# Configuration Manual

MSc Internship
Programme Name

Michael Oluwasegun Akinrele
Student ID: X18109489

School of Computing
National College of Ireland

Supervisor: Mr Vikas Sahni

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | |
|---|---|
| **Student Name:** | **Michael Oluwasegun Akinrele** |
| **Student ID:** | **X18109489** |
| **Programme:** | **CyberSecurity** **Year: 2019** |
| **Module:** | **MSc Internship** |
| **Lecturer:** | **Mr Vikas Sahni** |
| **Submission Due Date:** | **12/12/2019** |
| **Project Title:** | **Detection of Phishing and Spam Emails Using Ensemble Technique** |
| **Word Count:** | **324** **Page Count: 6** |

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.
I agree to an electronic copy of my thesis being made publicly available on NORMA the National College of Ireland's Institutional Repository for consultation.

**Signature:** ……

**Date:** ……

**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

MICHAEL OLUWASEGUN AKINRELE
X18109489

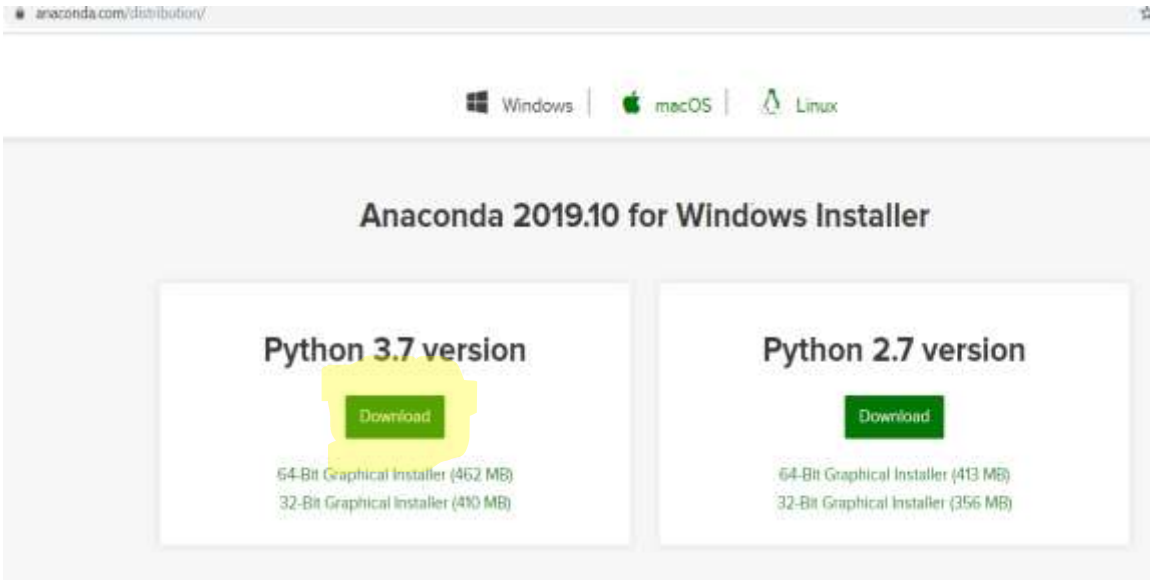
This configuration manual entails the tools capabilities and features used in the course of the project. It gives specific directions on the best replica of the experiment carried out.

The below steps shows how the installation is carried out on Windows Operating System.

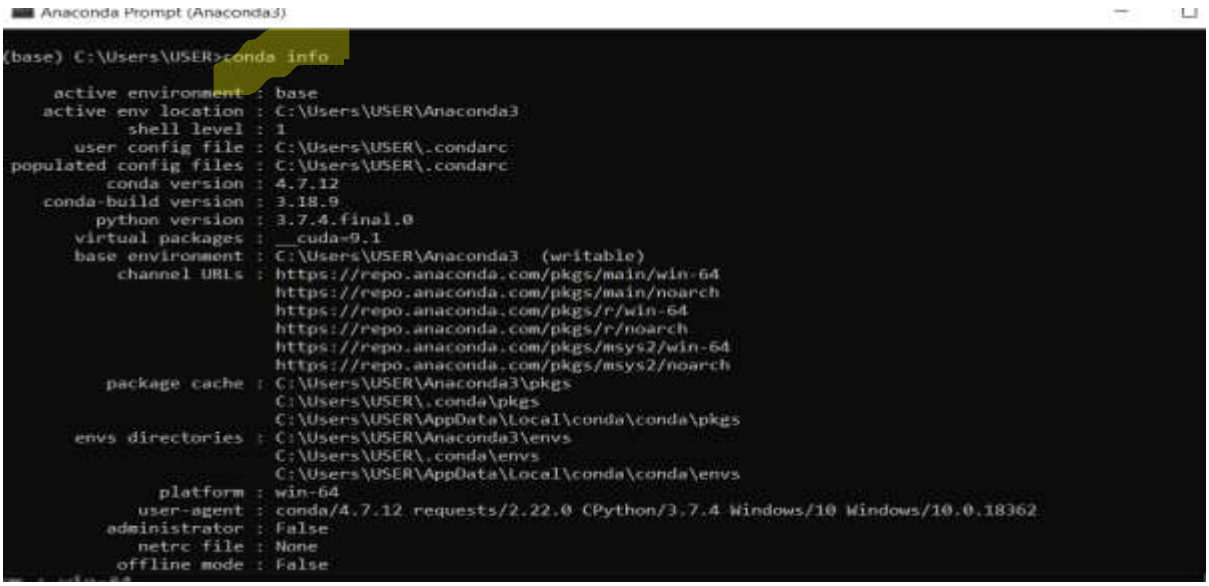
## 1 Installing Anaconda3

This section shows installation of Anaconda on an operating system of windows version 10. Anaconda3 is a designed package manager which is an open source for actualizing data analysis and machine learning project.

**Step 1:** Download the latest version of Anaconda installer with *https://anaconda.com/distribution*

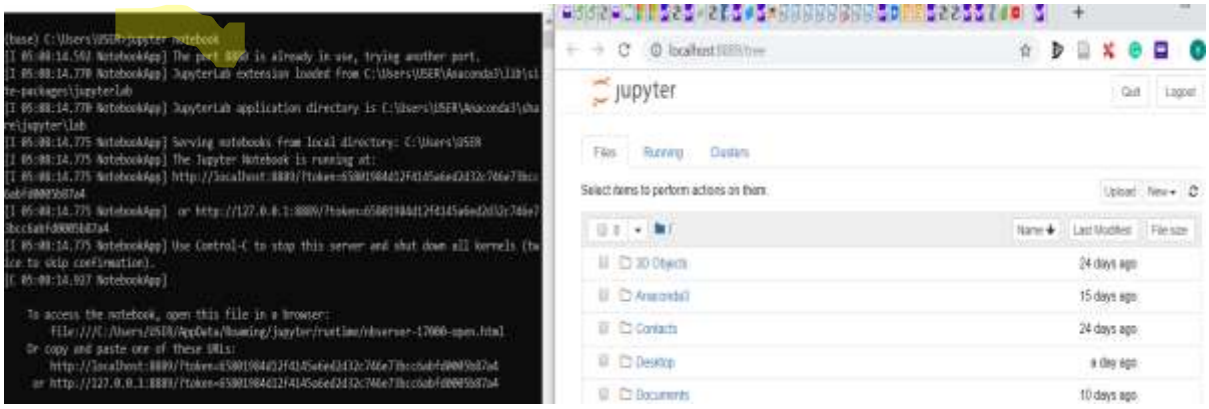


**Step 2:** After successful download, check the file integrity with the highlighted command to be sure that the download is right.




# 2 Jupyter Notebook IDE Installation using pip

**Step 1:** From the Anaconda prompt, run Jupyter notebook




# 3 Importing and Extraction of dataset

**Step 1:** Import necessary python libraries for dataset preprocessing

```
# Necessary Python Libraries
import pprint, os, re, email, csv
from bs4 import BeautifulSoup
from collections import Counter
from urllib.parse import urlparse
from IPy import IP
```

**Step 2:** Download dataset from the site as follows; (https://monkey.org/~jose/phishing/) and Spam Dataset (https://spamassassin.apache.org/old/publiccorpus/) and get path to spam, easy_ham, test_emails, and phishing from the Dataset imported.

```
spam_path = "C:\\Users\\USER\\Desktop\\phispam\\phish-spamming-detection\\datasets\\spam"
easy_ham_path = "C:\\Users\\USER\Desktop\\phispam\\phish-spamming-detection\\datasets\\easy_ham"
test_path = "C:\\Users\\USER\\Desktop\\phispam\\phish-spamming-detection\\datasets\\test_emails"
phishing_2017_path = "C:\\Users\\USER\Desktop\\phispam\\phish-spamming-detection\\datasets\\phishing\\phishing_2017"
phishing_2018_path = "C:\\Users\\USER\Desktop\\phispam\\phish-spamming-detection\\datasets\\phishing\\phishing_2018"
```

**Step 3:** Extracting all the 40 features from the dataset

```
def overall_feature_extraction(path, label, mail):
    necessary_fields = extract_necessary_fields(path, mail)
```

**Step 4:** Extract features of all mails in all paths and create a csv with the target attribute 'labels'

```
# Paths and corresponding labels
all_paths_labels = {spam_path : "Spam",
                    easy_ham_path : "Ham",
                    phishing_2017_path : "Phishing",
                    phishing_2018_path : "Phishing"}
pprint.pprint(all_paths_labels, width = 1)

# Datasets used in the paper
datasets = ["final_dataset_HamSpam", "final_dataset_HamPhishing", "final_dataset_HamSpamPhishing"]
```

**Step 5:** After successful feature extraction and dataset creation

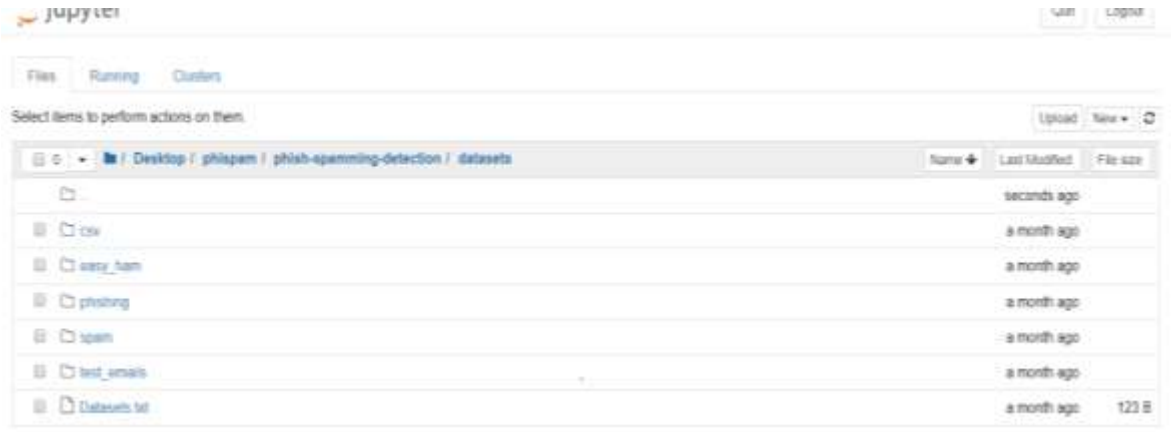
```
create_features_csv: success
extract_features_into_csv: success
dataset_creation: success
```

# 4 Analysis of dataset

**Step 1:** The below codes show the imported libraries in Jupyter notebook for analysis

```
In [1]: %matplotlib inline

import matplotlib.pyplot as plt
plt.style.use('ggplot')

import numpy as np
import os, math, sys
from collections import Counter
from IPython.display import display
import subprocess
import pandas as pd
from itertools import groupby
from operator import itemgetter
import timeit
```

**Step 2:** Insert easy_ham, phishing, spam and test_email files in the same folder, run the preprocessing files to download three csv files; which are test, final_dataset_HamPhishing, final_dataset_HamSpam, and final_dataset_HamSpamPhishing.




**Step 3:** Find below the view of our preprocessing and data analysis files in Jupyter notebook.

# 5 Read File and Sklearn Configuration

**Step 1:** Specify the directory that files are located and read **final_dataset_HamSpamPhishing** file only

```
csv_directory = "C:\\Users\\USER\\Desktop\\phispam\\phish-spamming-detection\\datasets\\csv\\"
files = ['final_dataset_HamPhishing', 'final_dataset_HamSpam', 'final_dataset_HamSpamPhishing']
```

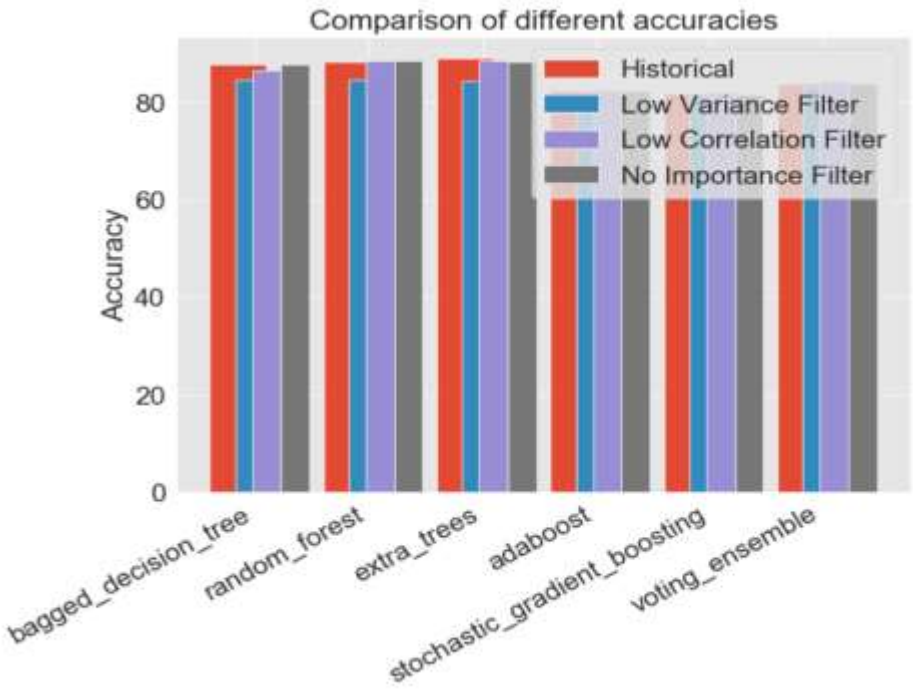
**Step 2:** Additional command to read head and tail of final dataset

```
global df
def get_data(filename):
    filename = csv_directory + filename + '.csv'
    df = pd.read_csv(filename)
    return df

filename = files[2]
df = get_data(filename)
display(df.head())
display(df.tail())
```

**Step 3:** Import all the algorithm from Sklearn

```
from sklearn import model_selection

from sklearn.ensemble import BaggingClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import ExtraTreesClassifier

from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier

from sklearn.svm import SVC
from sklearn.ensemble import VotingClassifier

from sklearn import metrics
from sklearn.naive_bayes import GaussianNB
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
```

**Step 4:** Knowledge chart of the classification model




# References

[1] https://docs.anaconda.com/anaconda-enterprise-4/ae-and-nav/

[2] https://scikit-learn.org/stable/install.html#installing-the-latest-release