Configuration Manual

MSc Internship

Cyber Security

Kiran Hassan Shivashankar

Student ID: 18184987

School of Computing

National College of Ireland

Supervisor: Prof. Ross Spelman

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Kiran Hassan Shivashankar
Student ID:	18184987
Programme:	MSc in Cybersecurity **Year:** 2019/2020
Module:	Academic Internship
Lecturer:	Prof. Ross Spelman
Submission Due Date:	17/08/2020
Project Title:	Accurate detection of malicious code in pdf files using Machine Learning Algorithms
Word Count:	320 **Page Count:** 2

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:	Kiran Hassan Shivashankar
Date:	17/08/2020

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

Kiran Hassan Shivashankar
Student ID: 18184987

1 Introduction

This document emphasises the system requirements and the configuration details for executing Malicious code detection in pdf files using Machine Learning Algorithm. The entire code has been written in Python programming language.

2 System Requirements

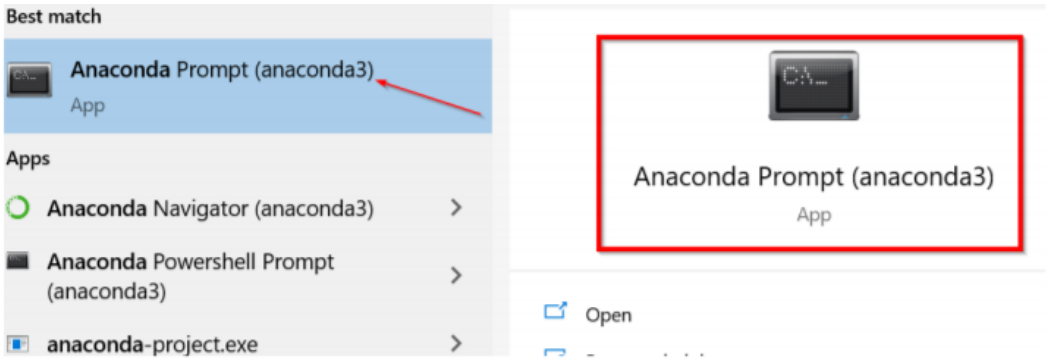
The user of this document requires to have basic python programming knowledge and python 2.7(which can be downloaded from python official website) installed on his system to execute this code. Anaconda Version 3.*.

2.1 Packages needed to install and run the code

- Pandas – Pandas is an open-source machine learning library used in python. It is a high performance easy to work with the library. Pandas is built on the Numpy library.
- Numpy – for array-based processing in python
- Matplotlib – is used for plotting the graphs
- sklearn - sklearn is a machine learning library build using Numpy and Matplotlib. It is used for data visualization and cross-validation. It is using in pre-processing steps on a dataset. Sklearn is a machine learning library that is composed of various regression, classification, and clustering-based algorithms.

2.2 How to install a package:

- To install a package first open Anaconda Prompt.



• Run the command – pip install <package name> E.g.: pip install pandas

3 Data Sources

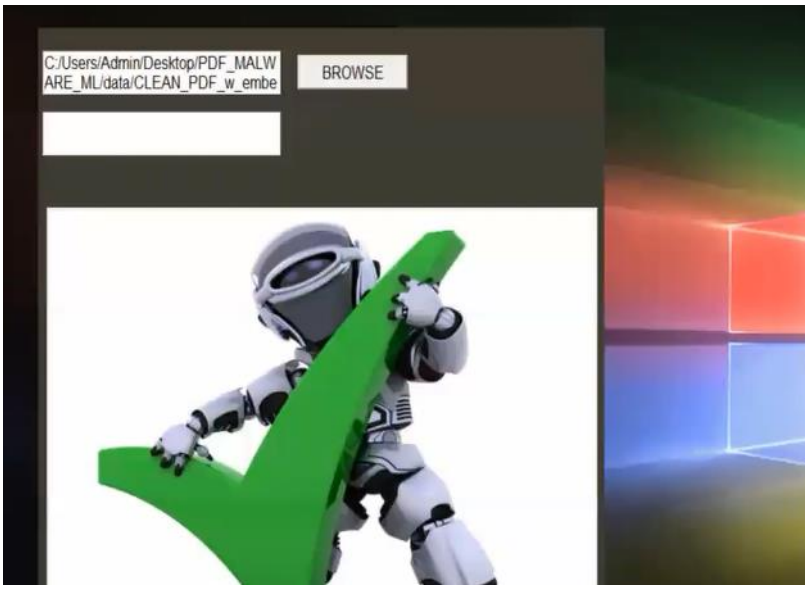
Data sets of 108 Benign pdf samples and 120 Malicious pdf documents are taken from contagio **malware dump**. You need to mail the author of the dataset for accessing the dataset to download that is password protected.

4 Code Execution

To execute the code, open Anaconda Navigator and create a virtual environment. Open Anaconda command prompt and activate the virtual environment created using **activate <environment_name>command**.

Once it is activated navigate to the folder where your code is stored using **cd** command. Next **run the python script create_dataset.py**. It will create the dataset in the data folder. Next **run the algo.py script**.

At the end you will get the window to check whether file you selected is malicious or not.


C:/Users/Admin/Desktop/PDF_MALW
ARE_ML/data/CLEAN_PDF_w_embe
BROWSE