National College of Ireland

Configuration Manual

MSc Research Project
Cybersecurity

Shruti Praveen Garg
Student ID: x23206047

School of Computing
National College of Ireland

Supervisor: Prof. Liam Mccabe

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Shruti Praveen Garg
Student ID:	X23206047
Programme:	MSc. Cybersecurity **Year:** 2024-2025
Module:	Practicum PT.2
Lecturer:	Prof. Liam Mccabe
Submission Due Date:	12th December 2024
Project Title:	Implementing Machine Learning Algorithms To Enhance Intrusion Detection Systems Across Computerised Networks Towards Pre-empting Cyber Attacks.
Word Count:	758 **Page Count:** 4

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: Shruti Praveen Garg

Date: 12th December 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

Shruti Praveen Garg
Student ID: x23206047

1 Introduction

This configuration manual provides setup instructions required to execute the implementation of enhancing IDS using machine learning techniques like Logistic Regression, Random Forest, Isolation Forest, One-Class SVM, and Voting Classifier on CIC-IDS 2017 dataset and implementing in Python language. This model performs classification and anomaly detection on network traffic data. The model uses Python 3.10.1, Visual Studio Code, and Jupyter for data preprocessing, model training and final evaluation.

2 System Requirements

2.1 Machine Info

- OS: Windows 11 Home Single Language
- System Type: 64-bit Operating System
- Processor: Intel Core i5 (4 cores)
- RAM: 8 GB
- GPU: NVIDIA GeForce MX450

3 Required Installations

The IDE used for the execution of this project was Visual Studio Code due to it's user friendly interface and support for all programming language, jupyter notebook, and much more. (*Visual Studio Code - Code Editing. Redefined*, no date)

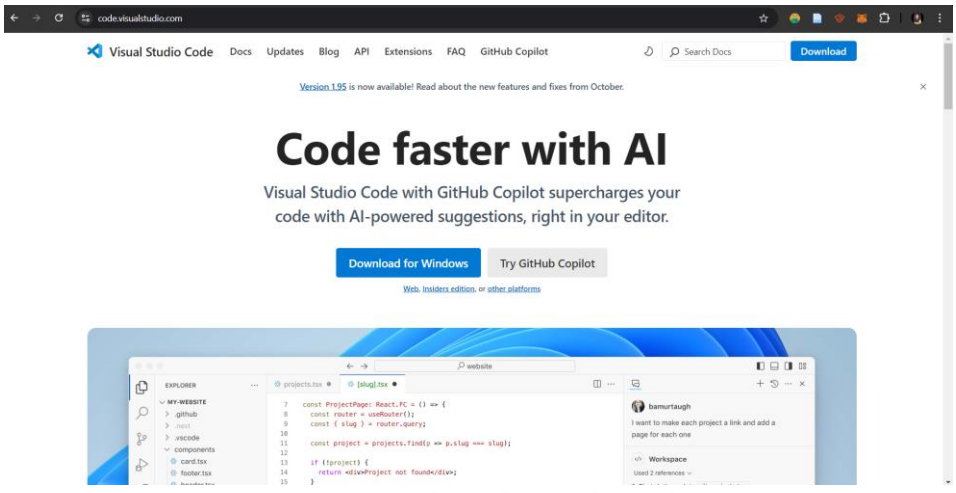

Image 1: Visual Studio Code Website

Since VSCode supports Python programming language which is used in this project, extension of Python was installed in the IDE itself. One can also install it from the official website. (https://www.python.org/downloads/) and import it in the IDE. (*Download Python*, no date)

Image 2: Python Extension

Jupyter notebook was also used for the execution of the python code as jupyter kernel supports to create, edit, run, plot a user friendly version of the notebook.

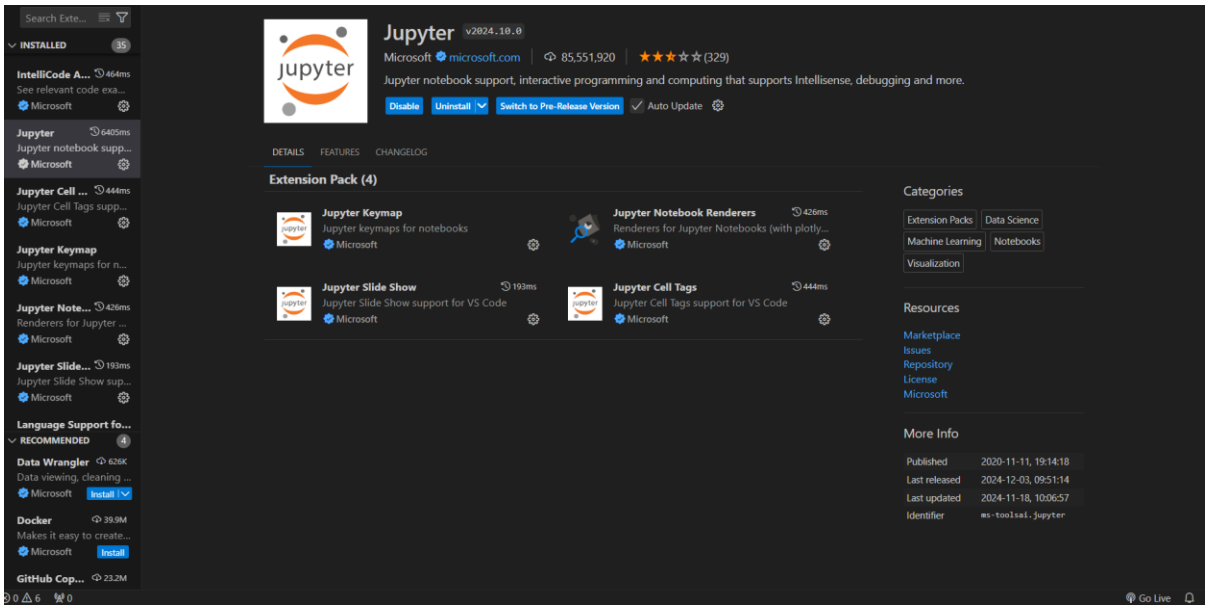

Image 3: Jupyter Extension

Certain python libraries are required to be installed for the project. The following re the commands to be used to install these libraries. To install these libraries, open terminal on upper left.

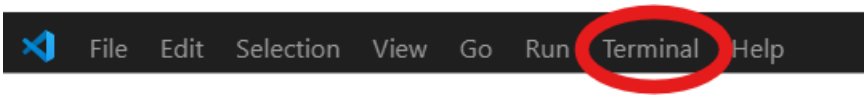

Image 4: Terminal

and run the following commands to install the libraries.

1. pip install numpy
2. pip install pandas
3. pip install scikit-learn
4. pip install matplotlib
5. pip install seaborn
6. pip install tensorflow

4 How to Run the Program

Post installing all the required libraries and dependencies;

1. Click on “File” and open a “New File” in the VSCode IDE.
2. Then select Jupyter Notebook and the file extension should be “.ipynb”.

Once the new file is open it will automatically be connected to python server.
The following python libraries are used in this project:

1. Pandas is used for data manipulation and handling.
2. NumPy is used for numerical operations.
3. Scikit-learn is used for machine learning models, feature selection, and evaluation.
4. Matplotlib and Seaborn is used for data visualization.
5. TensorFlow is used for deep learning integration for future practices.

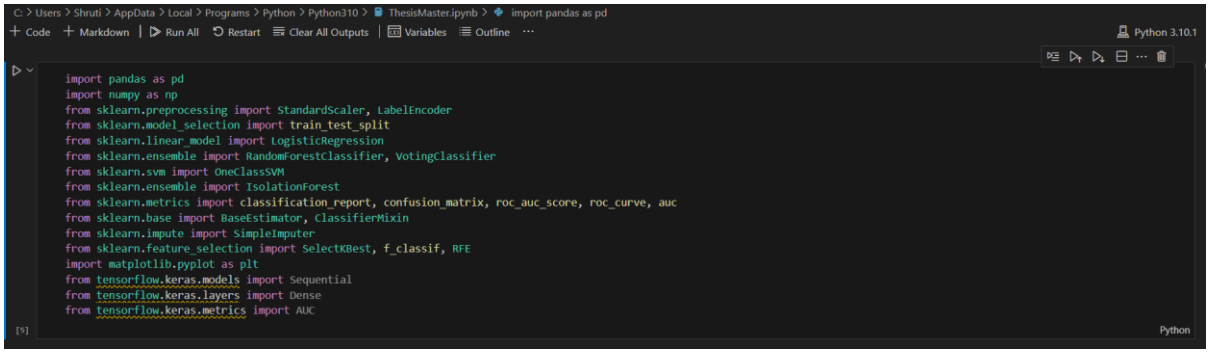

Image 5: Libraries Imported

Also import the CIC-IDS 2017 dataset using the following command;
df = pd.read_csv(“Path to the dataset/file name.csv”).

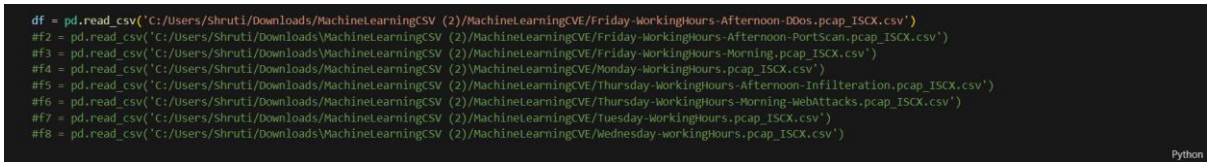

Image 6: Importing Dataset

Once the dataset is imported, preprocessing of the same is to be done using the preprocessing techniques. The following are the techniques used;

1. SimpleImputer from Scikit-learn is used to handle the missing values in the database. The missing values in the numeric columns are replaced by mean of the respective

column. This step makes sure that there are no missing values as it can impact the training process.
2. For infinite values in the dataset, **np.inf or -np.inf** fills the NaN values with 0.
3. Label Encoding is used to convert categorical target labels that is benign or malicious to binary/numeric values of 0 or 1. This step is important as machine learning models need numerical input values.
4. Feature Scaling is done using StandardScaler to standardize the features in the dataset which means to bring them to a standard scale of mean=0 and standard deviation=1.
5. Feature Selection is done using SelectKBest which selects the top 10 features based on ANOVA F-static using the command **f_classif.** This technique helps to decrease the dimensionality of the data and makes sure that only relevant features are utilized to train the model.
6. Recursive Feature Elimination which removes the least important feature of the dataset based on the chosen model.

```python
lr_model = LogisticRegression(class_weight='balanced')
selector_rfe = RFE(lr_model, n_features_to_select=10)
X_selected_rfe = selector_rfe.fit_transform(X_scaled, y)
```

Image 7: RFE for Logistic Regression

7. The dataset is divided into 70% for training and 30% for testing which means that the models are trained on a chunk of data and tested on the unseen data to assess its performance.

Post preprocessing of the dataset, the models will be trained and tested based on the CIC-IDS 2017 dataset(*IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB*, no date) which can be downloaded from (https://www.unb.ca/cic/datasets/ids-2017.html) website. The evaluation of the results are displaces using classification metrices like ROC-AUC scire, precision, recall, and F1-score of each model. The ROC curve of the models will also be plotted for performance visualization and comparing the Voting Classifier with other models.

References

Download Python (no date) *Python.org*. Available at: https://www.python.org/downloads/ (Accessed: 9 December 2024).

IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB (no date). Available at: https://www.unb.ca/cic/datasets/ids-2017.html (Accessed: 9 December 2024).

Visual Studio Code - Code Editing. Redefined (no date). Available at: https://code.visualstudio.com/ (Accessed: 9 December 2024).