifier from version 1.5.0 when using version 1.5.1. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(
C:\Python312\Lib\site-packages\sklearn\base.py:376: InconsistentVersionWarning: Trying to unpickle estimator RandomForestClassifier from version 1.5.0 when using version 1.5.1. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(
* Serving Flask app 'app'
* Debug mode: on
INFO:werkzeug:WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on http://127.0.0.1:5001
INFO:werkzeug:Press CTRL+C to quit
INFO:werkzeug: * Restarting with stat
C:\Python312\Lib\site-packages\sklearn\base.py:376: InconsistentVersionWarning: Trying to unpickle estimator DecisionTreeClassifier from version 1.5.0 when using version 1.5.1. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(
C:\Python312\Lib\site-packages\sklearn\base.py:376: InconsistentVersionWarning: Trying to unpickle estimator RandomForestClassifier from version 1.5.0 when using version 1.5.1. This might lead to breaking code or invalid results. Use at you

Fig 11

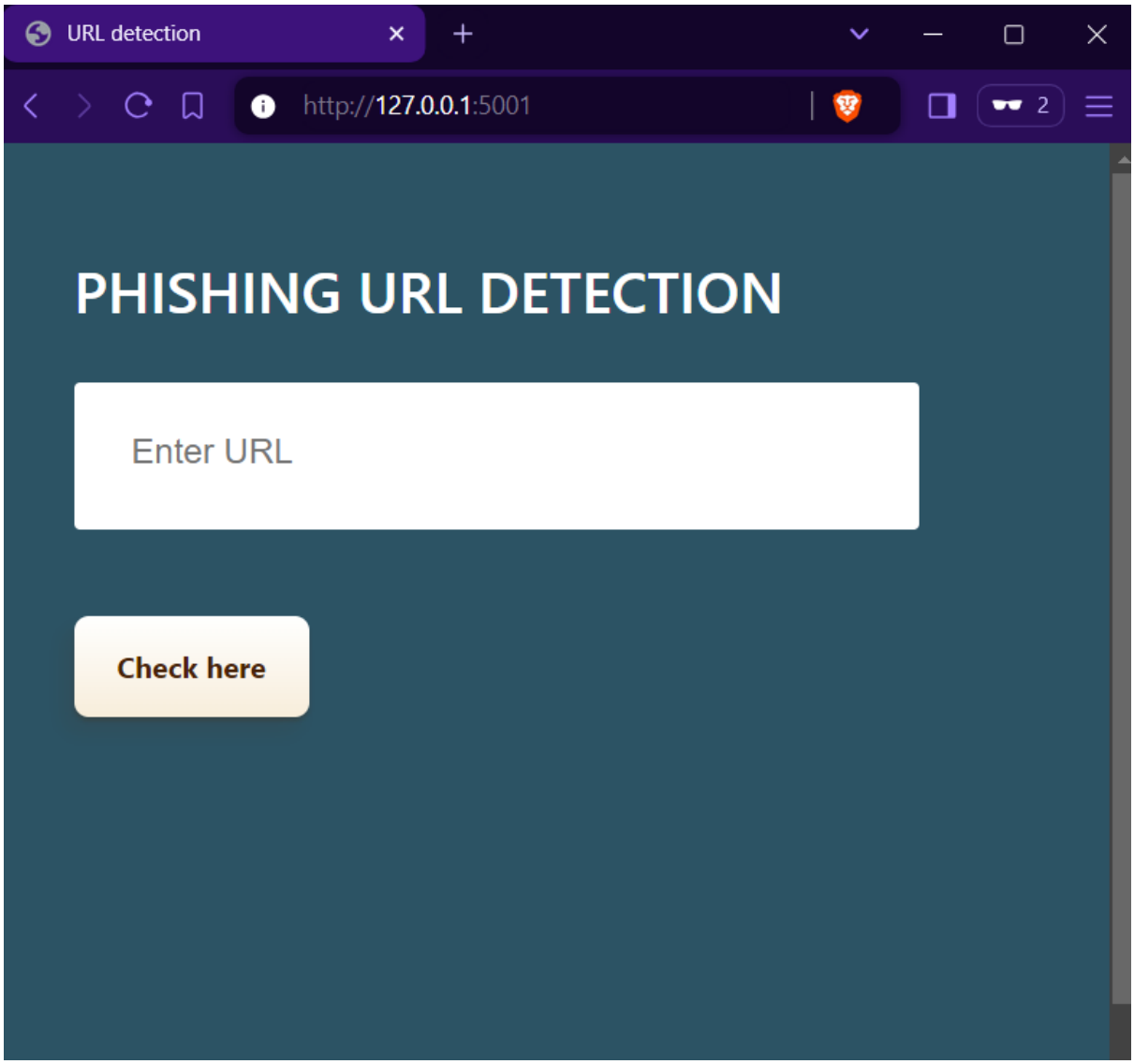
URL detection
http://127.0.0.1:5001
PHISHING URL DETECTION
Enter URL
Check here

Fig 12

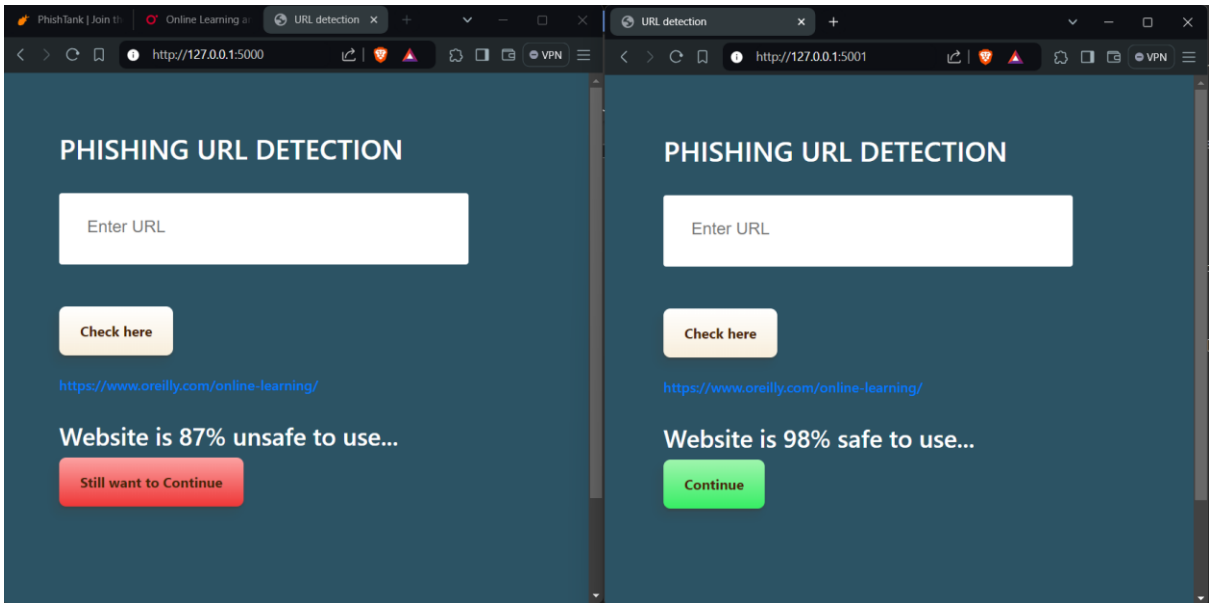

Fig 13

References

Anaconda (2024) Anaconda Documentation. Available at: https://docs.anaconda.com/ [Accessed 1.8.24].

Anaconda 2 (2024) Installing on Windows. Available at: https://docs.anaconda.com/anaconda/install/windows/ [Accessed 1.8.24].

Jupyter (2024) Jupyter. Available at: https://jupyter.org/ [Accessed 1.8.24].

V. Bichave (2022) Phishing URL Detection. Available at: https://github.com/vaibhavbichave/Phishing-URL-Detection [Accessed 1.8.24].