

# Configuration Manual

MSc Research Project  
MSc in Cybersecurity

Sivaprakash Sivakumar  
Student ID: 23378514

School of Computing  
National College of Ireland

Supervisor: Vikas Sahni

**National College of Ireland**  
**MSc Project Submission Sheet**  
**School of Computing**



**Student Name:** Sivaprakash Sivakumar  
**Student ID:** 23378514  
**Programme:** MSc in Cybersecurity **Year:** 2025  
**Module:** Practicum Part 2  
**Lecturer:** Vikas Sahni  
**Submission Due Date:** 11/12/2025  
**Project Title:** Detection of Malicious URLs through Lexical Feature Analysis using Machine Learning

**Word Count:** 477

**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:** Sivaprakash Sivakumar

**Date:** 11/12/2025

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

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

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

|                                  |  |
|----------------------------------|--|
| <b>Office Use Only</b>           |  |
| Signature:                       |  |
| Date:                            |  |
| Penalty Applied (if applicable): |  |

# Configuration Manual

Sivaprakash Sivakumar

23378514

## 1 System Environment Used

The entire project was executed on a MacBook Air M1 system with the following configuration:

- Device: MacBook Air M1 (2020)
- OS: macOS Sonoma
- RAM: 8 GB
- Python Version: 3.11 (ARM64 Homebrew build)
- Virtual Environment: venv (named `practicum.tf`)

## 2 Environment Setup

- Installed Homebrew (ARM64 version) to manage system packages.
- Installed Python 3.11 using Homebrew.
- Created a dedicated virtual environment named `urlenv` for the project.
- Activated the virtual environment for dependency isolation.
- Upgraded `pip`, `setuptools`, and `wheel` to ensure smooth package installations.

## 3 Libraries and Exact Versions Used

### 3.1 Core Scientific and Data Libraries

| Library | Version Used |
|---------|--------------|
| numpy   | 1.26.4       |
| pandas  | 2.3.3        |
| joblib  | 1.5.2        |

### 3.2 Machine Learning Libraries (Scikit-Learn)

All ML-related imports (`train_test_split`, `LabelEncoder`, `StandardScaler`, `MLPClassifier`, `RandomForestClassifier`, `metrics`, `class_weight`) come from:

| Library      | Version Used |
|--------------|--------------|
| scikit-learn | 1.7.2        |

### 3.3 Built-in Python Libraries

These modules are part of Python standard library (no version required):

- `re`
- `math`
- `urllib.parse`

## 4 Data Used

- Dataset: `spam_url_dataset.csv`
- Columns required: `url`, `label`

## 5 Feature Engineering Implemented

The following features were extracted per URL:

- URL length, host length, path length, query length
- Count of digits, letters, special characters
- Digit ratio, special character ratio
- Count of `"/`, `."`, `"-"`, `"="` and `"@"`
- Number of subdomains and subdirectories
- TLD length
- Shannon entropy
- Presence of HTTPS/HTTP
- Presence of IP address in host
- Suspicious keyword count

## 6 Model Training Configuration Used

### 6.1 Train/Test Split

- Test Set size: 20%
- Training set size: 80%

### 6.2 Scaling Used

- Standard Scaler applied to all features for MLP

### 6.3 Class Balancing

- Balanced class weights computed using `compute_class_weight`
- Sample weights applied to MLP training

### 6.4 Models Implemented

- **MLPClassifier** Hidden layers: (256, 128, 64), ReLU, Adam, early stopping enabled
- **RandomForestClassifier** 700 estimators, `balanced_subsample`, `max_features = sqrt`

### 6.5 Model Selection

- Both models evaluated on test set
- Best model chosen based on accuracy
- Saved as `offtech_best_model.pkl`

## 7 Outputs Generated During Training

- `offtech_best_model.pkl`
- `offtech_label_classes.txt`
- `offtech_scaler.pkl`
- Classification report (per model)
- Confusion matrix (per model)

## 8 References

1. Pandas Documentation. <https://pandas.pydata.org/docs/>
2. NumPy Documentation. <https://numpy.org/doc/>
3. Scikit-Learn Documentation. <https://scikit-learn.org/stable/>
4. Joblib Documentation. <https://joblib.readthedocs.io/en/latest/>
5. Python Standard Library (re, math, urllib). <https://docs.python.org/3/library/>
6. TensorFlow for Apple Silicon (Metal Acceleration).  
[https://www.tensorflow.org/install/pip#apple\\_silicon\\_macs](https://www.tensorflow.org/install/pip#apple_silicon_macs)
7. Streamlit Documentation. <https://docs.streamlit.io/>
8. OFFTECH Tool and End URL Finder (IRJET 2022).  
<https://www.irjet.net/archives/V9/i5/IRJET-V9I5637.pdf>
9. Kaggle Phishing URL Dataset.  
<https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset>
10. Random Forest Classifier — Original Paper.  
<https://link.springer.com/article/10.1023/A:1010933404324>
11. MLP (Multi-Layer Perceptron) Documentation.  
[https://scikit-learn.org/stable/modules/neural\\_networks\\_supervised.html](https://scikit-learn.org/stable/modules/neural_networks_supervised.html)
12. URL Parsing Reference (urllib.parse). <https://docs.python.org/3/library/urllib.parse.html>
13. Regular Expressions (re). <https://docs.python.org/3/library/re.html>