Configuration Manual

MSc Internship
Cybersecurity

Shreyas Sudhir Barde
Student ID: x18198350

School of Computing
National College of Ireland

Supervisor: Niall Heffernan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Shreyas Sudhir Barde
Student ID:	x18198350
Programme:	MSc Cybersecurity **Year:** 2019-20
Module:	MSc Internship
Lecturer:	Niall Heffernan
Submission Due Date:	17/08/2020
Project Title:	Cross Site Scripting detection using Random Forest and Dataset Ensemble Modelling.
Word Count:	582 **Page Count:** 05

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: Shreyas Sudhir Barde

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

Shreyas Sudhir Barde
Student ID: x18198350

1 Introduction

1.1 Objective of the document

The document has been created to state the required software, setups, and settings for successful execution of this research project.

1.2 Content of the Document

Module	***Description***
Overview	This section of a report gives the information of general system and software required.
Execution Process	This section of a report gives the information of set up and execution of the project artefacts.

2 Overview

2.1 Objective

The basic goal of this project is to give a detection framework where the user can check the Input and the Machine Learning based model will detect the real world XSS scripts and payloads if these are inserted in benign scripts or inputs. The project is based on three main pillars. The first one is the Balanced Ensemble dataset, which gives a better result by avoiding the over-fitting problems with the models and give more exact accuracy. The second one is feature extraction, where we have used a combined approach of multiple past related works which helps to detect the real-world payloads and attacks more precisely. Third and last is the Random Forest Bagging algorithm of machine learning. The trained datasets for a balanced dataset are saved in two folders named ‘Cheat Sheet Data’ and 'Github Data' in the dataset folder in an artifact zip file. 'Kaggle data' from the same folder is an unbalanced data that we have used for comparison with the balanced dataset. The file ‘Final.ipynb’ from the same zipped folder contains the actual ML code for the train and test the detection of XSS script. Secondly, the ‘Imbalance Dataset Model.ipynb' contains the code for the application of this same model on the unbalanced dataset.

2.2 General system and software requirements

In the module, required software installations and dependencies for the execution of the program have been discussed.

Python 3.8.5 has been used while implementation of this project hence same version of Python is recommended for running and execution of this project and can be downloaded from the official website of python. Some additional python libraries, packages as well as modules are required for execution such as:

- numpy
- pandas
- sklearn

Moreover, Anaconda and Jupyter Notebook are used for the implementation and execution of this project. Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. Where Jupyter Notebook is a web application that allows us to create and share documents that contain live code visualizations such as Python code visualizations. explanatory text that is written in markdown syntax.

3 Execution Process

3.1 Code and Project Execution

Four Simple steps are involved in the Execution of this project and they are as follows:

1) Download, Install, and Open **Anaconda** Software.

2) Open **Jupyter Notebook** and Navigate to the specific folder where all files are extracted from the zip file.

3) Open the file '**Final.ipynb**' from the same unzipped folder to train and test the cross-site script detection model. At the end of the code, some evaluations and comparisons with the other algorithms is given in same file. (Shortcut Shift + Enter can be used for line by line execution.)

4) Open the file '**Imbalance Dataset Model.ipynb**' for running and obtaining the output of the same model on imbalanced dataset.

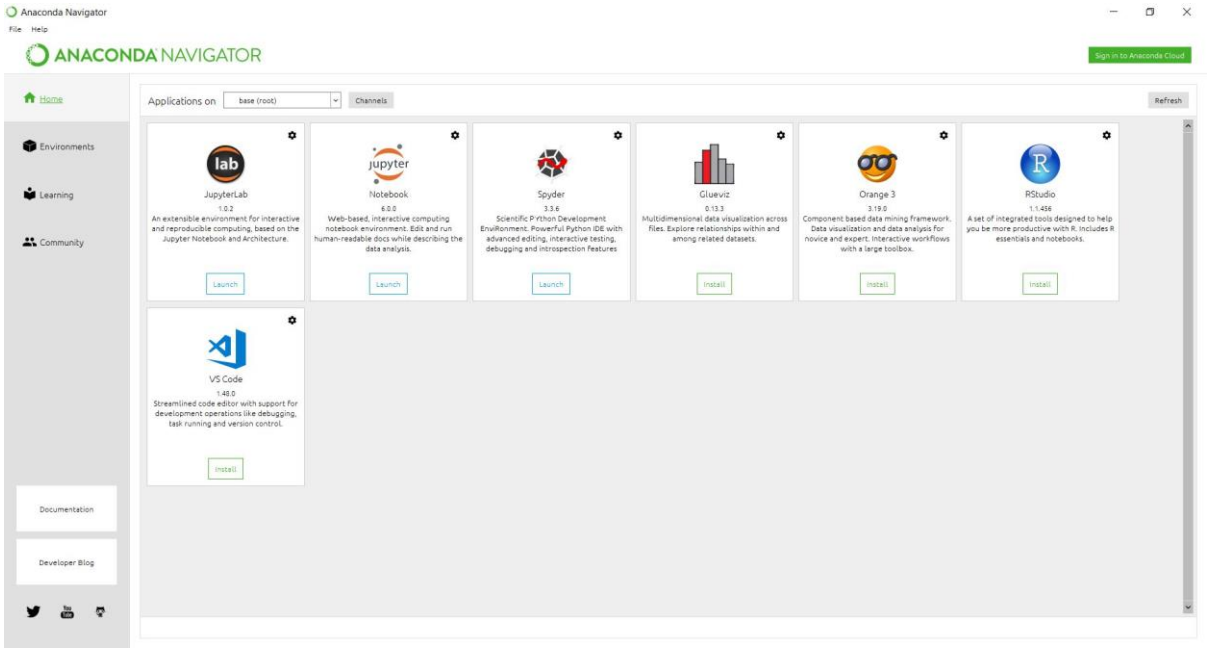

Figure1: ANACONDA NAVIGATOR

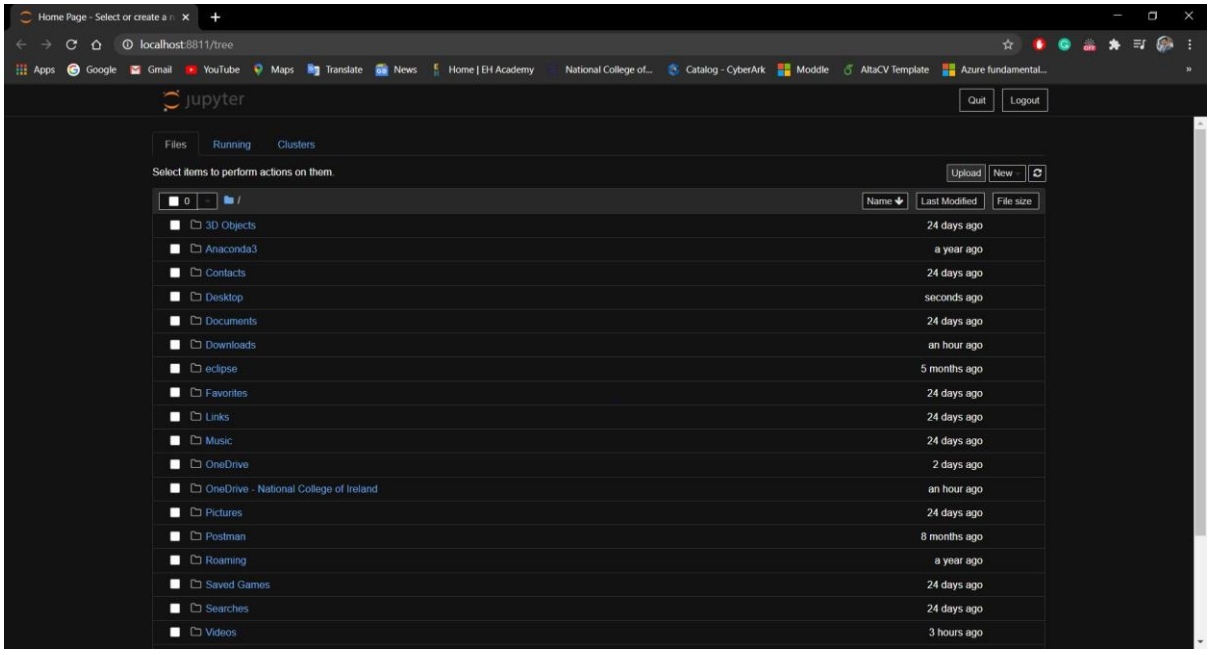

Figure2: Jupyter Notebook launch screen

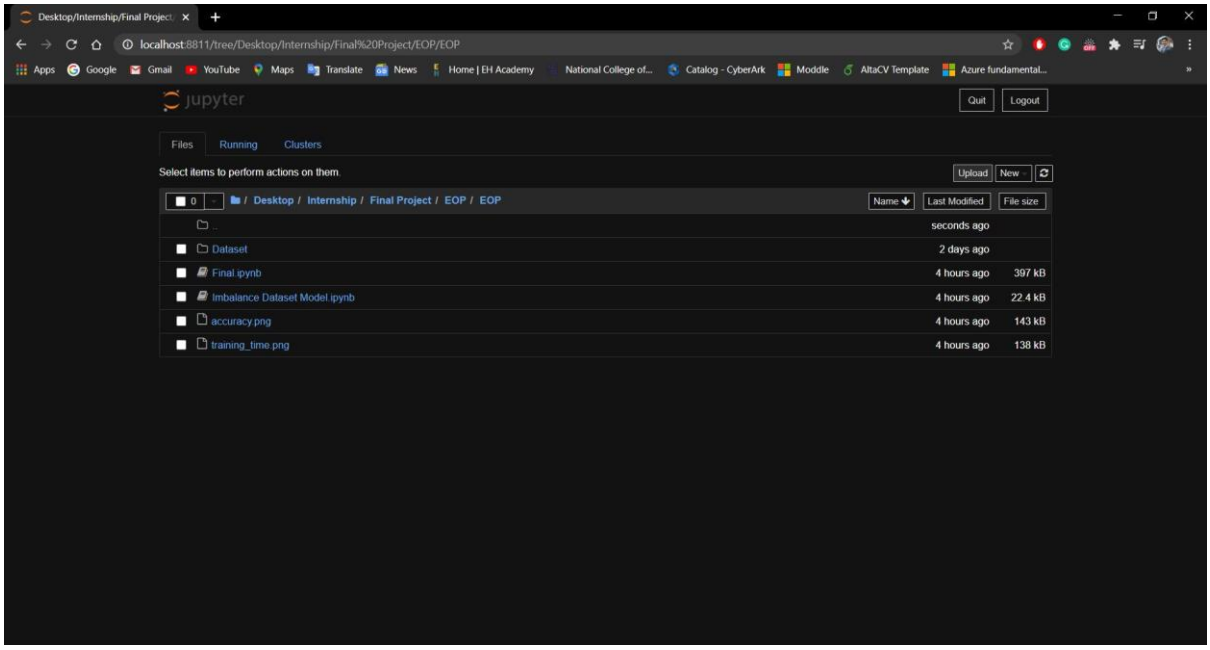

Figure3: Project Structure

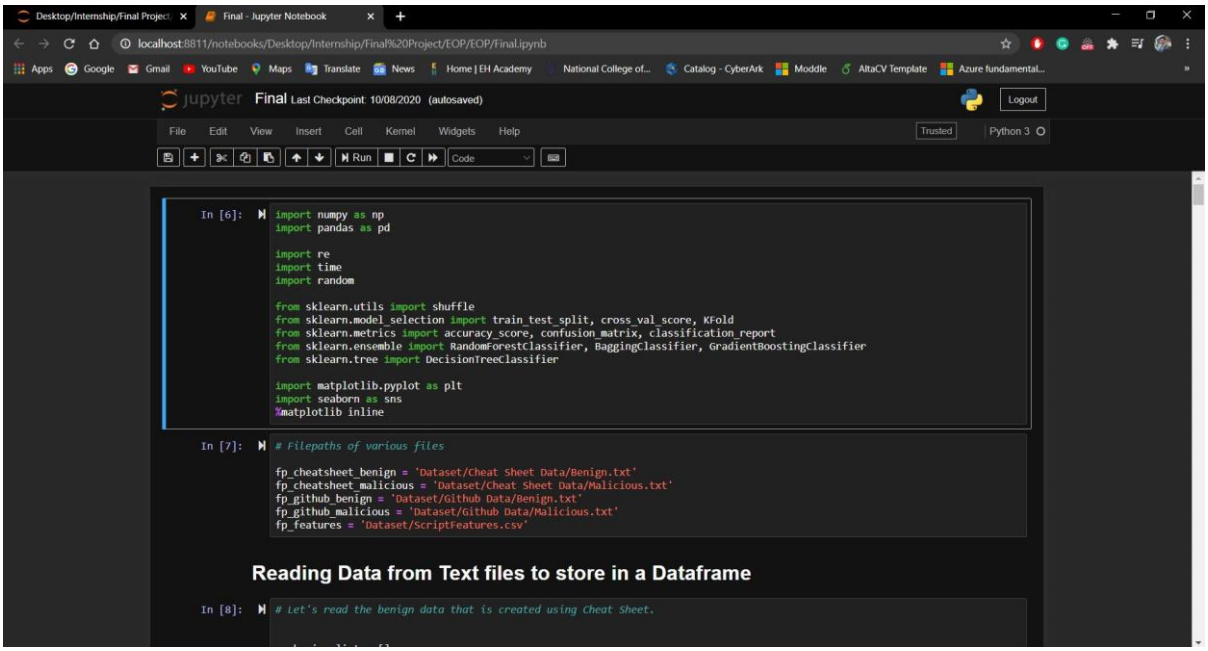

Figure4: Final project code snippet

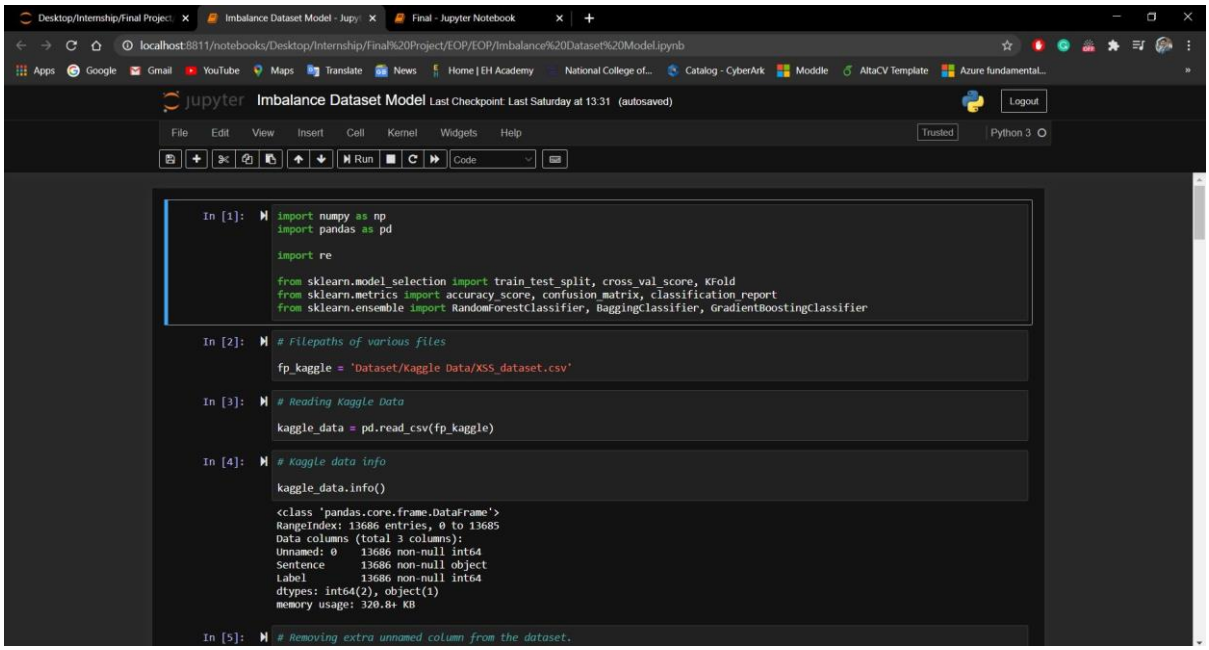

Figure5: Model with imbalance dataset for comparison with balanced dataset