National College of Ireland

Configuration Manual

Investigation of Machine Learning Algorithms for Malware Detection in PE and PDF Files

Masters in Cybersecurity

Rosemary Usoroh
Student ID: x21114374

School of Computing
National College of Ireland

Supervisor: Arghir-Nicolae Moldovan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	ROSEMARY USOROH
Student ID:	x21114374
Programme:	MASTER OF SCIENCE IN CYBERSECURITY **Year:** 2024
Module:	Msc Research Project
Lecturer:	Arghir-Nicolae Moldovan
Submission Due Date:	31-01-2024
Project Title:	Investigation of Machine Learning Algorithms for Malware Detection in PE and PDF Files
Word Count:	926 **Page Count:** 10

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: ROSEMARY USOROH

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

Rosemary Usoroh
Student ID: x21114374

1 Introduction

This configuration manual presents the requirements for creation of machine learning models in the detection of malware. The manual will give explanation of the software and hardware components that are needed for the completion of this project.

2 System Configuration

2.1 Hardware Requirement

The table below presents the hardware configuration used for the implementation of this project;

Hardware	**Configuration**
System	HP
Operating System	Windows 10(64 Bits) Pro
RAM	12GB
Hard Disk	1TB
Processor	Intel Core i5-10210U

Table 1 : Hardware requirement

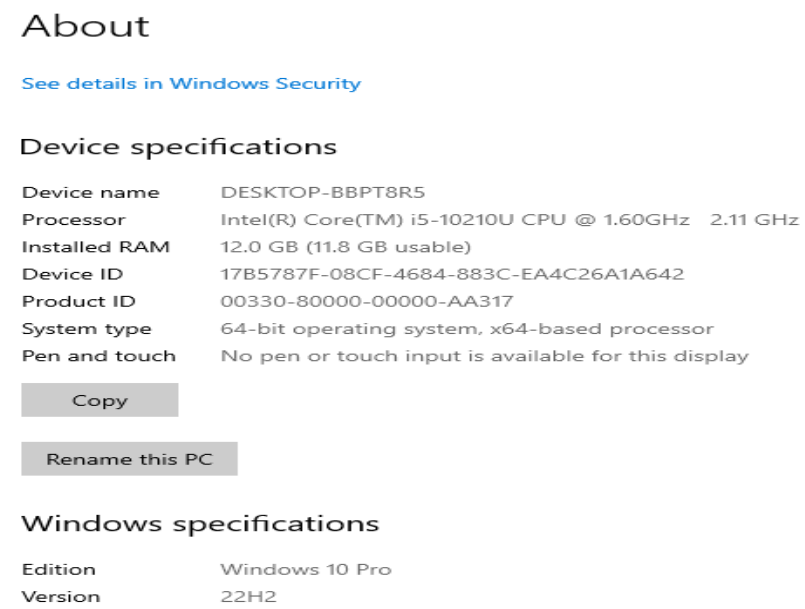
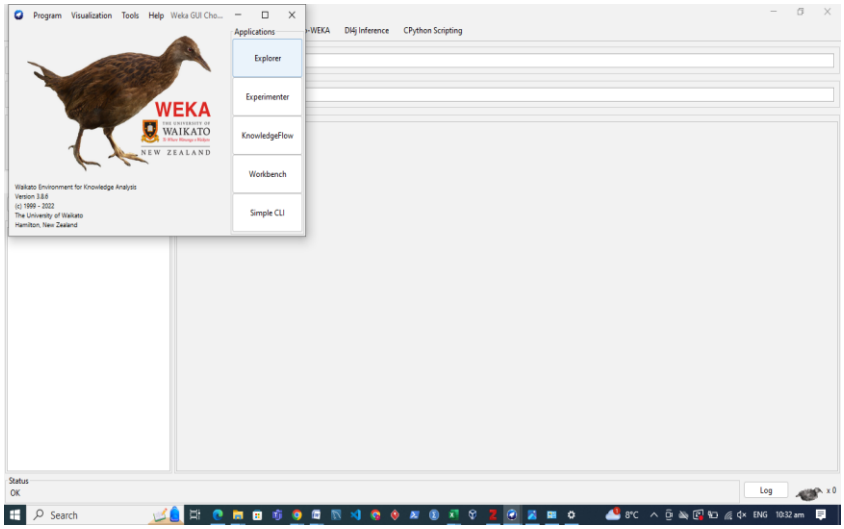


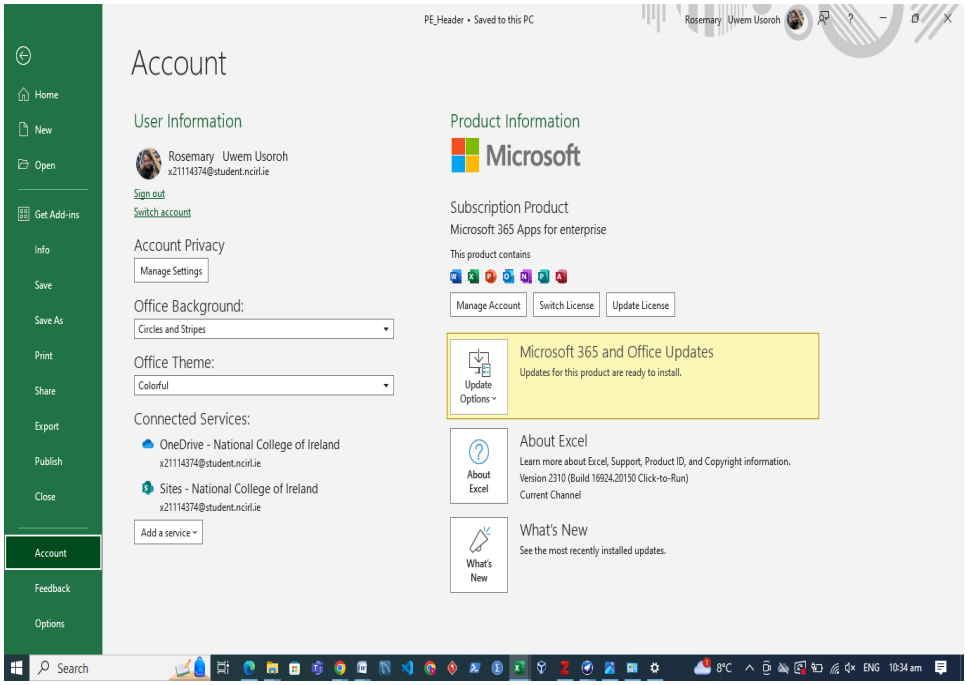
Figure 1: Operating System Configuration

2.2 Software Requirement

Software	Version
WEKA	3.8.6
Excel	2310
Oracle VM Machine	7.0.12

Table 2: Software requirements





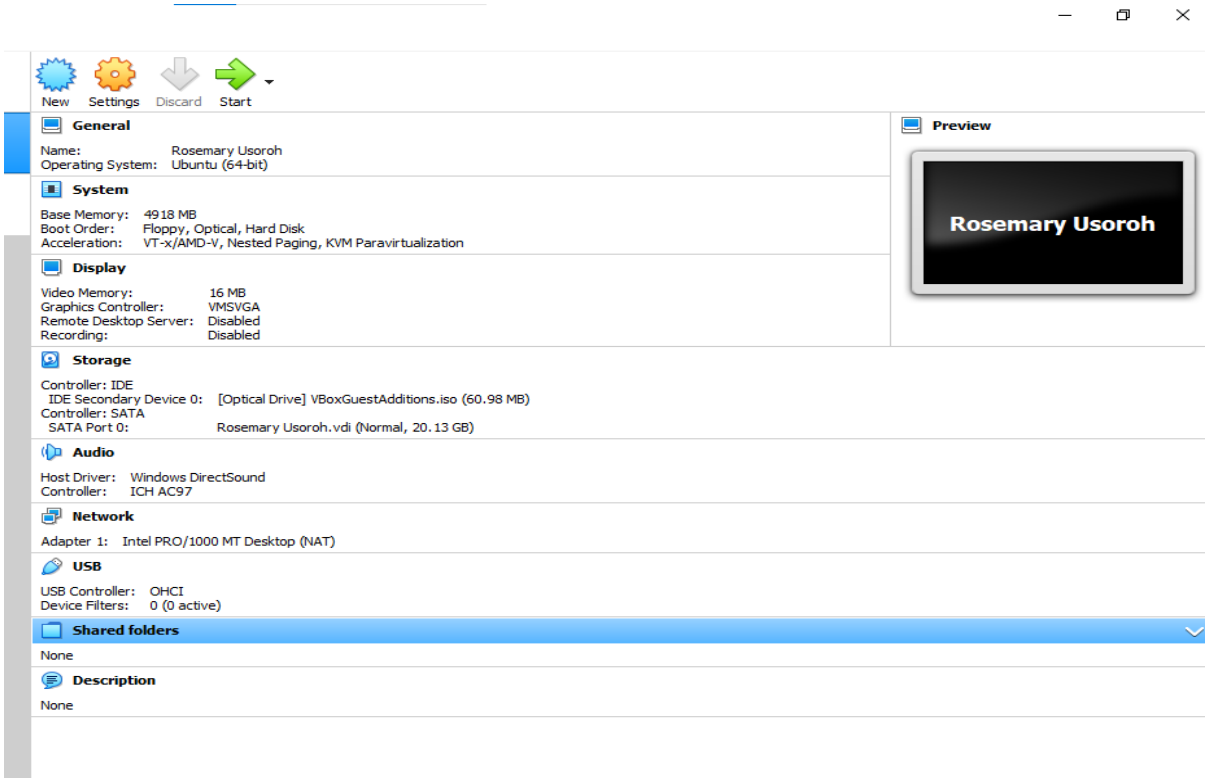


Figure 2: WEKA, Virtual Machine and Excel software

3 Project Implementation

The datasets used for this project are in the artefacts. This section will provide a summarized implementation of the research and pictorial presentation of the implementation.

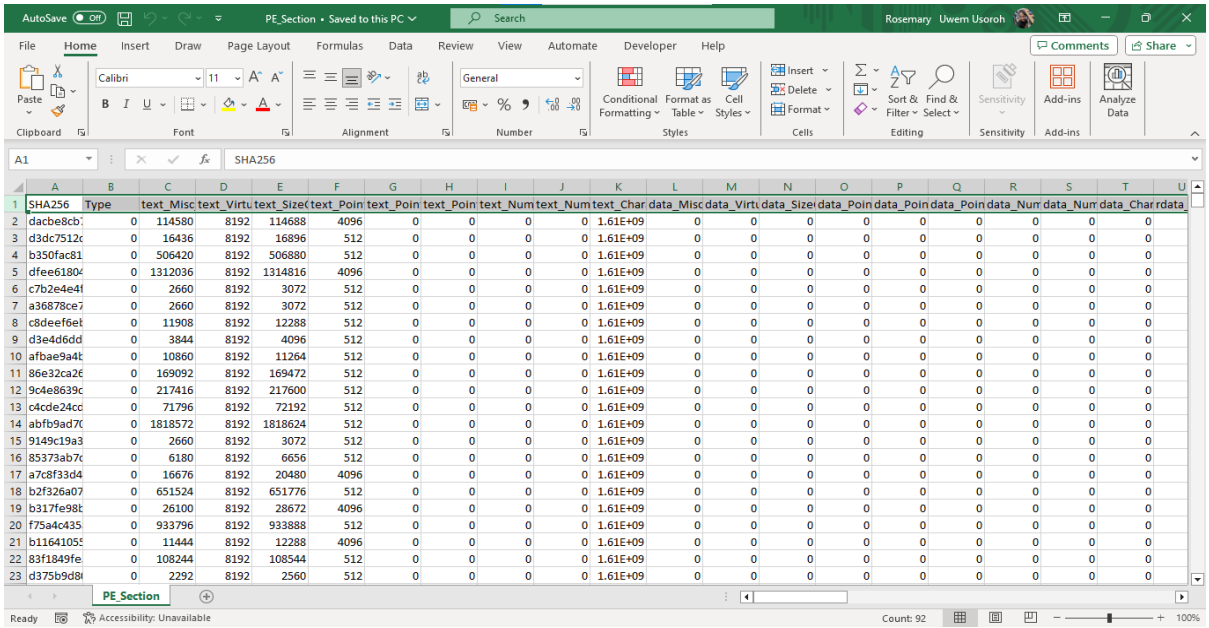


Figure 3: The figure above shows one of the dataset in the excel sheet.

In this research, four malware datasets were analyzed for this research. The first datasets was evasive PDF dataset (Evasive-PDFMal2022), it was downloaded from (Issakhani et al., 2022). The original file that was downloaded was in a CSV format. The second malware dataset which is Windows PE Malware (WinMal) samples was downloaded from (Yousuf et al., 2023). The samples were divided into four different parts which includes Imported DLLs, API Functions, PE Header and PE Section. The original files of these samples were all downloaded in CSV format.

3.1 Data Preprocessing

The Evasive PDFMal2022 and WinMal datasets came in a structured format, therefore needs little preprocessing. In the Evasive PDFMal2022, the row with NaN was removed on the excel sheet so that it is not treated as a separate class and for the classifier to be able to use the features. Inaccurate and corrupted data were also removed from the dataset on the excel sheet.

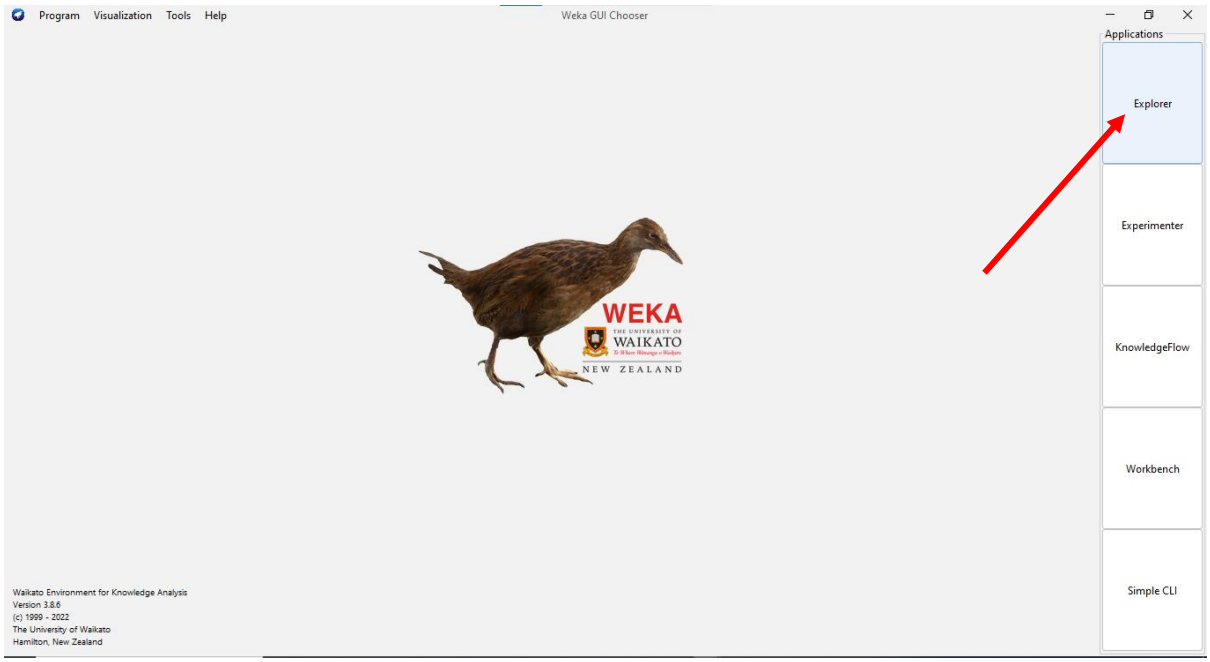


Figure 4 – The WEKA tool framework, the explorer section was were the main project was implemented.

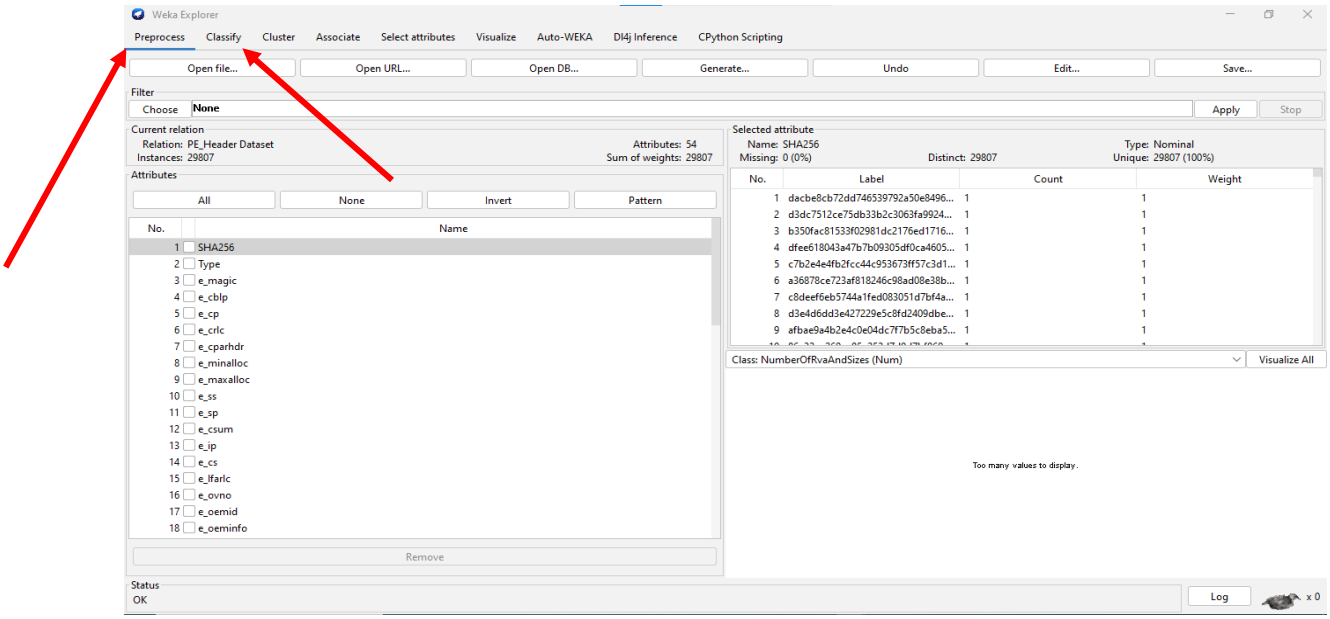


Figure 5 – The Weka explorer where the datasets can be preprocessed and classified with different models

In the WinMal dataset, the second feature set (API_Functions.csv files) was not included in the analysis. This is because the size of the dataset was too large (1. 21 GB size and 16384 features) and the WEKA tool will not be able to build models on it. The other datasets which are DLLs_Imported.csv file, PE_Header.csv file and PE_Section.csv file were used for the malware analysis. The SHA column in the datasets was removed in the WEKA tool to get accurate results from the classifiers.

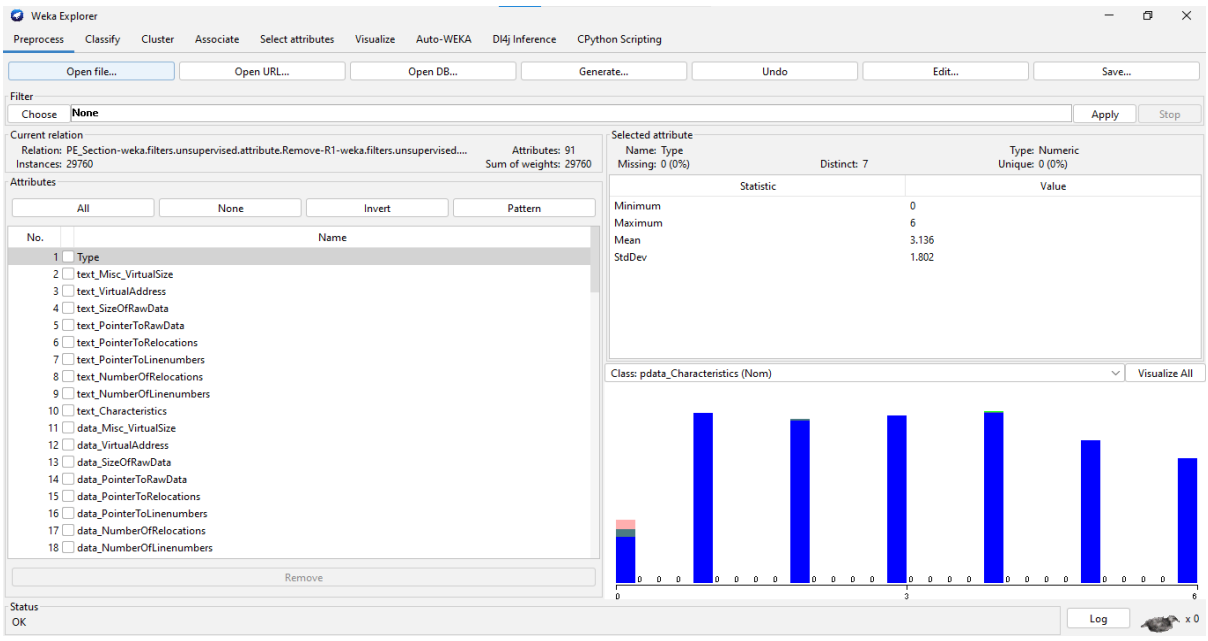


Figure 6: One of the WinMal (PE_Section) datasets after the SHA column has been removed on WEKA

3.2 Data Mining

Three machine learning algorithms were used to build the models. The machine learning algorithm includes; PART Rule, Ordinal Class Classifier and Bayes Net. 10-fold cross validation was used in the malware analysis. These machine algorithm are in the WEKA tool and these machine learning models were built on WEKA.

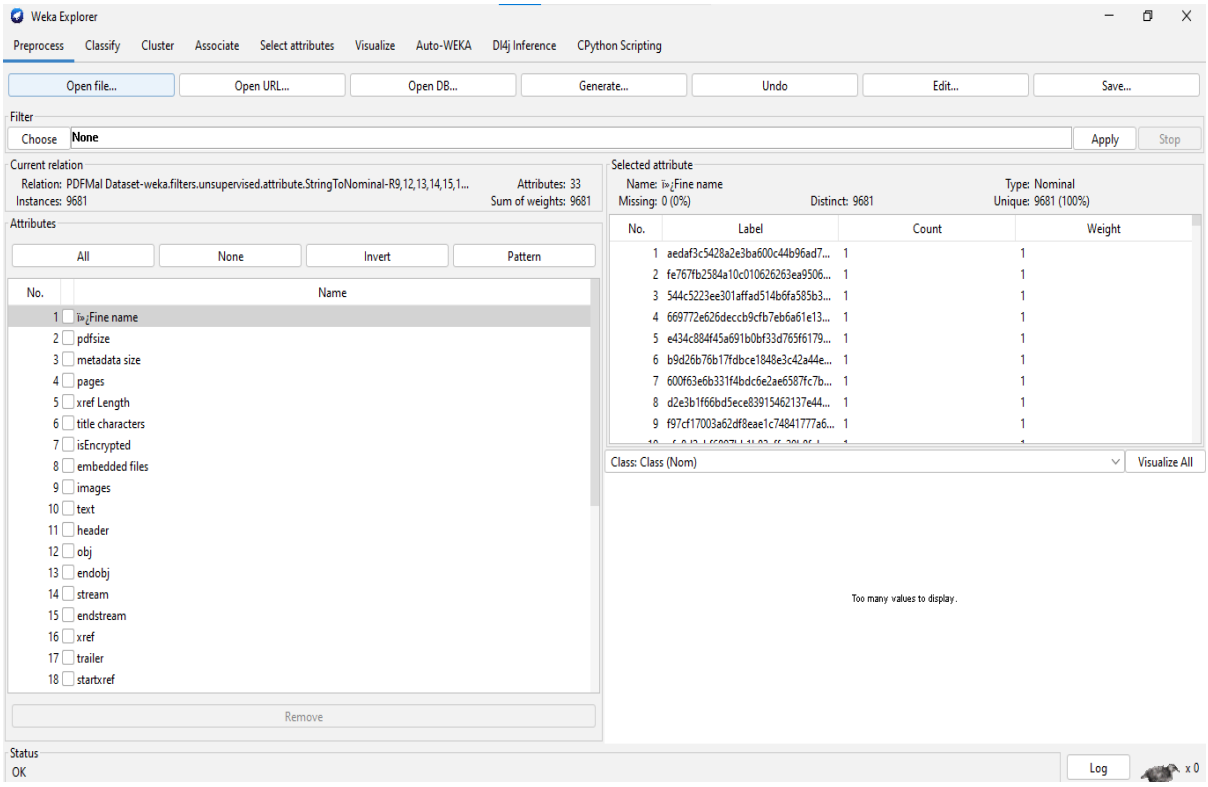


Figure 7: This is the PDFMal dataset opened on WEKA before classification

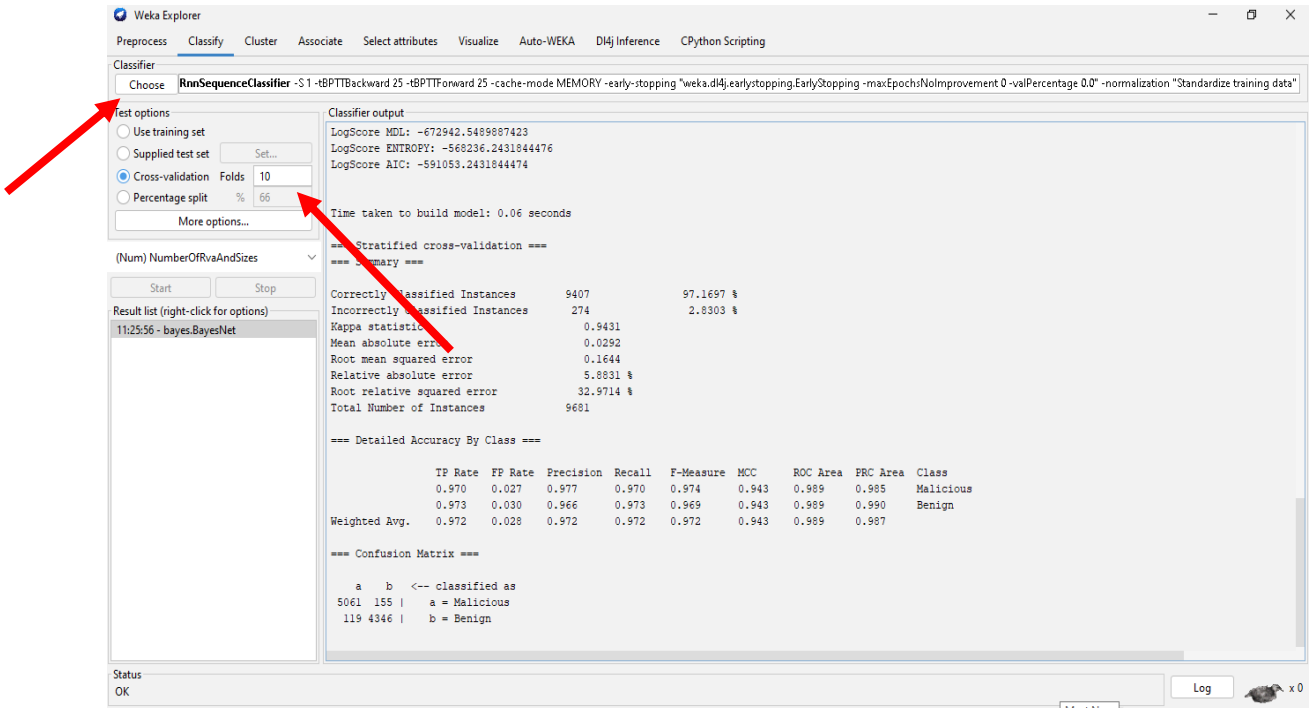


Figure 8 – The arrows showing where to choose the different machine learning algorithms to use and the type of cross-validation fold that was used.

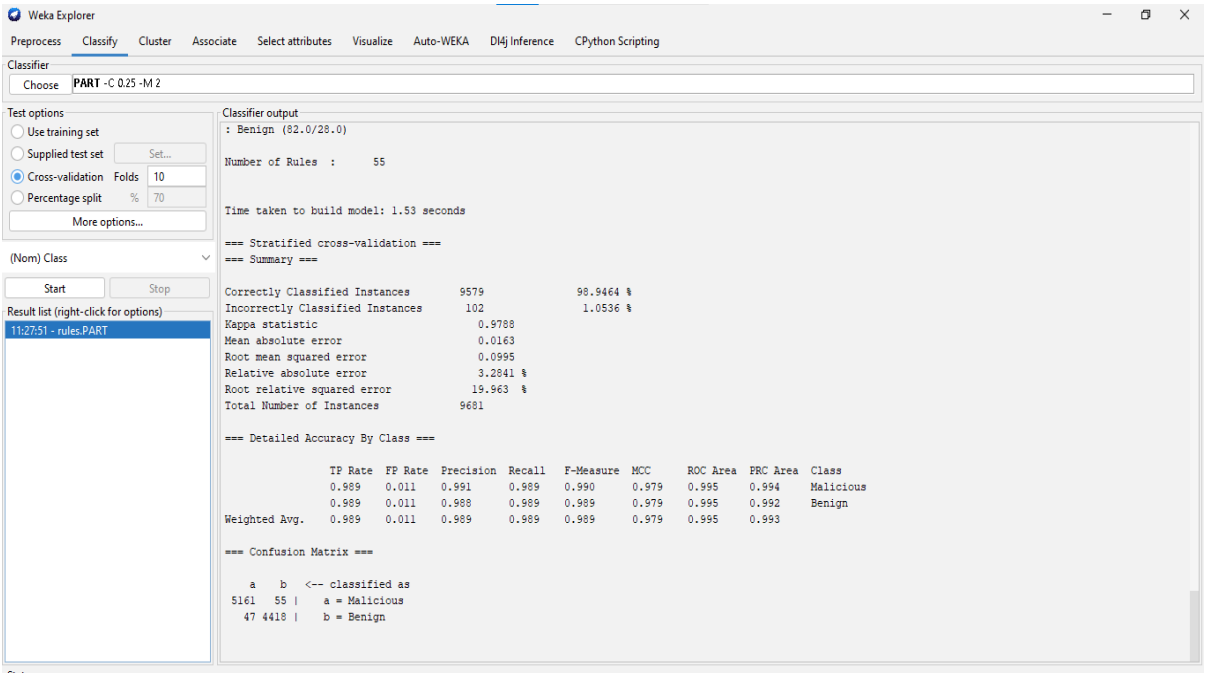


Figure 9: This is the PART model on PDFMal dataset. The accuracy and other metrics are displayed on the result.

4.0 Conclusion

The purpose of this research is to evaluate how accurately supervised machine learning algorithms can detect malware in PE files and PDF files. According to the result, Ordinal Class Classifier and PART Rule model achieved a detection accuracy of 100% on one of the WinMal dataset while Bayes Net achieved an accuracy of 99.98%. This means that PART Rule and Ordinal Class Classifier has greater accuracy rate when scanning malicious PE files and PDF files.

References

Issakhani, M., Victor, P., Tekeoglu, A., & Lashkari, A. (2022). PDF Malware Detection based on Stacking Learning: *Proceedings of the 8th International Conference on Information Systems Security and Privacy*, 562–570. https://doi.org/10.5220/0010908400003120

Yousuf, M. I., Anwer, I., Riasat, A., Zia, K. T., & Kim, S. (2023). Windows malware detection based on static analysis with multiple features. *PeerJ Computer Science, 9*, e1319. https://doi.org/10.7717/peerj-cs.1319