**Malware Detection**

END SEM PROJECT REPORT

***Submitted by***

BL.EN.U4AIE20009 ------ B Sai Kowshik

BL.EN.U4AIE20017 ------ G Mani Prakash Reddy

BL.EN.U4AIE20029 ------ K Balaji Reddy

**19AIE214 – Big Data Analytics**

**B.Tech. in Computer Science and Engineering (Artificial Intelligence)**

Logo

Description automatically generated with low confidence

AMRITA SCHOOL OF ENGINEERING, BANGALORE

AMRITA VISHWA VIDYAPEETHAM

BANGALORE 560 035

June – 2022

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.no** | **Topic** | **Page.no** |
| **1.** | **Problem Statement** | **3** |
| **2.** | **Dataset Description** | **4** |
| **3.** | **System Workflow** | **4-5** |
| **4.** | **Execution** | **5** |
| **4.** | **Conclusion** | **8** |
| **5.** | **References** | **8** |

**Problem Statement:**

Decision making is an important process today in almost all domains. In the specialty of cyber security, we must determine whether a file is malicious or benign. Malware is dangerous software, a programme, or a script that can damage any computing instance. Malware is the short form for ‘malicious software’, a technical word for noting some particular computer program or code which undertakes illegal tasks without the owner’s permission. In our Project, we have detected whether the applications are malicious or not.

In our Project, Given any unknown executable file we were able to detect whether it is malicious or benign.

We have appended the technology of Bigdata so as to improve the detection of malware based on the train and test data.

Overall, we were able to create a prediction model which consists a machine learning algorithm , as well as technology that operates on bigdata.

**What is Malware?**

Malware is a malicious software that gets installed in your device and performs unwanted tasks

It will be developed by cybercriminals to steal data and damage or destroy computers and computer systems.

It will also transmit information about our web browsing habits to the third party

**Types of Malware**

**Viruses:** It replicates itself and spreads by deleting files

**Trojan Horse:** It performs various tasks like corrupting files

**Spyware:** It transmits the user activities to the third party

**Adware**: It is used by companies for marketing purpose

**Worms**: It makes its own copies on local device

**Dataset Description**

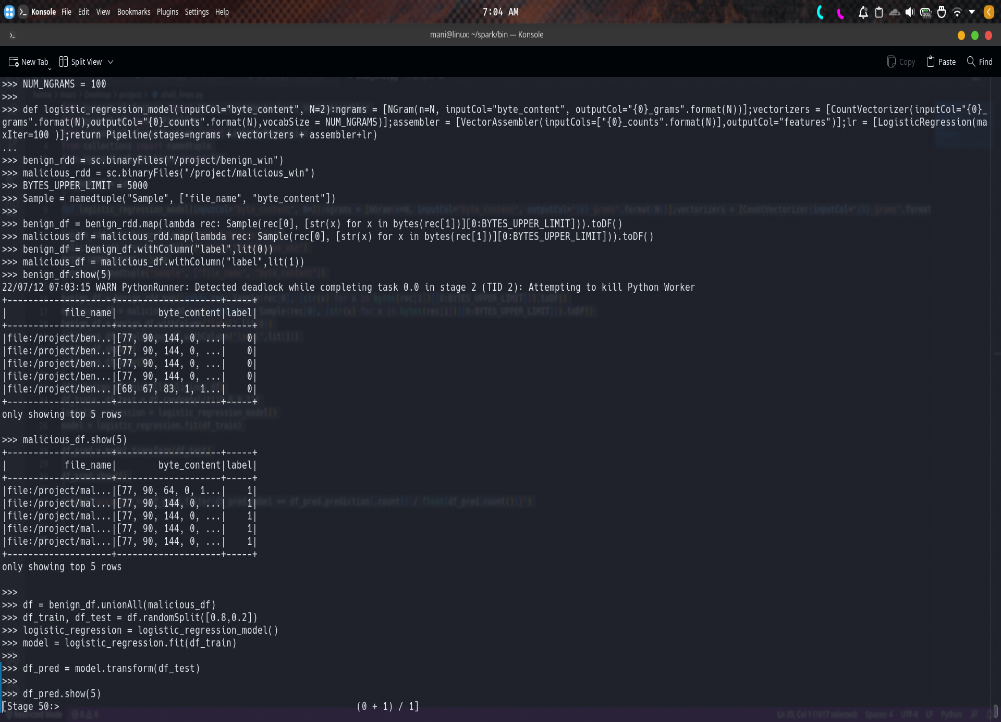
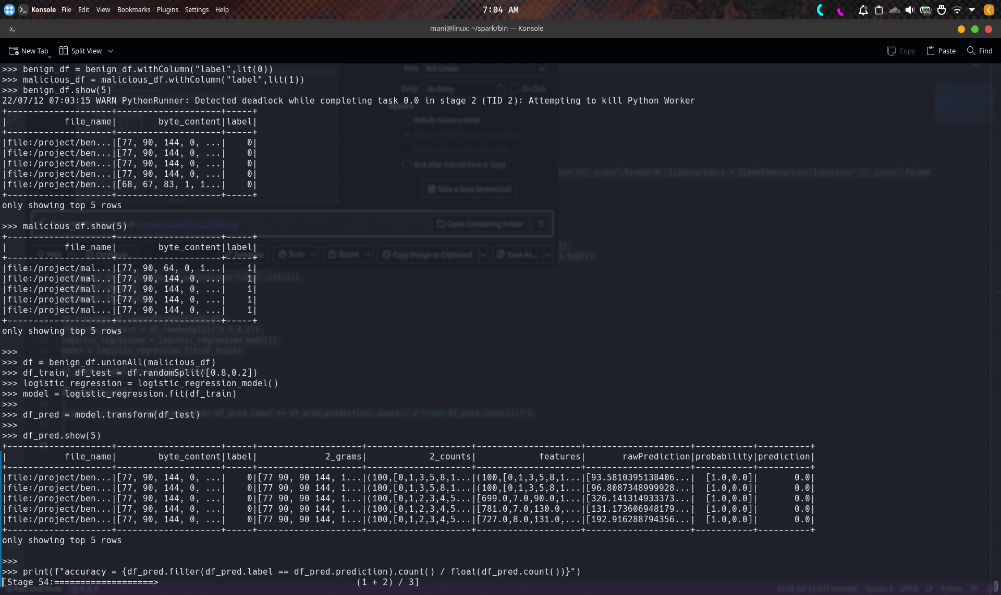
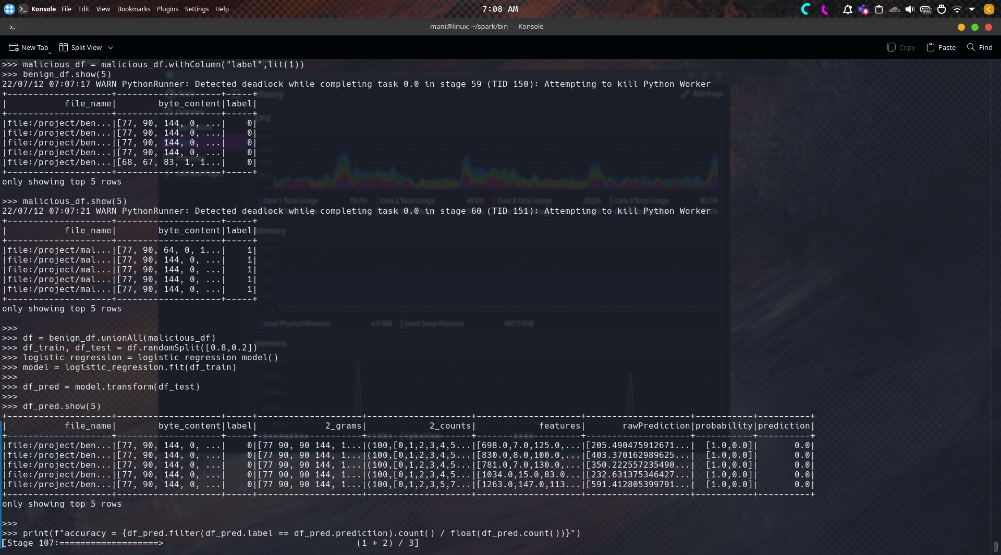
We have considered the dataset which consists of malicious as one class and benign as another class.

Both the malicious and benign class files are in the executable file format**.**

**System Workflow**

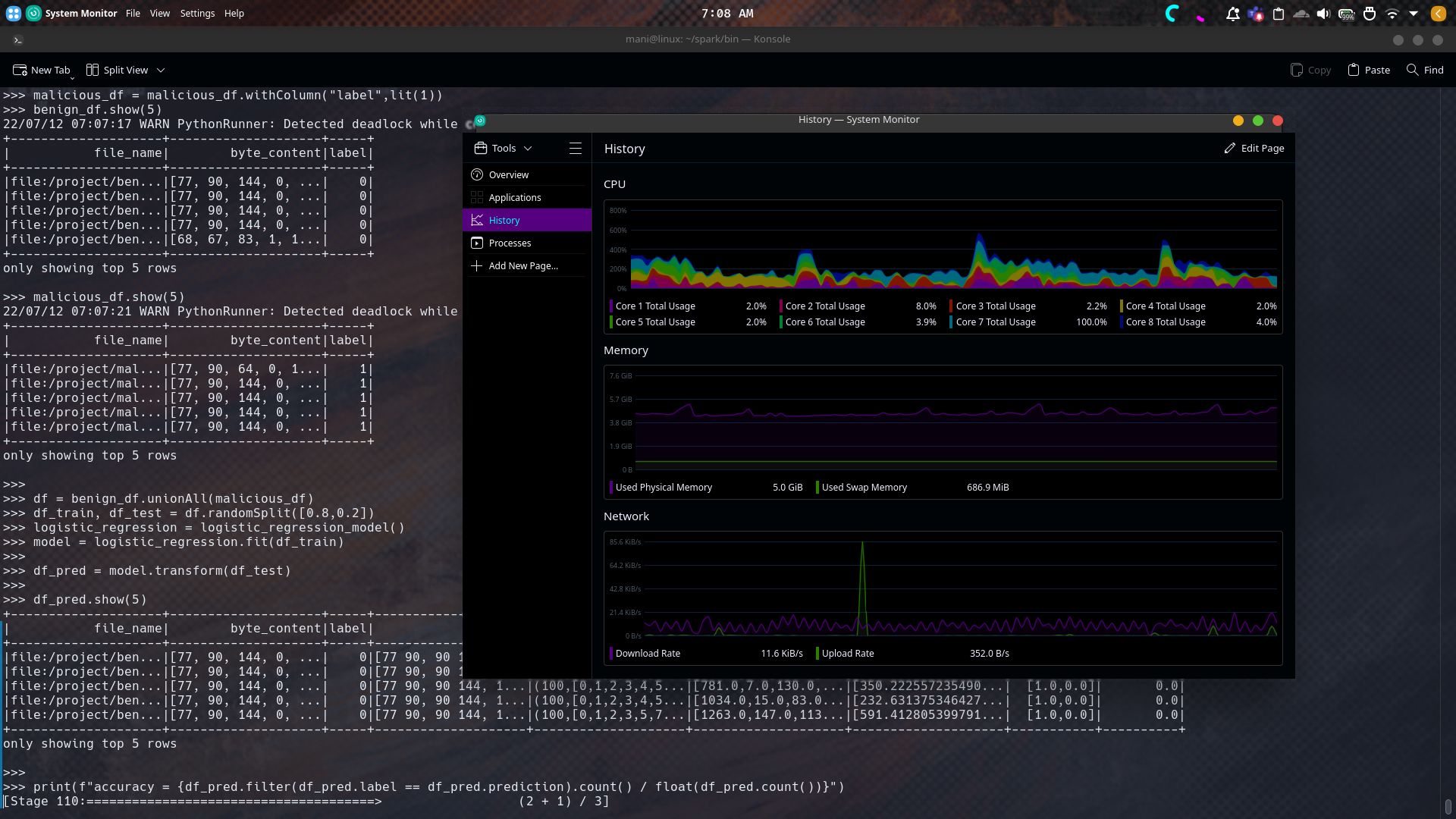
1. We will load malicious class into one RDD and benign to the another RDD
2. Then we will retrieve 500 bytes from the RDD’S and then we will convert them into data frames
3. After converting into individual data frames we will combine both as a single data frame by labelling 0 or 1
4. We will split the data and training data will be fit into the logistic regression model and we will find accuracy

**Execution:**

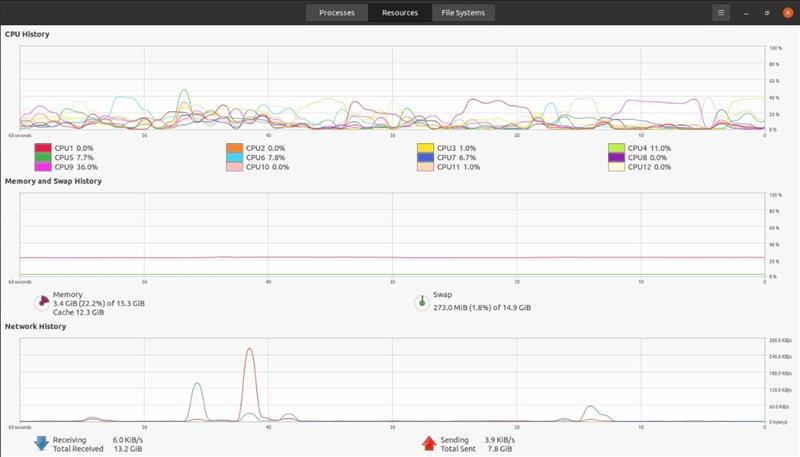
****

**A screenshot of a computer

Description automatically generated with medium confidence**



****

****

**Graphical user interface, application

Description automatically generated**

****

**Conclusion:**

We have implemented malware detection using machine learning algorithm and bigdata techniques on a dataset which is in executable file format and we can classify them into benign and malicious. We have tried to optimize the results by pre-processing of data and we have obtained an accuracy of 92 percent for Logistic Regression.

**References**

[1] Gavriluţ, Dragoş & Cimpoesu, Mihai & Anton, D. & Ciortuz, Liviu. (2009). Malware detection using machine learning. 4. 735 - 741. 10.1109/IMCSIT.2009.5352759.

[2] Kumar, Ajay & Abhishek, Kumar & Shandilya, Shishir K & Ghalib, Dr. Muhammad. (2020). Malware Analysis Through Random Forest Approach. Journal of Web Engineering. 19. 10.13052/jwe1540-9589.195610.

[3] Hughes, Kelly & Qu, Yanzhen. (2012). A Theoretical Model: Using Logistic Regression for Malware Signature based Detection.

[4] Deepak Gupta, Rinkle Rani,Improving malware detection using big data and ensemble learning,Computers & Electrical Engineering,Volume 86,2020.