**2024 HQ cyber security class**

**Individual project guideline**

**(modified version)**

1. **Topics : Cyber security coding using ML,DL Python library**
2. **Language : Python**
3. **Library : Scikit-learn, TensorFlow on python(choose one)**
4. **Format : guided format**
5. **Page volumes : 15pages at least**
6. **Submission address : class Google drive**
7. **Submission dead line : Sep 24, 25**
8. **Mission :**

* **Survey AI security Python source code example from ref. site**
* **Clone/download the code**
* **Execute the code, find the error and fix it as far as you do**
* **Explain the code**

1. **Reference site**

|  |
| --- |
| **AI chatbots: In Vietnam, various generative AI chatbots such as ChatGPT, Gemini AI and Bing Chat are popular.** |

* **Gemini AI**

Google's latest AI chatbot model, multimodal function

Understands and processes data in various formats.

* **Gemini official site**: [Gemini AI](https://gemini.google.com/?hl=ko) 1
* **DeepMind Gemini tech.page**: [DeepMind Gemini](https://deepmind.google/technologies/gemini/) 2
* **Gemini supporting**: [Gemini 지원](https://support.google.com/gemini/answer/13278668?hl=ko) 3
* **Google Cloud Gemini** : [Google Cloud Gemini](https://cloud.google.com/ai/gemini?hl=ko) 4

https://gemini.google.com/app/cae785ec9a0df194

* **Bing AI Chat**

<https://www.bing.com/?toWww=1&redig=A519D80C0451458E816B246AB9A14D8D>

Microsoft's conversational AI service Bing Chat is an AI-based chat tool that can be used for various purposesin Vietnam. When a user asks a question or makes a request, this tool provides answers or helps generate the necessary information.

* **WRTN**

Korean site, a platform that provides useful information and opportunities

for communication in Vietnam

<https://wrtn.io/en/overview/>

[https://wrtn.ai/](https://wrtn.ai/)%EC%97%90)

[https://wrtn.ai/chat/u/663397a3208f03982f4f7dae/c/66c9955571626b4b88890188?type=u](https://wrtn.ai/chat/u/663397a3208f03982f4f7dae/c/66c9955571626b4b88890188?type=u&fbclid=IwZXh0bgNhZW0CMTAAAR1dhuzIs2qZvgRYI-0TnskHW2bj_7YWLh0iJIDcsghMB10yRqL0mS47LOM_aem_XOShWVwdozpXILLR-pyYow)

* **Chatgpt**

[https://chatgpt.com/share/f63baf30-74c6-4dcb-8908-f8a1893da6ad](https://chatgpt.com/share/f63baf30-74c6-4dcb-8908-f8a1893da6ad?fbclid=IwZXh0bgNhZW0CMTAAAR10HFqoiOyV8WH4ZqBOwwPFDnCG5PY4JZMBBWMzhh8B6NImvOBD0mJOlsc_aem_iFWMrkL0ge5zwPIh5KsI6g)

* **git HUB**

Repository exploration: On the main page of GitHub, you can search for

topics or languages ​​of interest to explore related repositories. You can check the project overview and usage through the README file of each repository.

* **Process of source code exercise**

|  |
| --- |
| **Sign - up -> log-in -> search the topic code -> compile the code -> find error -> explain the error -> explain the code** |

**Individual project topic**

(Topic NO. 01) “ code example of detecting DoS IP Spoofing using Python Scikit-learn”

(Topic NO. 02) “code example of detecting malware file on my computer using Scikit-learn”

(Topic NO. 03) “code example of detecting DDoS attack using Python Scikit-learn”

(Topic NO. 04) “code example of detecting DoS attacks using TensorFlow”

(Topic NO. 05) “SQL injection Python code example using TensorFlow”:

example

(Topic NO. 06) “malware detection code example on my computer using TensorFlow”

(Topic NO. 07) “ code example of how to load other files(excel,html) on scikit learn program”

(Topic NO. 08) “ code example of basic process of sklearn.preprocessing using Python Scikit-learn”

**(Topic NO. 09) To perform an scanning test on a specific server IP (e.g., 111.111.111.111) using Scikit-Learn in Python**

**(Topic NO. 10) To perform an security risk analysis on a specific server IP (e.g., 111.111.111.111) using Scikit-Learn in Python**

**(Topic NO. 11) To perform a web system security check on a specific URL (e.g., https://test.com) using Scikit-Learn**

Report format

|  |  |
| --- | --- |
| class | CT201H M03 |
| Student id | B2111957 |
| Student name | Phan Trung Thuan |
| Student email | Thuanb2111957@student.ctu.edu.vn |
| Submitting date | 20/09/2024 |

**1. Title : Cyber security attacks detection using Machine learning Algorithms**

**2. Purpose of study(under five lines)**

The purpose of this study is to develop and enhance methods for detecting cyber security attacks, including Denial of Service (DoS), Probe, User to Root (U2R), and Cybil attacks. By leveraging advanced machine learning and deep learning techniques, the study aims to create a robust detection framework that can identify these attacks in real-time. The goal is to improve accuracy, reduce false positives, and provide scalable solutions for network security, ensuring better protection against evolving cyber threats.

**3. Scope of survey(list up the scope of the searching by items)**

Types of Cyber Attacks:

* Denial of Service (DoS): Attacks aimed at overwhelming network resources.
* Probe: Scanning techniques used to gather information about a network.
* User to Root (U2R): Exploiting system vulnerabilities to gain root access.
* Cybil Attacks: Distributed attacks where an attacker manipulates multiple fake identities.

Detection Techniques: Signature-based Detection (identifying known attack patterns)

Machine Learning Approaches: Supervised Learning, including RandomForest, GradientBoost, XGBoost, LogisticRegression, KnearestNeighbor, GaussianNaiveBayes, SupportVectorMachine and Stacking model.

Evaluation Metrics: Accuracy, F1-score.

Datasets for Research: KDD CUP 99, a widely used dataset for network intrusion detection.

**4. Results of exercise**

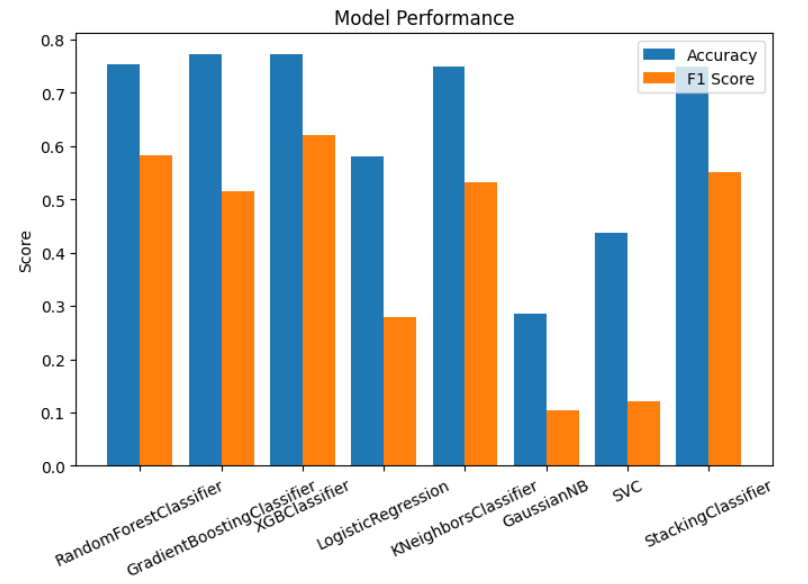
1. **Install Python Library**

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21 | *# module imports*  import numpy as np  import pandas as pd  import matplotlib.pyplot as plt  import seaborn as sns  import itertools  import random  *# model imports*  from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, StackingClassifier  from xgboost import XGBClassifier  from sklearn.neighbors import KNeighborsClassifier  from sklearn.linear\_model import LogisticRegression  from sklearn.naive\_bayes import GaussianNB  from sklearn.svm import SVC  *# processing imports*  from sklearn.preprocessing import LabelEncoder  from sklearn.model\_selection import cross\_val\_score  from sklearn.metrics import accuracy\_score, f1\_score, make\_scorer |

1. **Clone the Code**

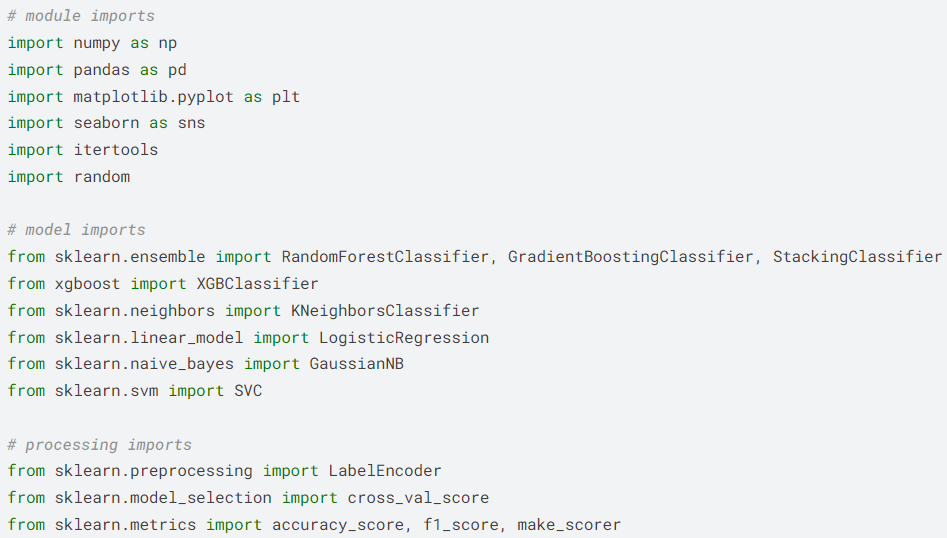
This research base on this code (https://www.kaggle.com/code/maryamanwer/ddos-attack-detection-using-ml/notebook)

1. **Execute the program**

****

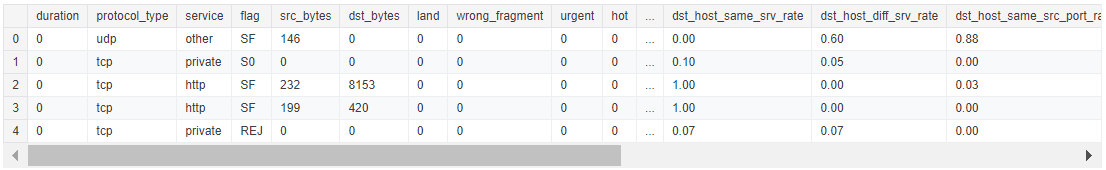
1. **Check error and modify the logics**
2. **Explain your code**

Firstly, I import some necessary library:



Second, load dataset:



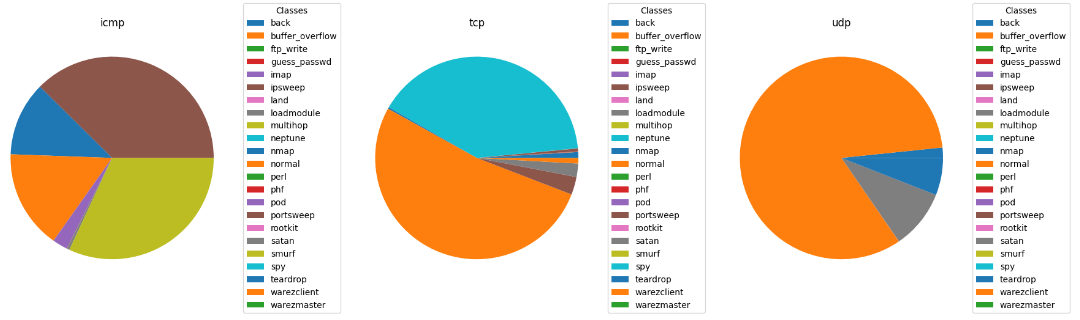


The raw data need to be standardized:

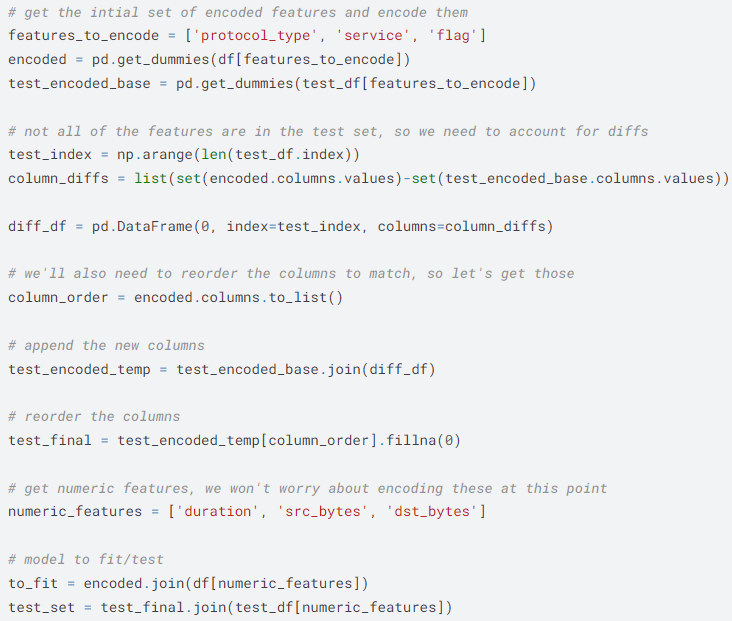




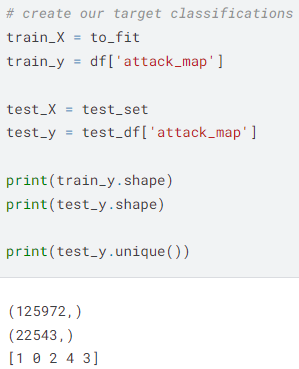
Visualize the dataset distribution:



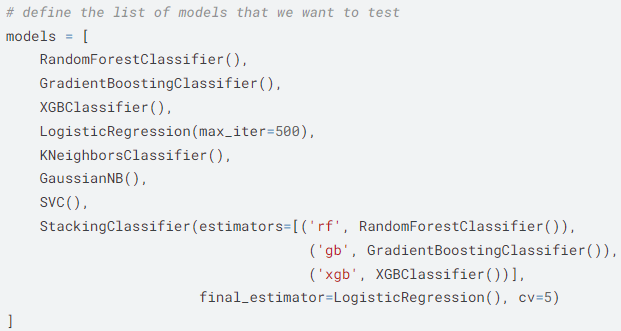
Apply one-hot encoding for every categorical features because some model require numeric features for calculation. Therefore, the number of columns increase to 87

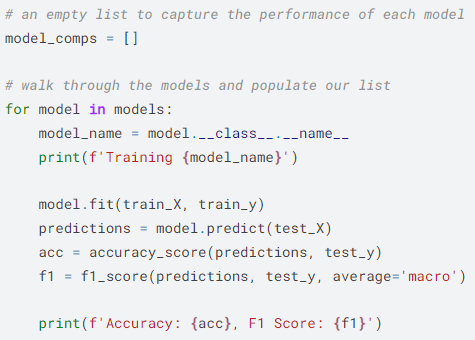


Checking number of samples for training and testing. So there are almost 126000 samples for training and about 22500 samples for testing

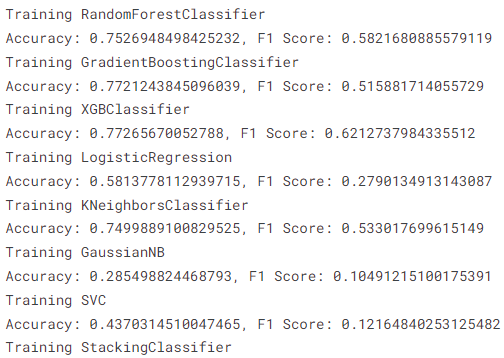


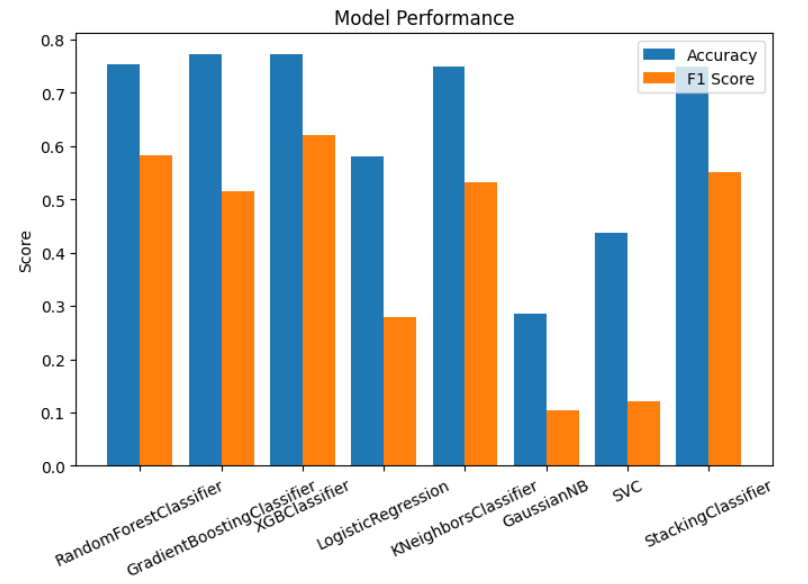
Define a list of models I want to testing on it and a training loop.





Last but not least, checking the result:



****

1. **Conclusion**

In conclusion, this research presents a comprehensive approach to detecting various types of cyber security attacks, including DoS, Probe, U2R, and Cybil attacks. By utilizing advanced machine learning techniques, the study demonstrates how XGBoost can significantly improve the accuracy and efficiency of identifying these malicious activities. The implementation of robust detection mechanisms not only enhances real-time threat detection but also reduces false positives, contributing to the overall security of network infrastructures.

1. **References**

[1] Scikit-learn: https://scikit-learn.org/stable/

[2] Colab: https://colab.research.google.com/

[3] KDD dataset: https://www.kaggle.com/datasets/hassan06/nslkdd