Applying Machine Learning Technique for Knowledge Discovery in Network Database

Ayorinde I. T.
Department of Computer Science, University of Ibadan, Ibadan, Nigeria.
Email: temiayorinde@yahoo.com

---ABSTRACT---

Cyber attacks are malicious activities conducted in the digital realm with the intent of exploiting vulnerabilities, stealing sensitive information, disrupting operations, or causing damage to computer systems and networks. Network security is one of the viable ways of mitigating against cyber attacks. This study, through machine learning technique, was able to discover certain parameters that need to be taken cognizant of while working in a networking environment. An online network database retrieved from Kaggle was used in this study. Six inputs were used for the prediction of cyber attacks severity levels which was simulated with Naive Bayes algorithm in the Rapidminer Studio. The results show that the severity level "High" has the highest values for both raw and predicted data. It was also recorded that TCP has the least value (34%) for the predicted "High" severity level which shows that it is a good protocol to be used. On the other hand, HTTP had the highest value (67%) for the predicted "High" severity level which means that it is highly vulnerable to attack. With these results, internet users should make it of high priority to secure their data and network always by choosing the right protocols.

Keywords **- Cyber attack, Machine learning, Naive bayes, Severity level.**

Date of Submission: June 24, 2024. Date of Acceptance: July 3, 2024.

1. INTRODUCTION

Machine learning (ML), a subset of artificial intelligence, empowers systems to learn and make predictions or decisions without being explicitly programmed. Naïve Bayes is one of the fast and easy Machine Learning algorithms that is used to predict a class of datasets. It can be used for Binary as well as Multi-class Classifications. Naive Bayes performs well in Multi-class predictions as compared to the other Algorithms [1]. Discovering knowledge comes with the ability of a ML algorithm to infer new knowledge from a new set of data after training it with some data [2,3].

Cyber attacks are malicious activities conducted in the digital realm with the intent of exploiting vulnerabilities, stealing sensitive information, disrupting operations, or causing damage to computer systems and networks. These attacks encompass a wide range of tactics, techniques, and procedures employed by cyber criminals to compromise the confidentiality, integrity, or availability of digital assets. Cyber attacks can come in different forms like Malware, Phishing, Denial of Service (DoS) and Distributed Denial of Service (DDoS), Man-in-the-Middle (MitM) Attacks, SQL Injection and Zero-Day Exploits among others [4,5].

Cyber attacks can result in grave consequences such as financial loss, reputation damage, data breaches and operational disruption among others. It can be prevented through the use of firewalls and antivirus software, employee training, regular software updates, network segmentation and incident response plans and the use of appropriate network protocols among others [6,7].

The Naive Bayes algorithm is widely applied in various fields, including network databases [8] [9] [10]. It is utilized for classification tasks based on the assumption of feature independence within a dataset, making it particularly useful for large data sets and outperforming more complex classification methods [9]. In the context of network databases, the Naive Bayes algorithm has been employed in academic information systems for data classification based on protocols with different categories, showcasing high accuracy levels in throughput, delay, and packet loss classifications [10]. It has also been used for prediction [11]. Additionally, the algorithm has been utilized in the field of information retrieval for document searches and metadata descriptions, demonstrating its effectiveness in reducing information overload and enhancing search capabilities in databases of texts, images, and sounds [9].

Hence, this study uses the Rapid Miner studio to discover some patterns that inform decision making while predicting a network database using Naive Bayes algorithms.

2. RELATED WORKS

In [8], the work underscores the importance of data mining, particularly using Naive Bayes algorithm for extracting insights and making informed decisions in the e-commerce industry and beyond. Patterns were identified within the dataset used which aids in prompt decision making.

The authors in [10] used the Naive Bayes algorithm, a machine learning method that utilizes probability and statistical calculations. Classification is carried out on data protocols which have low, medium and high categories. The results in this study were the throughput on the server was 38.8% in the medium category, the delay on the server was 2.80 ms in the very good category, and the packet loss was 0% in the very good category. The results of classification on the protocol have two confidence that is producing an average accuracy value that is right for classification on the long protocol of 94.92% and the protocol counting of 81.35%. These findings indicate the effectiveness of using machine learning algorithms like Naive Bayes for network status classification in academic information systems, thereby providing valuable insights for system optimization and performance enhancement.

In order to improve the intrusion detection ability of multi-dimensional node combination mixed topology network, the author in [12] proposes an intrusion detection method based on naive Bayes algorithm. The work builds a distributed structure model of intrusion data in the network, and conduct traffic statistics and feature analysis on the network through low-speed monitoring and combined frequency scanning, so as to extract abnormal traffic label features of data in the network. According to the types of attacks, the fuzzy clustering center of intrusion data was detected while the fusion model of anomaly feature distribution of intrusion traffic sequence was also established. The test results show that the intrusion data detection results obtained by this method have high accuracy, so it has good detection performance and strong anti-interference ability, which can be used to improve the network security and anti attack ability.

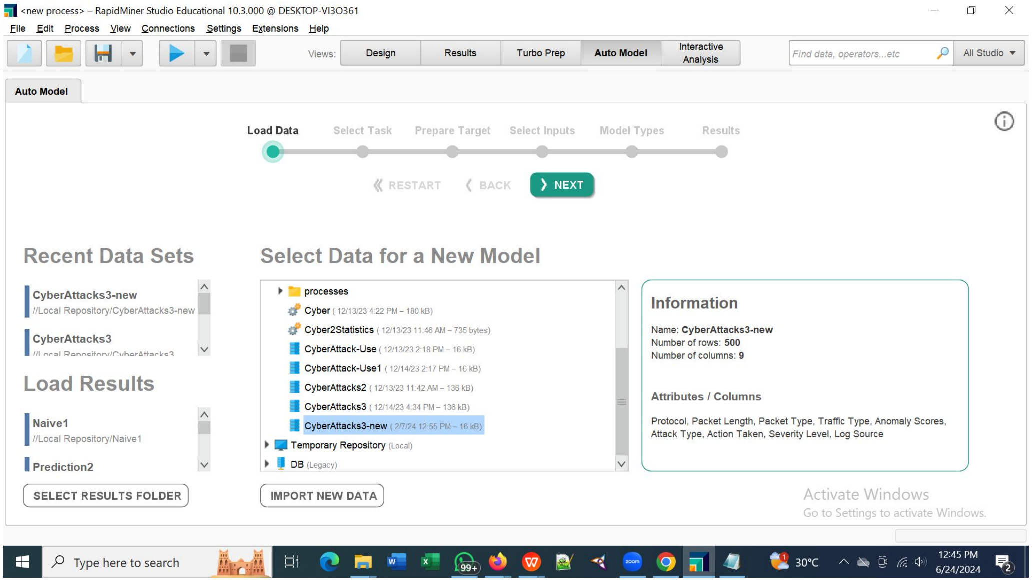
The research in [13] enables the authors to realize the best neural network algorithm which helps in predicting the pervasive developmental disorder (PDD) using Naïve Bayes neural network algorithm with accuracy and reducing both time and cost for prediction. The paper also gives the comparison between machine learning algorithms and helps to know which algorithm gives better output with more accuracy.

The authors in [14], through their study, demonstrated the usefulness of unsupervised machine learning techniques with the potential to uncover latent clinical phenotypes. The study gave a better understanding of the different clinical phenotypes across the disease spectrum in patients with COVID-19. The work could serve as a more robust classification for patient triaging and patient-tailored treatment strategies.

3. METHODOLOGY

The dataset used for this study is the cybersecurity-attacks dataset from Kaggle open datasets [15]. The dataset was preprocessed and cleaned. The Naive Bayes Simulator in the Rapidminer Studio was used to predict the severity levels at various variation values. Feature extraction was performed on the data set and six attributes that were used as input values are: Traffic Type, Protocol, Log Source, Attack Type, Packet Length and Packet Type while the output was Severity Level. Each of the six inputs were classified into different categories as seen in the dataset by the naive bayes algorithm.

In modeling the dataset, data was first loaded into the rapid miner studio as seen in Figure 1.


<new process> – RapidMiner Studio Educational 10.3.000 @ DESKTOP-VI3O361
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Auto Model
Load Data Select Task Prepare Target Select Inputs Model Types Results
RESTART BACK NEXT
Recent Data Sets
Load Results
Select Data for a New Model
SELECT RESULTS FOLDER
IMPORT NEW DATA
Activate Windows
Go to Settings to activate Windows.
Type here to search
99+
30°C
12:45 PM
6/24/2024
2

Recent Data Sets

CyberAttacks3-new //Local Repository/CyberAttacks3-new

CyberAttacks3 //Local Repository/CyberAttacks3

Load Results

Naive1 //Local Repository/Naive1

Prediction2

Select Data for a New Model

- processes
- Cyber (12/13/23 4:22 PM – 180 kB)
- Cyber2Statistics (12/13/23 11:46 AM – 735 bytes)
- CyberAttack-Use (12/13/23 2:18 PM – 16 kB)
- CyberAttack-Use1 (12/14/23 2:17 PM – 16 kB)
- CyberAttacks2 (12/13/23 11:42 AM – 136 kB)
- CyberAttacks3 (12/14/23 4:34 PM – 136 kB)
- CyberAttacks3-new (2/7/24 12:55 PM – 16 kB)
- Temporary Repository (Local)
- DB (Legacy)

Information

Name: CyberAttacks3-new
Number of rows: 500
Number of columns: 9

Attributes / Columns

Protocol, Packet Length, Packet Type, Traffic Type, Anomaly Scores, Attack Type, Action Taken, Severity Level, Log Source

Figure 1: Snapshot of the Loaded Data Environment.

The initial data consists of 500 rows and 9 columns which are the attributes or parameters found in the dataset. They are: Protocol, Packet Length, Packet Type, Anomaly Score, Attack Type, Action Taken, Traffic Type, Severity level and Log Source. This is followed by selection of data where the output value, which is severity level was selected. In preparing the target output, the severity level was automatically classified into three which are: High

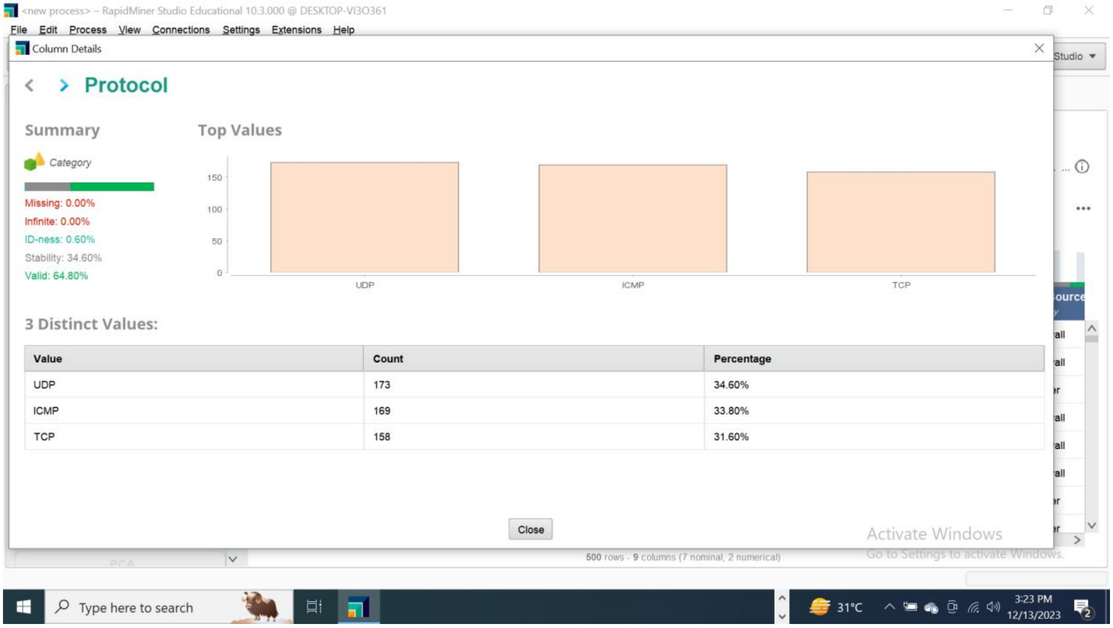
(191), Medium (158) and Low (151). This is followed by the selection of the input parameters. Six out of the overall nine were eventually selected for the modeling. This was guided by the level of importance and correlation shown. The six input variables are: Traffic Type, Protocol, Log Source, Attack Type, Packet Length and Packet Type. Each of the six input variables were also classified. Traffic Type was classified into DNS (178), FTP (171) and HTTP (151). Protocol was classified into UDP (173), ICMP (169) and TCP (158). Log Source was classified into Firewall (257) and Server (243). Attack Type was classified into DDoS (168), Malware (167) and Intrusion (165) while the Packet type was classifies into Control (253) and Data (247). Prediction was performed and results were generated.

4. DISCUSSION OF RESULTS

The Rapidminer studio was used to implement the Naive Bayes algorithm. Section 4.1 explains the classification of each of the input variables while section 4.2 explains the prediction results.

4.1 Discussion of Results for The Data Classification

Figure 2 shows the output for the attribute “Protocol”.


<new process> – RapidMiner Studio Educational 10.3.000 @ DESKTOP-VI3O361
File Edit Process View Connections Settings Extensions Help
Column Details
Studio
Summary
Category
Top Values
150
100
50
0
UDP
ICMP
TCP
Close
Activate Windows
Go to Settings to activate Windows.
Type here to search
31°C
3:23 PM
12/13/2023

Protocol

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.60%
Stability: 34.60%
Valid: 64.80%

3 Distinct Values:

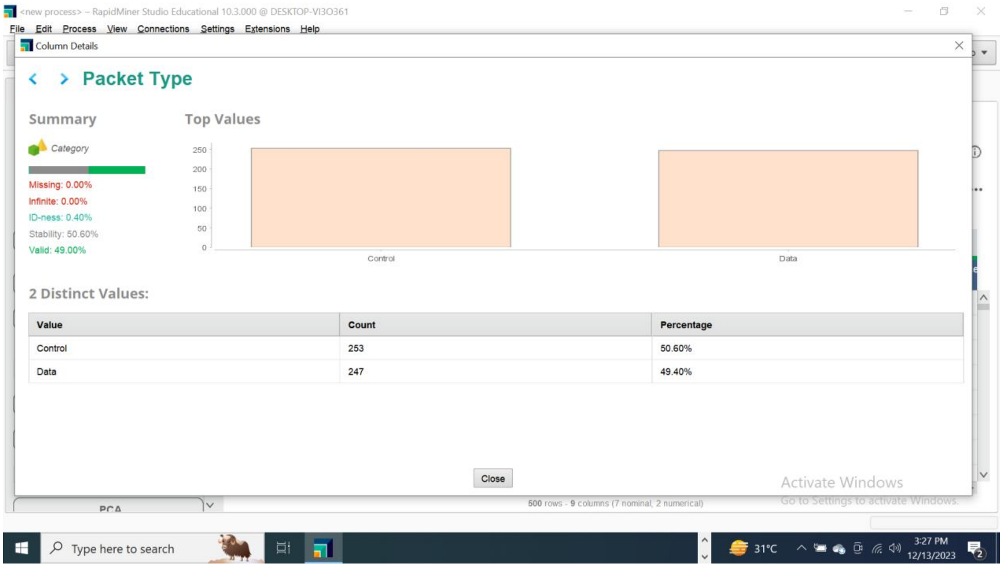
Value	Count	Percentage
UDP	173	34.60%
ICMP	169	33.80%
TCP	158	31.60%

500 rows · 9 columns (7 nominal, 2 numerical)

Figure 2: Result for Protocol

The protocol was classified majorly into three which are User Datagram Protocol (UDP), Internet Control Message Protocol (ICMP) and Transmission Control Protocol (TCP). The result shows that UDP is most widely used followed by ICMP while TCP has the minimum count. The distribution here does not imply the importance of the protocols as each often has specific roles, advantages and disadvantages, depending on the application and the network conditions. It simply shows the variation of usage in this dataset.

Figure 3 shows the outcome for the attribute “Packet Type”


<new process> – RapidMiner Studio Educational 10.3.000 @ DESKTOP-VI3O361
File Edit Process View Connections Settings Extensions Help
Column Details
Top Values
250
200
150
100
50
0
Control
Data
Close
Activate Windows
Go to Settings to activate Windows.
PCA
Type here to search
31°C
3:27 PM
12/13/2023

Packet Type

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.40%
Stability: 50.60%
Valid: 49.00%

2 Distinct Values:

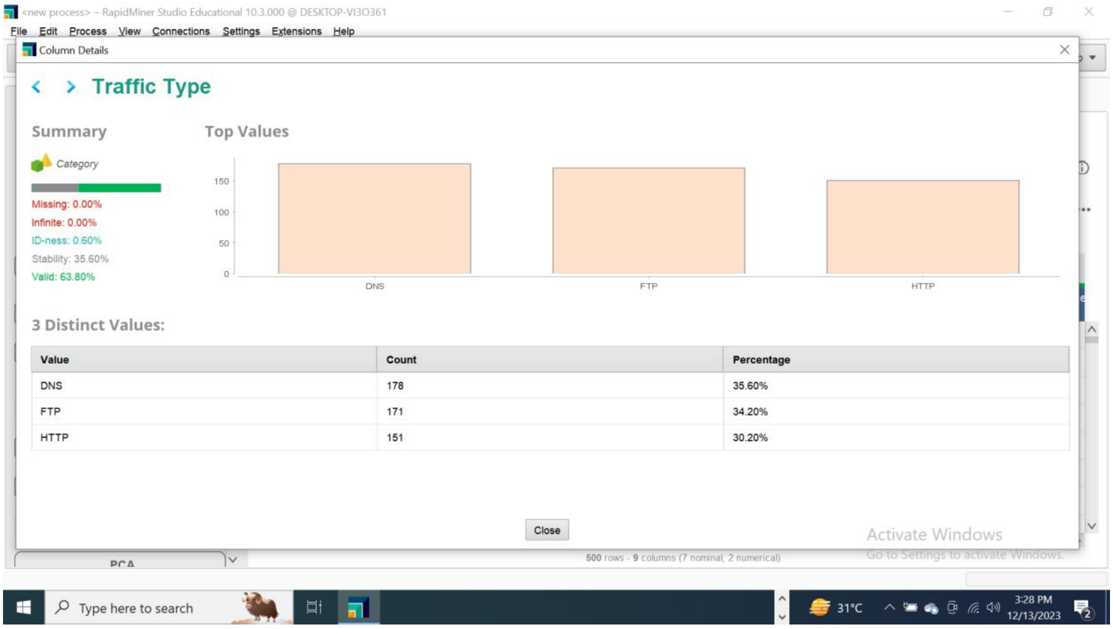
Value	Count	Percentage
Control	253	50.60%
Data	247	49.40%

500 rows - 9 columns (7 nominal, 2 numerical)

Figure 3: Result for Packet Type

The result shown in Figure 3 gives two major types of packets that are being used on the network. They are control and data packets. For any protocol used, the control plane is the one that determines how packets should be forwarded while the data plane is the one that actually forwards the packets.

Figure 4 shows the result for Traffic Type.


<new process> – RapidMiner Studio Educational 10.3.000 @ DESKTOP-VI3O361
File Edit Process View Connections Settings Extensions Help
Column Details
Top Values
150
100
50
0
DNS
FTP
HTTP
Close
Activate Windows
Go to Settings to activate Windows.
PCA
Type here to search
31°C
3:28 PM
12/13/2023

Traffic Type

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.60%
Stability: 35.60%
Valid: 63.80%

3 Distinct Values:

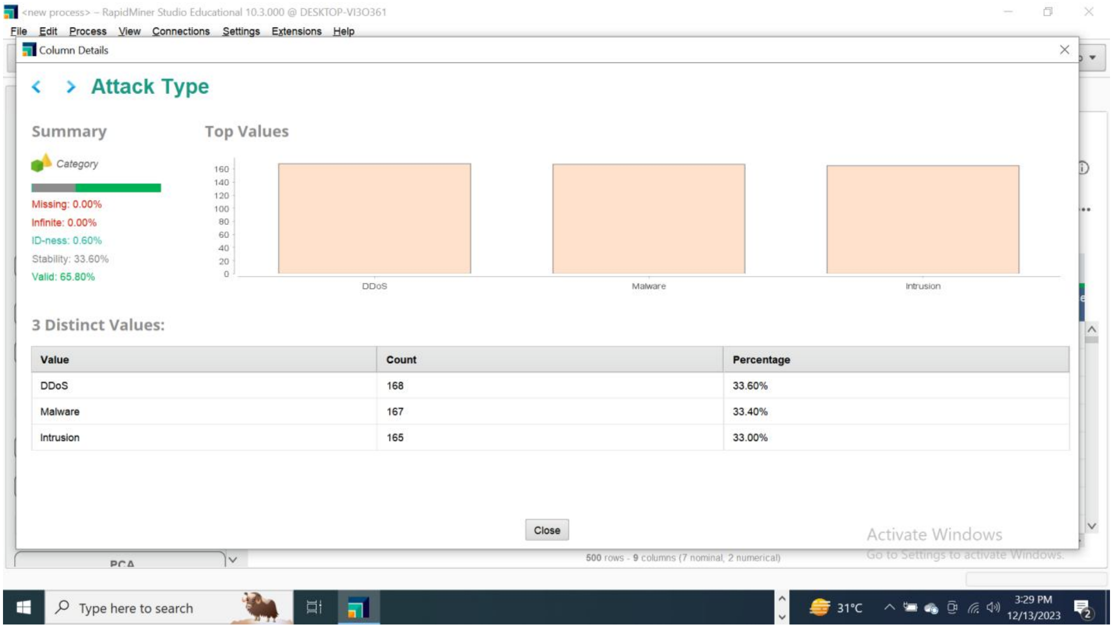
Value	Count	Percentage
DNS	178	35.60%
FTP	171	34.20%
HTTP	151	30.20%

500 rows - 9 columns (7 nominal, 2 numerical)

Figure 4: Result for Traffic Type

The traffic type attribute was classified majorly into three which are Domain Name System (DNS), File Transfer Protocol (FTP) and Hypertext Transfer Protocol (HTTP). Information is accessed on the internet through domain names which are translated to Internet Protocol (IP) addresses. The File Transfer Protocol is a standard network protocol used for the transfer of files from one host to another over a TCP-based network, such as the Internet while the Hypertext Transfer Protocol is used to load web pages using hypertext links.

Figure 5 shows the Attack Type that is being experienced on the internet. Three major cyber attacks were identified from the dataset. They are: Distributed Denial-Of-Service (DDoS), Malware and Intrusion. A distributed denial-of-service (DDoS) attack occurs when multiple systems flood the bandwidth or resources of a targeted system, usually one or more web servers. Malware is any program or file that is intentionally harmful to a computer, network or server while Intrusion is an unauthorized penetration of your enterprise's network, or an individual machine address in an assigned domain.


<new process> – RapidMiner Studio Educational 10.3.000 @ DESKTOP-VI3O361
File Edit Process View Connections Settings Extensions Help
Column Details
Close
PCA
Activate Windows
Go to Settings to activate Windows.
Type here to search
31°C
3:29 PM
12/13/2023
160
140
120
100
80
60
40
20
0
DDoS
Malware
Intrusion

Attack Type

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.60%
Stability: 33.60%
Valid: 65.80%

Top Values

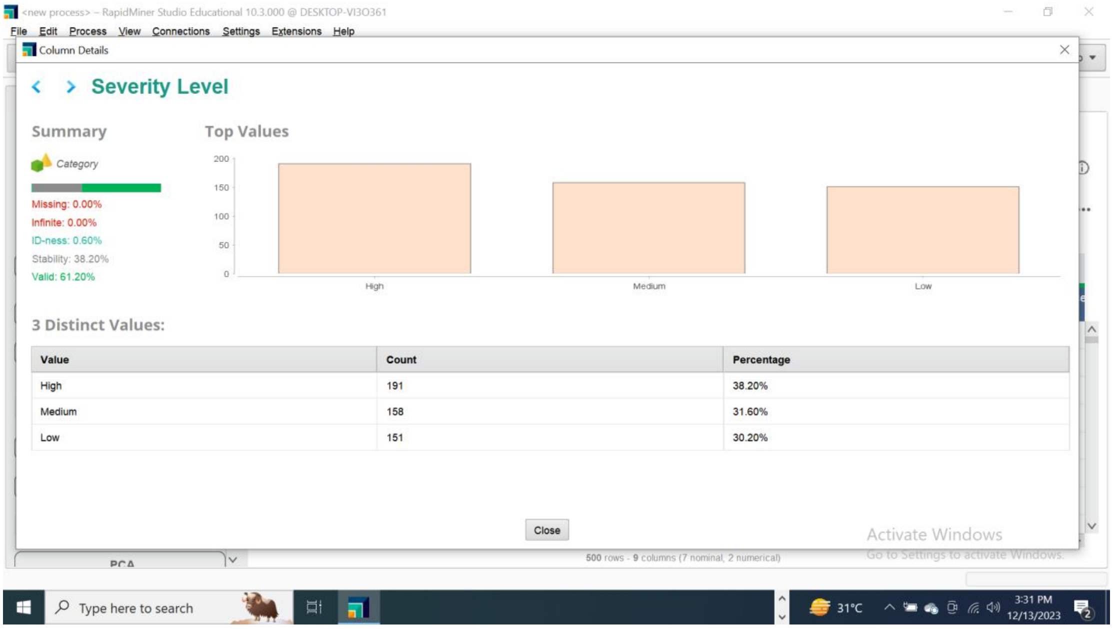
3 Distinct Values:

Value	Count	Percentage
DDoS	168	33.60%
Malware	167	33.40%
Intrusion	165	33.00%

500 rows - 9 columns (7 nominal, 2 numerical)

Figure 5: Result for Cyber Attack Types.

Figure 6 shows the result for severity level. Three major levels were detected and they are High, Low and Medium. From this result, it is obvious that the severity label “High” has the highest value. This simply indicates that the rate of cyber attacks are high.


<new process> – RapidMiner Studio Educational 10.3.000 @ DESKTOP-VI3O361
File Edit Process View Connections Settings Extensions Help
Column Details
Close
PCA
Activate Windows
Go to Settings to activate Windows.
Type here to search
31°C
3:31 PM
12/13/2023
200
150
100
50
0
High
Medium
Low

Severity Level

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.60%
Stability: 38.20%
Valid: 61.20%

Top Values

3 Distinct Values:

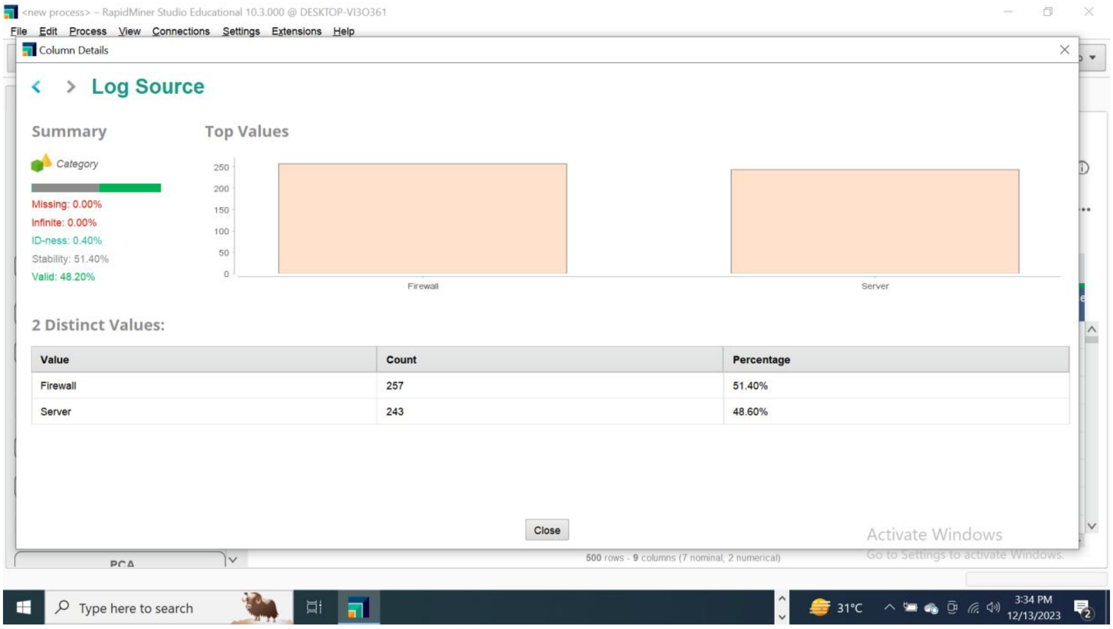
Value	Count	Percentage
High	191	38.20%
Medium	158	31.60%
Low	151	30.20%

500 rows - 9 columns (7 nominal, 2 numerical)

Figure 6: Result for Severity Level.

Figure 7 shows the Log Source classification.

Two major log sources that were identified in the dataset are Firewall and Server.


<new process> – RapidMiner Studio Educational 10.3.000 @ DESKTOP-VI3O361
File Edit Process View Connections Settings Extensions Help
Column Details
Top Values
250
200
150
100
50
0
Firewall
Server
Close
Activate Windows
Go to Settings to activate Windows.
PCA
Type here to search
31°C
3:34 PM
12/13/2023

Log Source

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.40%
Stability: 51.40%
Valid: 48.20%

2 Distinct Values:

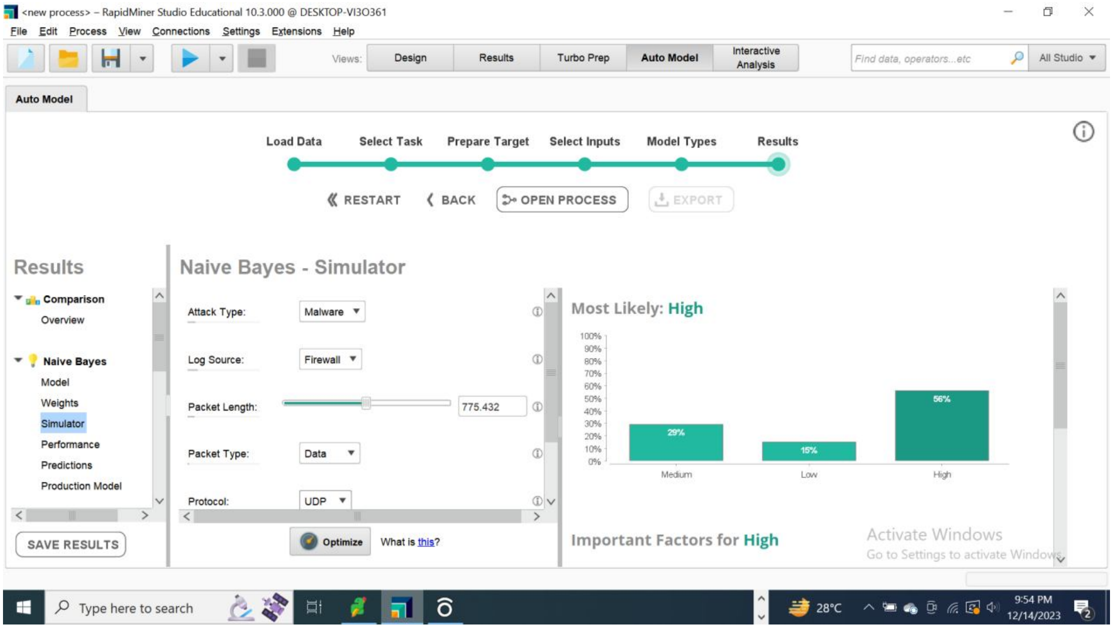
Value	Count	Percentage
Firewall	257	51.40%
Server	243	48.60%

500 rows · 9 columns (7 nominal, 2 numerical)

Figure 7: Result for Log Source

4.2 Discussion of Results for The Prediction

Figure 8 shows the Naive Bayes simulator predicted values for severity level. The severity level "High" has the highest value and hence, the prediction of Most Likely High.



Naive Bayes - Simulator

Attack Type: Malware

Log Source: Firewall

Packet Length: 775.432

Packet Type: Data

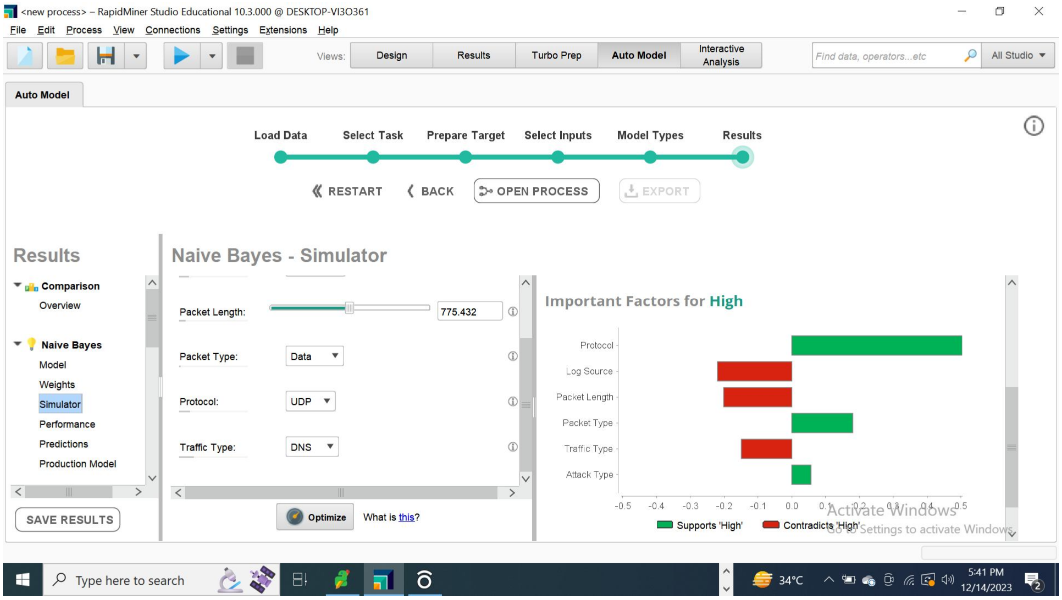
Protocol: UDP

Most Likely: High

Important Factors for High

Figure 8: Predicted Values for the Severity Levels

Figure 9 shows the factors responsible for the outcome of the High prediction for the severity level.



Packet Length: 775.432

Packet Type: Data

Protocol: UDP

Traffic Type: DNS

Figure 9: Important Factors for High Severity Level

The important factors that responsible for the High prediction of the severity level are Protocol, Packet Type and Attack Type.

Figure 10 shows the various predicted values for severity value as output and the six inputs values which are Traffic Type, Protocol, Log Source, Attack Type, Packet Length and Packet Type.

Predicted Severity Level Values					
	Attributes		**High (%)**	**Low (%)**	**Medium (%)**
1	Attack Type	Malware	56	15	29
		DDoS	54	16	29
		Intrusion	49	21	30
2	Log Source	Firewall	56	15	29
		Server	62	11	28
3	Packet Type	Control	53	18	30
		Data	56	15	29
4	Protocol	UDP	56	15	29
		ICMP	40	28	32
		TCP	**34**	34	32
5	Traffic Type	DNS	56	15	29
		FTP	47	23	30
		HTTP	67	7	26
6	Packet Length	775.432			

Figure 10: Result for the Predicted Values for Severity Level

Varying the different metrics under the six attributes gave different values for the High, Low and Medium severity levels. Here, it is seen that the feature that gives the minimal percentage for High severity level is TCP while the one that gave the highest level is HTTP. The predicted value with the TCP simply confirms the fact that it is more reliable as seen in the literature [16], but slower and more complex than UDP while the UDP is faster and simpler but less reliable and more prone to errors. ICMP is not used for data transmission, but for network management and troubleshooting. Also, the predicted value for HTTP confirms the fact that it is less secure. Possibly if it is hypertext transfer protocol secure (HTTPS), the value would not have been as high as we have it here. Further studies can be done for HTTPS to ascertain this as seen in the literature [17].

5. CONCLUSION

In conclusion, it is evident from this study that cyber attacks are on the high side as both raw and predicted data indicated High for the security level. Hence, security measures must be put in place in order to mitigate against cyber attacks. Also, users of the internet should work proactively to protect their data on the network by using more secured resources.

REFERENCES

[1]. Tpoint Tech. (2023). Machine Learning Tutorials. (Retrieved from https://www.javatpoint.com/machine-learning on 14 December, 2023)

[2]. I.T Ayorinde, and A. O. Osofisan (2010). Application of Artificial Neural Networks in lassifying Medical Database. Journal of Science Research Vol. 9: 12-18.

[3]. O. Osunade, I. T. Ayorinde and B. I. Ayinla (2024). Knowledge Discovery in Research Security Practices among Scientists Using Machine Learning Techniques (A Case Study of Faculty of Science, University of Ibadan). *International Journal of Computer Applications (IJCA).* Vol. 186. No. 9: 19-28.

[4]. A. S. Gillis, and M.K. Pratt (2023) Cyber Attack. An Online Tutorial (Retrieved from https://www.techtarget.com/searchsecurity/definition/cyber-attack on April 12, 2024).

[5]. I. Andersen (2023). The 12 Most Common Types of Cyber Security Attacks Today (Retrieved on December 13, 2023 from https://blog.netwrix.com/types-of-cyber- attacks)

[6]. J. Mackay (2023). 5 Damaging Consequences Of Data Breach: Protect Your Assets. (Retrieved on 13 December, 2023 from https://www.metacompliance.com/blog/data-breaches/5-damaging-consequences-of-a-data-breach)

[7]. UK Government Survey. (2023). Cyber security for business: Impact of cyber attack on your business (Retrieved on 12 December, 2023 from https://www.nibusinessinfo.co.uk/content/impact-cyber-attack-your-business)

[8] K. Nanda-Kumar and B. H. Alamma (2022). Data mining using naive bayes in e- commerce. Indian Scientific Journal Of Research In Engineering And Management, 06(06) doi: 10.55041/ijsrem14245

[9]. N. R. Nayak (2020). Application of Naive Bayes Classifier for Information Extraction. Center for Open Science. 10.31219/osf.io/z7q2e.

[10]. S. Budiyanto and I. Pratama (2020). Classification of Network Status in Academic Information Systems using Naive Bayes Algorithm Method. 2nd International Conference on Broadband Communications, Wireless Sensors and Powering (BCWSP), Yogyakarta, Indonensia, 2020, pp. 107-112.doi: 10.1109/BCWSP50066.2020.9249398

[11]. T. Almanie (2023). Quantitative Study of Traffic Accident Prediction Models: A Case Study of Virginia Accidents. *Int. J. Advanced Networking and Applications.* Volume: 14 Issue: 05 Pages: 5582 - 5589 (2023) ISSN: 0975-0290

[12]. Y. Huang (2022)."Network Intrusion Detection Method Based on Naive Bayes Algorithm," 2022 6th Asian Conference on Artificial Intelligence Technology (ACAIT), Changzhou, China, 2022, pp. 1-10, doi:10.1109/ACAIT56212.2022.10137846.

[13]. K. R. Nataraj R. Rakshitha, Trupti, K. Monika (2022). Detection of Pervasive Developmental Disorder (PDD) Using Naïve Bayes Neural Network Algorithm. In: Kumar, A., Ghinea, G., Merugu, S., Hashimoto, T. (eds) Proceedings of the International Conference on Cognitive and Intelligent Computing. Cognitive Science and Technology. Springer, Singapore. https://doi.org/10.1007/978-981-19-2350-0_44

[14]. L. Y. Lau, K. S. Ng, K. W. Kwok, K. M. Tsia, C. F. Sin, C. W. Lam, and V. Vardhanabhuti (2022). An Unsupervised Machine Learning Clustering and Prediction of Differential Clinical Phenotypes of COVID-19 Patients Based on Blood Tests—A Hong Kong Population Study. *Front. Med.* 8:764934. doi: 10.3389/fmed.2021.764934.

[15]. Kaggle Dataset link (https://www.kaggle.com/datasets)

[16]. B. Gorman (2023). Blog Post on TCP vs UDP: Differences between the protocols (https://www.avast.com/c-tcp-vs-udp-difference) - Retrieved 24 June, 2024)

[17]. Cloudflare, Inc.(2024).Blog Post on HTTP Vs. HTTPS (https://www.cloudflare.com/learning/ssl/why-is-http-not-secure/ - Retrieved 24 June, 2024)