Configuration Manual for A Hybrid Ensemble Model using XGBoost and AdaBoost to detect and distinguish zero-day attacks

MSc Research Project
MSc Cyber Security

Ajay Krishna Edakkat Parambil
Student ID: X22110674

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ajay krishna Edakkat Parambil		
Student ID:	X22110674		
Programme:	MSc Cyber Security	**Year:**	2023-2024
Module:	MSc Academic Internship		
Lecturer:	Vikas Sahni		
Submission Due Date:	January 31		
Project Title:	A Hybrid Ensemble Model using XGBoost and AdaBoost to detect and distinguish zero-day attacks		
Word Count:	…………445…………… **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: AJAY KRISHNA EDAKKAT PARAMBIL……………………………

Date: …31-01-2024…………………………………………………………………

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

Ajay Krishna Edakkat Parambil
x22110674

1 Introduction

This document discusses about the Hybrid Ensemble Model using base classifiers Random Forest, Decision Tree and AdaBoost as well as a meta-classifier called XGBoost was implemented and executed. The project is run in python programming language in google collab.

2 System Specifications

The Application is run in the following specifications
Code Editor : Google Collab
Python Version : 3.6.3
Operating System : MacOs Sonoma Version 14.1.2

2.2 Software Requirements

Google Colaboratory: cloud-based jupyter notebook, python version 3.6.3

3 Package Details

NumPy and Pandas for data manipulation, Seaborn for data visualisation, imbalanced-learn (imblearn) for addressing class imbalance using SMOTE, scikit-learn for various machine learning functionalities, matplotlib for plotting and charting, XGBoost for gradient boosting, mlxtend for stacking classifier implementation, and other specific modules for tasks such as classification reports, train-test splitting, decision tree and random forest classifiers label.

Package	Version
numpy	1.23.5
pandas	1.5.3
seaborn	0.12.2
matplotlib	3.7.1
xgboost	2.0.2
mlxtend	0.22.0

numpy : It is used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices

pandas : It offers data structures and operations for manipulating numerical tables and time series matplotlib: It is used for plotting various graphical visualisations

4 DataSet

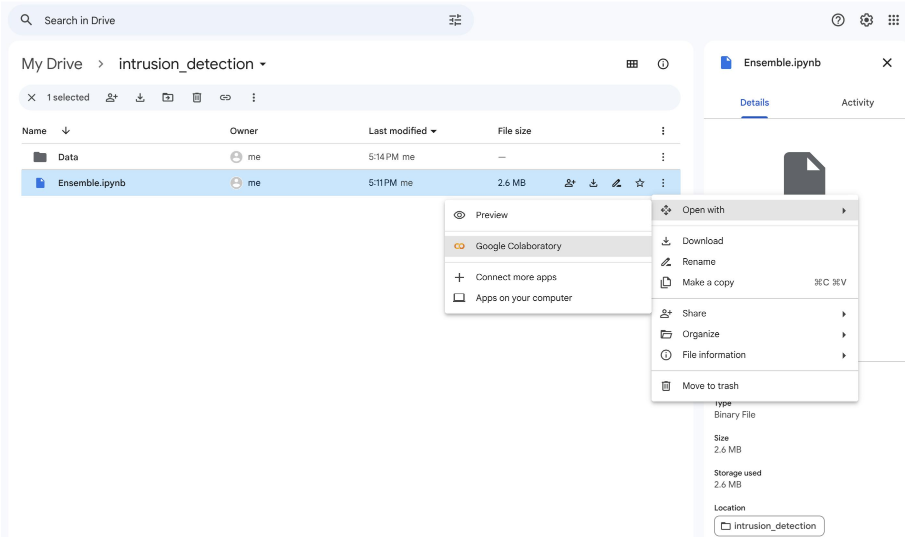
The dataset is from Kaggle[1], UNSW-NB15 a network intrusion dataset that contains raw network packets. It was created by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security at the University of New South Wales (UNSW) Canberra for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviors. A data training set named “UNSW_NB15_training-set.csv” was taken from this dataset.

5 Execution

Environment Setup

Google Collaboratory is used for the process of developing models. It is necessary to have a working Gmail account in order to access the Google Collaboratory. Python version 3.6.3 is used during the whole Model creation process.

1) After uploading the project file to your google repository, Now select the project file “open with -> Google Colaboratory



Name	Owner	Last modified	File size
Data	me	5:14 PM me	—
Ensemble.ipynb	me	5:11 PM me	2.6 MB

Type
Binary File

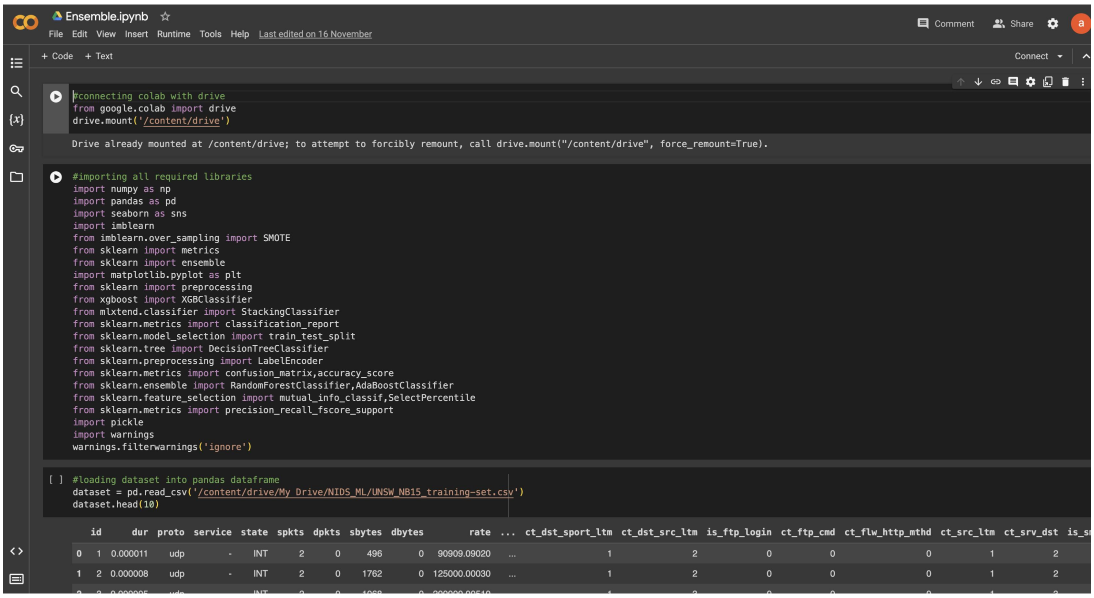
Size
2.6 MB

Storage used
2.6 MB

Location
intrusion_detection

[1] https://www.kaggle.com/datasets/mrwellsdavid/unsw-nb15/

2) After opening the project file with Colaboratory



Ensemble.ipynb ☆
File Edit View Insert Runtime Tools Help Last edited on 16 November
Comment
Share
+ Code + Text
Connect

```
#connecting colab with drive
from google.colab import drive
drive.mount('/content/drive')
```

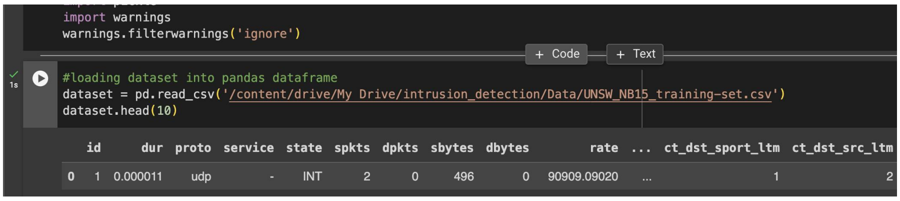
```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
#importing all required libraries
import numpy as np
import pandas as pd
import seaborn as sns
import imblearn
from imblearn.over_sampling import SMOTE
from sklearn import metrics
from sklearn import ensemble
import matplotlib.pyplot as plt
from sklearn import preprocessing
from xgboost import XGBClassifier
from mlxtend.classifier import StackingClassifier
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import confusion_matrix,accuracy_score
from sklearn.ensemble import RandomForestClassifier,AdaBoostClassifier
from sklearn.feature_selection import mutual_info_classif,SelectPercentile
from sklearn.metrics import precision_recall_fscore_support
import pickle
import warnings
warnings.filterwarnings('ignore')
```

```
[ ] #loading dataset into pandas dataframe
    dataset = pd.read_csv('/content/drive/My Drive/NIDS_ML/UNSW_NB15_training-set.csv')
    dataset.head(10)
```

```
   id       dur proto service state  spkts  dpkts  sbytes  dbytes          rate  ...  ct_dst_sport_ltm  ct_dst_src_ltm  is_ftp_login  ct_ftp_cmd  ct_flw_http_mthd  ct_src_ltm  ct_srv_dst  is_s
0   1  0.000011   udp       -   INT      2      0     496       0   90909.09020  ...                 1               2             0           0                 0           1           2
1   2  0.000008   udp       -   INT      2      0    1762       0  125000.00030  ...                 1               2             0           0                 0           1           2
```

3) Now change the location of the dataset to the correct location in your drive and click run

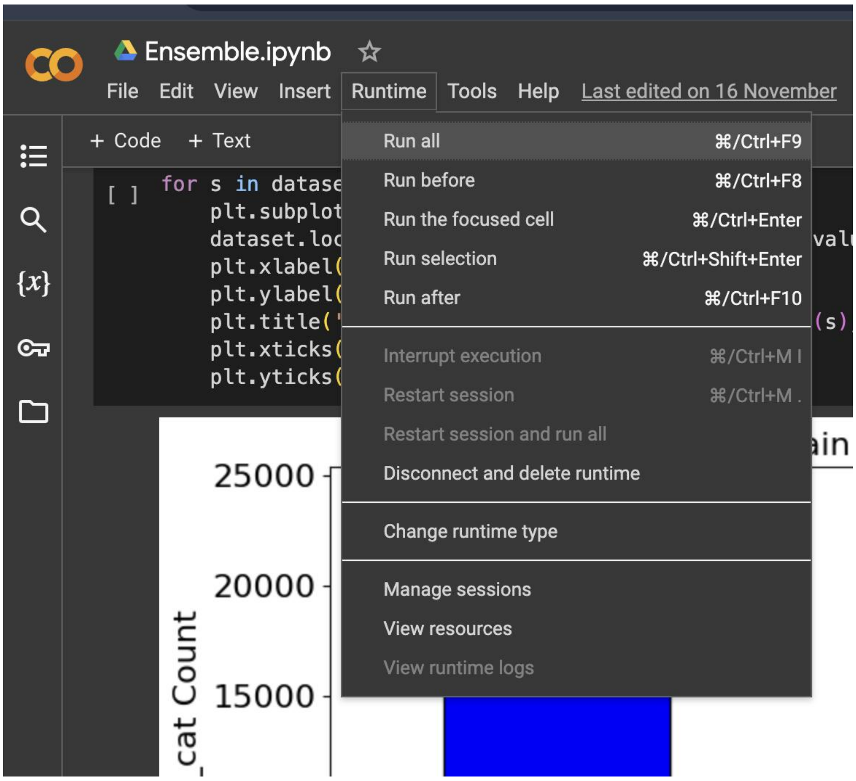


```
import warnings
warnings.filterwarnings('ignore')
```

```
#loading dataset into pandas dataframe
dataset = pd.read_csv('/content/drive/My Drive/intrusion_detection/Data/UNSW_NB15_training-set.csv')
dataset.head(10)
```

```
   id       dur proto service state  spkts  dpkts  sbytes  dbytes         rate  ...  ct_dst_sport_ltm  ct_dst_src_ltm
0   1  0.000011   udp       -   INT      2      0     496       0  90909.09020  ...                 1               2
```

4) Now click on Run time and select run all



Ensemble.ipynb ☆
File Edit View Insert Runtime Tools Help Last edited on 16 November
+ Code + Text
Run all ⌘/Ctrl+F9
Run before ⌘/Ctrl+F8
Run the focused cell ⌘/Ctrl+Enter
Run selection ⌘/Ctrl+Shift+Enter
Run after ⌘/Ctrl+F10
Interrupt execution ⌘/Ctrl+M I
Restart session ⌘/Ctrl+M .
Restart session and run all
Disconnect and delete runtime
Change runtime type
Manage sessions
View resources
View runtime logs
25000
20000
15000
cat Count
ain

```
[ ] for s in datase
        plt.subplot
        dataset.loc                val
        plt.xlabel(
        plt.ylabel(                (s)
        plt.title('
        plt.xticks(
        plt.yticks(
```

5) Now the dataset will be balanced with Smote(Juanjuan et al., 2007) all the models will start running and you can find the precision, accuracy, f1 score and recall of each model.

6) Additionaly the explanation for the code is provided in the link given below:

x22110674_academic_thesis_presentation&demo.mp4

References

Juanjuan, W., Mantao, X., Hui, W. and Jiwu, Z. (2007) 'Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding', *In International Conference on Signal Processing Proceedings, ICSP*.