People's Democratic Republic of Algeria

Ministry of Higher Education & Scientific Research

20 Aout 1955 University of Skikda

Faculty of Sciences

Department of Computer Science

End of Studies Project's Report for Obtaining Professional Master's Degree in Computer Science 2022

Specialty: Network and Distributed Systems

TOPIC

REINFORCING INTRUSION DETECTION SYSTEM USING MACHINE LEARNING

Authored by :

Tadjine Yahya

Djebablah Douaa

Supervised by :

Touil Ghassen

Acknowledgments

First of all we would like to thank God Almighty who gave us strength and patience to accomplish this humble work.

We would like to take this opportunity and extend our sincere thanks and gratitude to all those who helped us from near or far in achieving this thesis.

We thank Mr. Touil Ghassen for guidance, assistance and encouragement. Thanks to his advice and guidance, we were able to complete this work.

We also add this thanks to the members of the jury for their interest and interest in this work.

We would also like to express my sincere thanks to all the teachers of the Computer Science Department who contributed to our training.

Finally, and above all, we extend our sincere thanks to all our family, especially our mothers, who have always encouraged us to continue our studies, as well as for their help, understanding and support without forgetting to thank our friends for their support and good humor while preparing this letter.

For fear of forgetting someone, we heartily thank all those we owe her.

ABSTRACT

The increasing of security attacks and unauthorized intrusion have made network security one of the main subjects that should be considered in present data communication environment. Intrusion detection system is one of the suitable solutions to prevent and detect such attacks. This project aims to study for Machine learning techniques have been massively applied in an open source java project for network intrusion detection. The complete NSL-KDD dataset is used for training and testing data. Number of different experiments have been done. Java programming language is used for system implementation.

ملخص

إن أمنية الشبكات هي إحدى المواضيع الرئيسية التي يجب الاهتمام بها في بيئة الاتصالات الحالية بسبب زيادة التهديدات الأمنية والتطفلات الغير المخولة. نظم كشف التطفل هي إحدى الحلول المناسبة لصد و كشف هذه الهجمات. يهدف البحث الى دراسة نظام كشف تطفل شبكي (NIDS) يعتمد على تعلم الالة، تم استخدام تقنياتها في مشروع مفتوح المصدر مطور بلغة الجافا اعتماد مجموعة بيانات NSL-KDD كاملة في تدريب واختبار البيانات

CONTENT TABLE

Title	Page
General introduction	1

Chapter 1: CyberSecurity

1- Introduction	2
2- Cybersecurity	2
2.2- Disadvantages of Cyber Security	3
2.3- Advantages of Cybersecurity	3
2.4- Types of Cybersecurity	3
2.5- Principles of Information Security	4
2.6- Importance of Cyber Security	4
2.7- The Benefits of Cybersecurity	4
3- Cyber Attack	5
3.1- Types of attacks	6
3.1.1- Network attack	6
3.1.2- Application attack	6
3.1.3- Denial-of-Service (DoS)	7
3.1.4- Data attacks	7
4- Common Cyberattacks	7
4.1- TCP SYN Flood Attacks	8
4.2- Smurf Attacks	8

4.3- Botnets	9
4.4- Social engineering	10
4.5- Phishing Attacks	10
4.6- Man-in-the-Middle (MitM) Attacks	11
4.7- Cookie Theft	11
4.8- Session Hijacking	12
4.9- Malware Attacks	12
4.10- Logic Bombs	12
4.11- Trojans	12
4.12- SQL injection Attacks	13
4.13- Click jacking	13
5- Information Security Governance	13
6- Information Audit	14
6.1- Risk Assessment	15
6.2- Types of Audits	16
6.2.1- Security Review	16
6.2.2- Security Assessment	16
6.2.3- Security Audit	17
6.3- Security Auditing Tools	17
6.3.1- Service Mapping Tools	17
6.3.2- Vulnerability Assessment Tools	18

6.3.3- Packet Capture Tools	18
6.3.4- Penetration Testing Tools	19
7- Technical control	19
8- Difference between Firewall and Intrusion Detection System	20
Conclusion	21

Chapter 2: Intrusion Detection Systems

1- Introduction	22
2- Definition of Intrusion	22
3- Definition of an intrusion Detection system	22
4- How does an IDS work?	23
5- Types of intrusion detection systems	24
5.1-NIDS (Network-based Intrusion Detection System)	24
5.2- HIDS (Host-based Intrusion Detection System)	26
5.3- Hybrid IDS	26
6- Characteristics of IDS	27
7- The architecture of an IDS	27
7.1- Sensor	27
7.2- Analyzer	28
7.3- Manager	28
8- Setting up an IDS	28
8.1- The positioning of the IDS	28

9- Criteria for testing an IDS	29
10- Mode of operation	30
10.1- Detection of anomalies	30
10.2- Recognition of Signatures	30
11- Active and passive responses	32
11.1- Passive response	32
11.2- Active response	32
12- Methods of detection	33
12.1- The behavioral approach (Anomaly Detection)	33
13- Anti-IDS Techniques	33
14-The limits of the IDS	34
15- Security Experts	34
Conclusion	35

Chapter 3: Machine learning techniques

1- Introduction	36
2- Machine Learning	36
3- Supervised Learning	37
4- WEKA Environment	37
5- Classifiers	38
5.1- Introduction	38
5.2- Learning Algorithms	39

6- Decision trees	39
6.1- Definition	39
6.2- Algorithm	42
6.3- Criticisms of the method	42
6.4- Fields of application	42
7- J48 algorithm	43
8- Random Forests	44
8.1- Advantages of Random Forests	44
8.2- Disadvantages of Random Forests	45
9- Voting	45
9.1- Benefits of Voting	45
9.2- Drawbacks of voting	46
10- Support Vector Machines (or VSM)	46
Conclusion	49

Chapter 4: Learning and Testing

1- Introduction	50
2- NSL-KDD Dataset	50
2.1- BASIC FEATURES OF EACH NETWORK CONNECTION VECTOR	50
3- Software	53
4- Experimentation	55
4.1- Use Testing set	55

4.2- Cross validation	56
5- Testing with J-48	57
5.1- J48 results	58
6- Combining VSM and Random Forest	59
6.1- Result of combining VSM and RF	60
7- Discussion of results	61
Conclusion	61

General Conclusion	**62**
Bibliography	**63**

LIST OF FIGURE

Figure 1.1: All in cybersecurity.	5
Figure 1.2: Top-network-attack-types-2016-chart-copy.	6
Figure 1.3: Denial-of-Service (DoS)	7
Figure 1.4 : Common-cyber-threats-cybersecurity-interview-questions	8
Figure 1.5: TCP SYN Flood Attacks	9
Figure 1.6: Smurf-attacks-simplified	9
Figure 1.7: How a botnet works –steps-.	10
Figure 1.8: Phishing Attacks steps	11
Figure 1.9 : Cookie Theft steps.	12
Figure 1.10: SQL injection Attack (SQLI).	13
Figure 1.11: Information Security Gouvernance Infrastructure	14
Figure 1.12: IT governance and IT audit.	15
Figure 2.1: Some IDS implementation approaches	23
Figure 2.2: What does an IDS do	24
Figure2.3: NIDS with its components.	25
Figure 2.4: HIDS (Host Intrusion Detection Systems).	26
Figure 2.5: IDS architecture	27
Figure 2.6: The principle of detection.	28
Figure 2.7: The position of IDS.	29
Figure 3.1: Machine learning environment	36
Figure 3.2: Types of machine learning	37

Figure 3.3: Weka environnement.	38
Figure 3.4: Steps-carried-out-in-the-machine-learning-process	39
Figure 3.5: The decision tree.	40
Figure 3.6: Example of the decision tree	41
Figure 3.7: The decision tree algorithm.	42
Figure 3.8: J48 filter (classifier).	43
Figure 3.9 : Random Forests Simplified	44
Figure 3.10: cCse of text categorization with VSM.	47
Figure 3.11: Soft Margin VSM (Cortes and Vapnik 1995).	48
Figure 4.1: Main window interface	53
Figure 4.2: OptionHAN interface	54
Figure 4.3: OptionNAN interface	54
Figure 4.4: OptionGEN interface	55
Figure 4.5: Testing set approach	56
Figure 4.6: Cross-validation approach	57
Figure 4.7: Classification J-48 results	57
Figure 4.9: Decision tree for J-48	59

LIST OF TABLES

Table1.1: Difference between Firewall and Intrusion Detection System	21
Table 2.1: Advantages and Disadvantages between IDS types	31
Table 2.2: True positives and negatives.	32
Table 4.1: NSL-KDD basic features	52
Table 4.2: The result of classification with J-48	58
Table 4.3: The result by class of classification with J-48	58
Table 4.4: The result of classification using voting between VSM and RF	60
Table 4.5: The result by class of classification using voting between VSM and RF	61

GENERAL INTRODUCTION

General introduction

Cybersecurity is a sweeping the world by storm, with a number of the greatest & most advanced companies in the world falling victim to cyber-attacks in just the most recent 5 years. Against that backdrop, truly personal and vulnerable information like social security numbers were definitely consumed by the Equifax hack, affecting much more than 145 million people.

When the network system suffers from attacking, it relies on some devices like Intrusion Detection System (IDS) to monitor the network behaviors and to analyze the attack methods. By the report from IDS, system administrators are able to fix the insufficiencies and to tighten the security of the system.

This report discusses the research done on the chosen topic, which is Network Intrusion Detection System. In this project we present a study for Machine learning techniques have been massively applied in an open source java project for network intrusion detection.

The problem of this project is an unauthorized access and an attack into a networking system that may cause harm by stealing private and confidential information as firewall and anti-virus won't be sufficient against a determine attacker.

The objective is to show which classifier is best suited to the data of the KDD 20persent intrusion database. To do so, we will test three classifiers with supervised learning, namely the J-48 classifier and the combining between Random Forest and VSM classifier. The considered dataset is a set of network traffic data sequences, labeled according to the type of attack it conveys.

This report is organized as follows:

Chapter 1 is about Cybersecurity essentially definitions of cyber security and presents the different types of attacks.

Chapter 2 presents the interest of intrusion detection systems and their architecture.

Chapter 3 introduces Machine Learning techniques and the tools we used in the testing.

Chapter 4 presents the software we used and the dataset and the classifiers used for learning and testing along with the results

Chapter 1

CYBERSECURITY

1- Introduction

Cybersecurity will be the shelter of internet-connected strategies, such as hardware, software, and data from cyberattacks. In a computing context, security comprises actual physical protection and cybersecurity – both are employed by businesses to guard against unauthorized access to info centers and other computerized systems. The goal of cybersecurity is restricting risk as well as protect IT home from attackers with malicious intent. Info security, designed to always keep the confidentiality, integrity, and availability of advice is a subset of cybersecurity. Throughout this chapter, our focus will be on the main threats to cybersecurity as well as the defense mechanisms and technique of hackers.

2- Cybersecurity

The Internet age has produced a lot of jargon. One of those is the term "cybersecurity." The following are a few of the current examples of dictionary attempts to define cybersecurity: The state of being protected against the criminal or unauthorized use of electronic data, or the measures taken to achieve this (Oxford English Dictionary)

Measures taken to protect a computer or computer system (as on the Internet) against unauthorized access or attack (Merriam-Webster) The body of technologies, processes, and practices designed to protect networks, computers, programs, and data from attack, damage, or unauthorized access (WhatIs.com) Refers to preventative methods used to protect information from being stolen, compromised, or attacked (Technopedia) The concept of cybersecurity can be applied in various contexts, from general business operations to firewall technologies, but it can be divided into a few general categories.

- **Information security**: to protect and secure the privacy and integrity of data at rest or at movement.
- **Network security**: to secure a computer network from bad actors that might be a targeted attack or malicious malware.
- **Operational security**: to create and maintain the processes, procedures and decision making for treatment and protecting data assets.
- **Application security**: to concentrate on maintaining the safety of software and devices clear of threats.
- **Business continuity and disaster recovery:** to decide how an organization responds to a cybersecurity incident or breach of data. These are the policies and procedures that

dictate how the organization reestablishes control of its operations and information to the same level prior to the event because resources may be lacking post event.

- **Risk Management:** to manage organizational risk in the company's information security program itself, which offers an operative framework for setting the risk appetite and security controls for systems.
- **Security Awareness training:** to address the education of people who often cause security vulnerabilities based on their actions or lack thereof. People can unintentionally introduce a virus or malware to an otherwise secure system if they are not knowledgeable of security best practices, such as deleting suspicious attachments in emails, refrain from inserting unidentified USB drives, etc.

2.2- Disadvantages of Cyber security

- Strict regulations.
- Hard to deal with for nontechnical users.
- Restrictive to resources.
- Constantly needs patching.
- Constantly being attacked.

2.3- Advantages of Cybersecurity

- Protection from malicious attacks on your system.
- Deletion or perhaps guaranteeing malicious parts within a preexisting group.
- Prevents owners from unauthorized access to the product.
- Denies uses from particular resources which might be infected.
- Securing confidential info.

2.4- Types of Cybersecurity

- Critical infrastructure security.
- Application security.
- Network security.
- Cloud security.
- Internet of things (IoT) security.
- Information security.
- Disaster recovery.

- Website security.
- Endpoint security.

2.5- Principles of Information Security

● **Confidentiality:** Information is only available to the people or systems that need access to it. This is done by encrypting information that only certain people are able to decrypt or denying access to those who don't need it. This might seem simple at first, but confidentiality must be applied to all aspects of a system. This means preventing access to all backup locations and even log files if those files can contain sensitive information.

● **Integrity:** Information can only be added or updated by those who need to update that data. Unauthorized changes to data cause it to lose its integrity, and access to the information must be cut off to everyone until the information's integrity is restored. Allowing access to compromised data will cause those unauthorized changes to propagate to other areas of the system.

● **Availability:** The information needs to be available in a timely manner when requested. Access to no data is just as bad as access to compromised data. No process can be performed if the data on which the process is based is unavailable.

2.6- The Importance of Cyber Security

Cyber security is important because government, military, corporate, financial, and medical organizations collect, process, and store unprecedented amounts of data on computers and other devices. A significant portion of that data can be sensitive information, whether that be intellectual property, financial data, personal information, or other types of data for which unauthorized access or exposure could have negative consequences. Organizations transmit sensitive data across networks and to other devices in the course of doing business, and cyber security describes the discipline dedicated to protecting that information and the systems used to process or store it. As the volume and sophistication of cyber-attacks grow, companies and organizations, especially those that are tasked with safeguarding information relating to national security, health, or financial records, need to take steps to protect their sensitive business and personnel information. As early as March 2013, the nation's top intelligence officials cautioned that cyber-attacks and digital spying are the top threat to national security, eclipsing even terrorism.

2.7- The Benefits of Cybersecurity

Benefit 1 – Protect the Standing of Yours

In numerous cases, on account of the public nature of the site of yours, a thriving breach is going to be difficult to conceal from the buyers of yours, potential customers, and associates. Sometimes hackers will bring your entire website down, while, in others, you may be made to do this yourself to limit the exposure and the destruction of yours.

Benefit 2 – Protect Your Information

If an assault leads to private info being stolen, including username, passwords, or perhaps payment details, airers4you in question will be made owning almost as the breach. To protect the personal information of yours and, far more particularly, your clients' info, by deploying strong, robust, and regularly updated cybersecurity measures, can be seen as an enterprise priority. To guard the information of yours, while your rivals do not do so, it may elevate your condition and act the hassle of differentiation.

Benefit 3 – Protect Your Account Balance

A single, highly effective cyber-attack can possess an amazing impact on the business organization's bank harmony of yours. A number of the financial hits are going to be experienced promptly. In contrast, others involve a great deal more being experienced, which helps make the actual costs of good cyberattacks difficult to totally quantify.

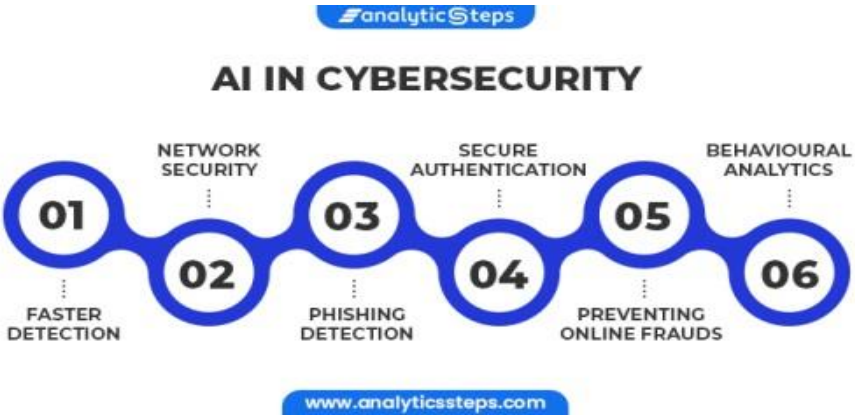

Figure 1.1: All in cybersecurity.

3- Cyber Attack

When it comes to computers and networks, attacks largely attempt to expose, disable, alter, steal, destroy, or gain access to the use of assets. Any such offensive maneuver that tries to target computer systems, networks, infrastructures, and personal devices can be called a cyber-attack. A

cyber-attack can be carried out by a single individual or a group with malicious intent putting hardware, data, and functionalities at risk. Cyber-attacks can be categorized as cyber warfare or cyber-terrorism, depending on the nature of the attack and its threat.

Furthermore, cyber-attacks can be carried out by various societies and sovereign states operating in anonymity. During a cyberattack, susceptible systems and devices are hacked into while achieving the attacker's malicious intents or the attackers. A cyberattack scale may also vary from a single computer, device, individual, or company is the primary target to infrastructures of entire nations. [1]

3.1- Types of attacks

There exists 04 essentially types of attacks

3.1.1- Network attack

A network attack is an attempt to gain unauthorized access to an organization's network, with the objective of stealing data or perform other malicious activity. There are two main types of network attacks:

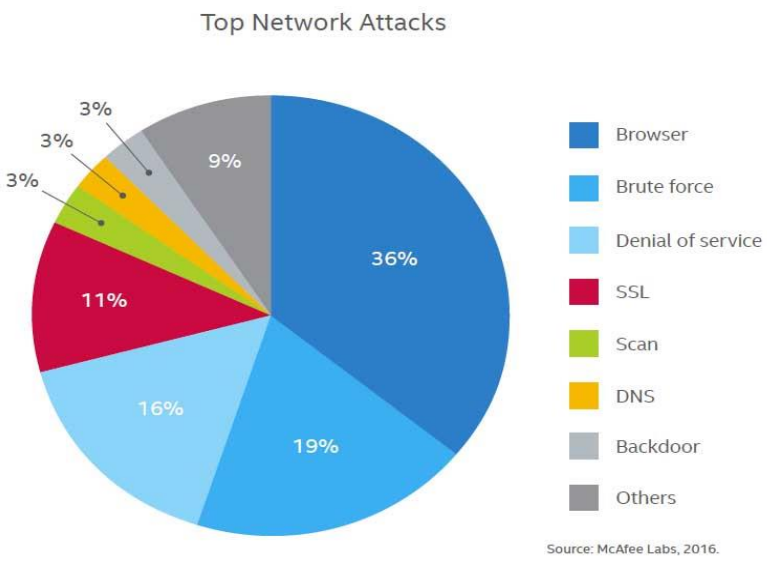

Figure 1.2: top-network-attack-types-2016-chart-copy.

Passive: Attackers gain access to a network and can monitor or steal sensitive information, but without making any change to the data, leaving it intact.

Active: Attackers not only gain unauthorized access but also modify data, either deleting, encrypting or otherwise harming it.

3.1.2- Application attack

Application attacks are based on flaws in the programs used, or even configuration errors. However, as before, it is possible to classify these attacks according to their origin.

3.1.3- A Denial-of-Service (DoS)

DoS attack is an attack meant to shut down a machine or network, making it inaccessible to its intended users. DoS attacks accomplish this by flooding the target with traffic, or sending it information that triggers a crash. In both instances, the DoS attack deprives legitimate users (i.e. employees, members, or account holders) of the service or resource they expected.

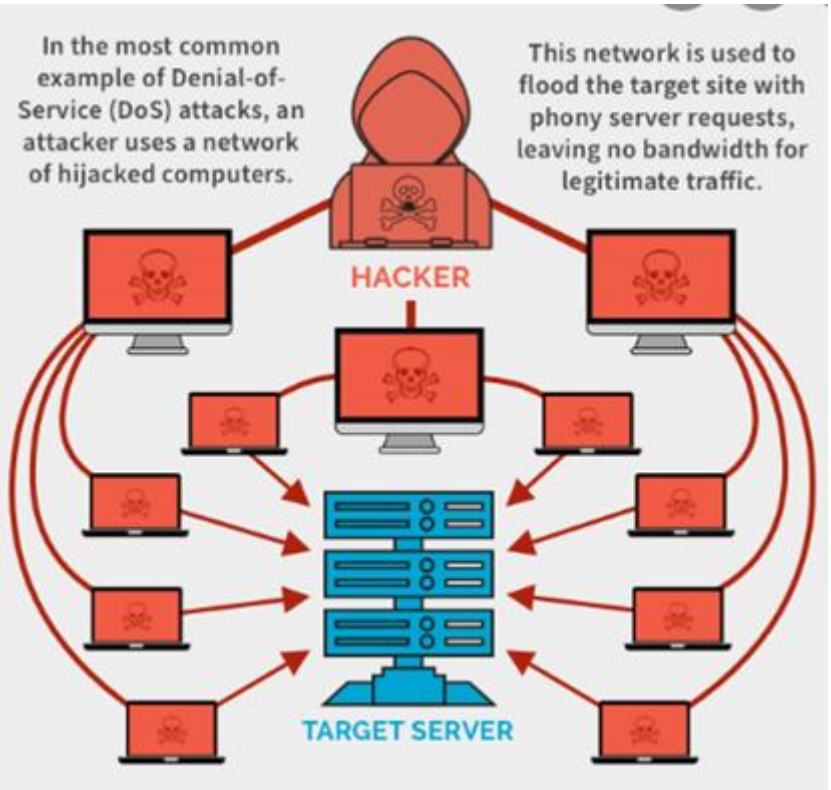

Figure 1.3: A Denial-of-Service (DoS)

3.1.4- Data attacks

The data transported by the application protocol can constitute a threat to the integrity of the system that receives them. The main attacks of this type, we find: virus, worm, Java Applet, Trojans…designated by malicious codes or Malware.

4- Common Cyberattacks

Figure 1.4: Common-cyber-threats-cybersecurity-interview-questions-

4.1- TCP SYN Flood Attacks

Attackers utilize the buffer space in the Transmission Control Protocol or TCP session initialization handshake in this type of attack. The target system is flooded with connection requests by the attacker's device. However, the attacker's device does not respond to the target system's replies to its requests. As a result, the target computer system times out upon waiting for responses from the attacker's device, leading it to become unusable or even crash due to the filling up of connection queues. [2]

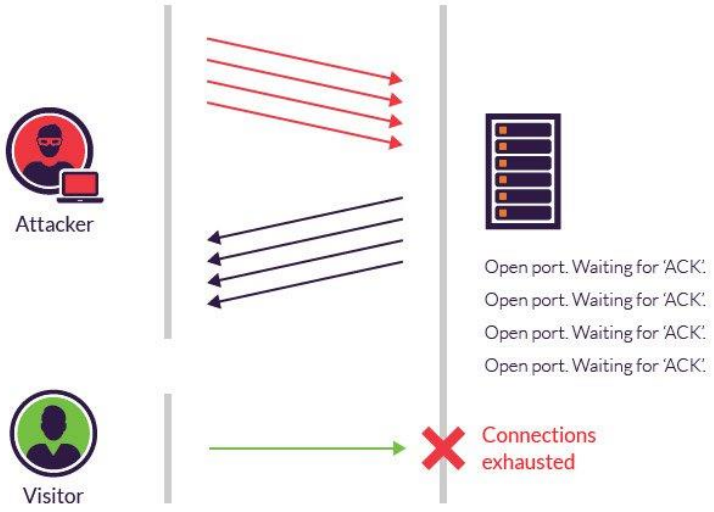

Figure 1.5: TCP SYN Flood Attacks

4.2- Smurf Attacks

The target network is saturated with traffic by the attackers using ICMP and IP spoofing in a Smurf attack. ICMP echo requests are used in this method directed at broadcast IP addresses. The attacker creates spoof ICMP echo requests from a spoofed IP address giving the broadcast IP addresses the impression that the target victim is sending the request.

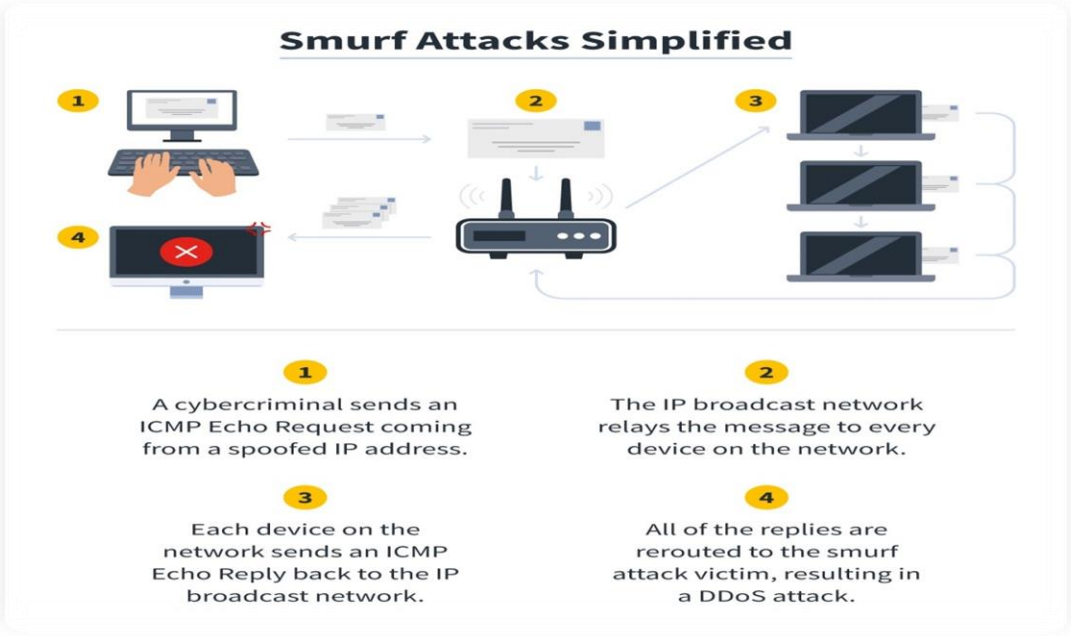

Figure 1.6: Smurf-attacks-simplified

Multiple requests are sent to broadcast IP addresses, and the target address receives multiple responses that overwhelm it. Spoofed ICMP echo requests are automated and generated in bulk by the attacker, causing serious network congestion that forces the target network to deny services.

4.3- Botnets

Hackers infect millions of systems and gain control over them to be used as bots to launch DoS attacks. The target system or network is attacked using these bots or zombie systems under the control of the attacker. Attackers often use botnets to overwhelm the bandwidth of the target system and its processing capabilities. DoS attacks are usually difficult to trace since they use infected bots or computer systems located in numerous geographic locations. [3]

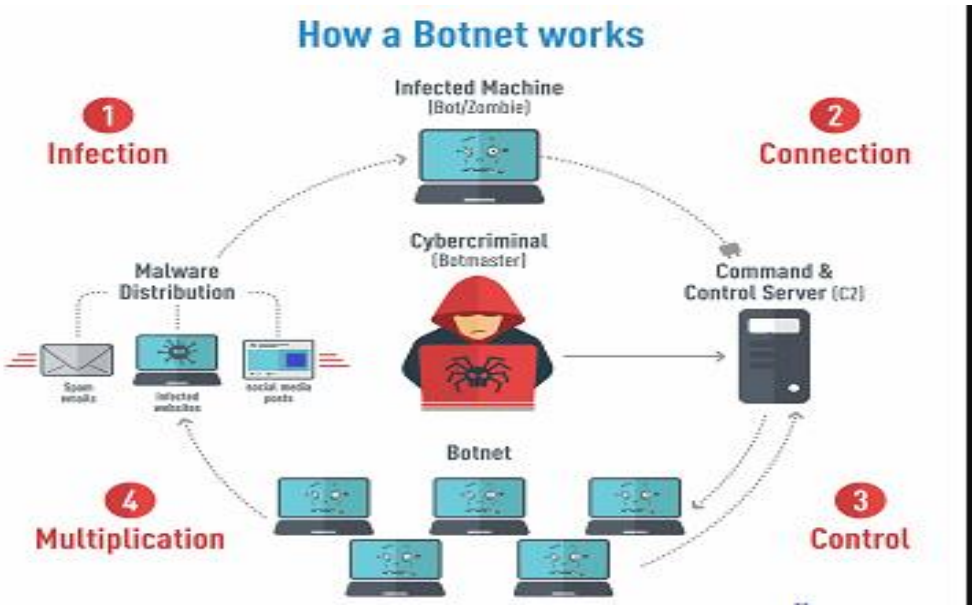

Figure 1.7: How a botnet works –steps-.

4.4- Social engineering

It is the art of manipulating people through personal interaction to gain unauthorized access to something.

4.5- Phishing Attacks

These cyber-attacks are carried out by sending highly realistic emails that seem to be from trustworthy sources to gain sensitive personal information from the receivers or influence them to carry out an action the attacker's desire. Therefore, phishing is a combination of technical trickery and social engineering.

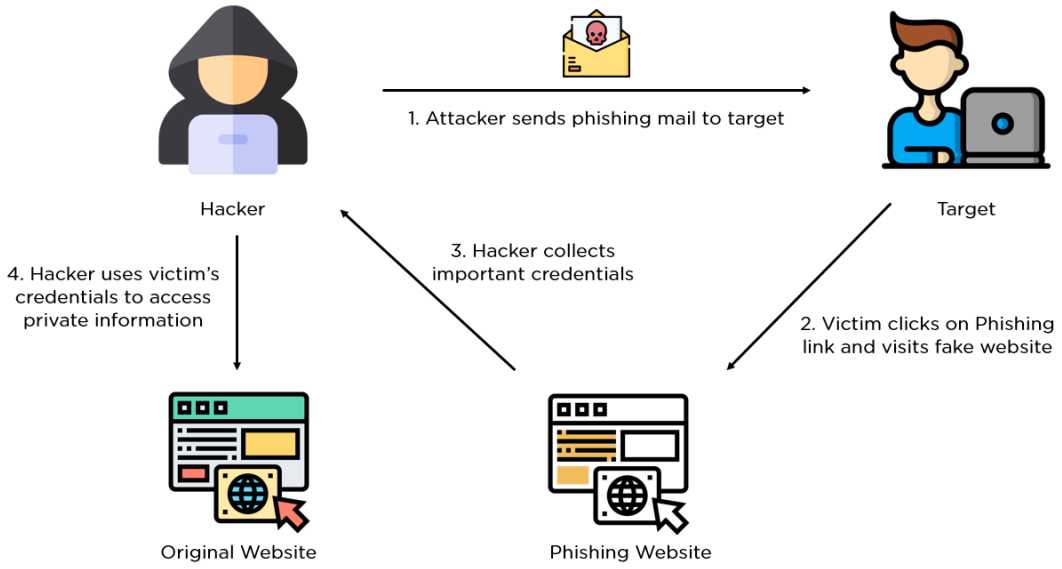

Figure 1.8: Phishing Attacks steps

4.6- Man-in-the-Middle (MitM) Attacks

These attacks place the attacker between clients and servers' communication, enabling them to interfere and tamper with those communications, including requests and responses. Various types of MitM cyber-attacks exist.

4.7- Cookie Theft

Websites use cookies to modify your browsing experience to make it tailored to your needs as well as for proper ad placement. Cookie thefts are used by hackers to gain access to that information. Cookies are one of the most natural methods of hacking, they can be stolen through public Wi-Fi networks

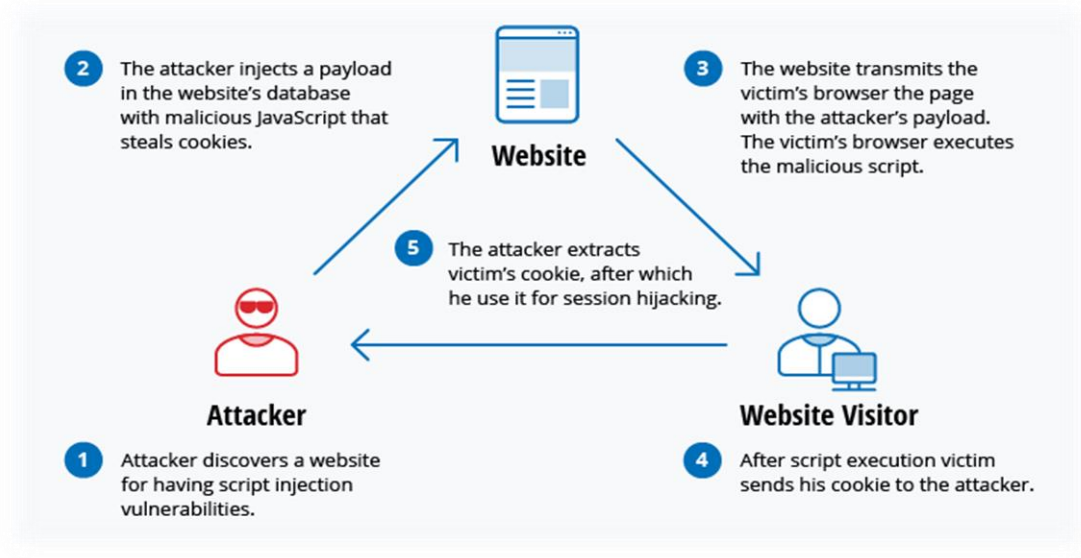

Figure 1.9: Cookie Theft steps.

4.8- Session Hijacking

Attackers launch these attacks by hijacking sessions between trusted clients and network servers. The attacking computer uses the trusted client's IP address while the server proceeds with the session under the impression that it is communicating with the client. This confusion caused by the attack enables the attacker to communicate with the server and carry out their intentions without the actual client or the server being aware of what is happening. [1]

4.9- Malware Attacks

Any unwanted software installed in a computer without the user's consent or the administrator of the said computer is called malware. These malicious software attaches themselves to the system's legitimate codes and its useful Applications while replicating across the internet. Many types of malware exist with various capabilities and intentions.

4.10- Logic Bombs

This malicious software is programmed to be activated by a specific occurrence on a computer. The existence of logic, such as specific data, time, or any other occurrence to trigger them, has given them the name "Logic Bombs." [1]

4.11- Trojans

Also known as Trojan Horses, these programs hide in trustworthy and useful programs while performing malicious functions. However, unlike viruses, Trojans do not self-replicate themselves.

Trojans can launch attacks on a computer and establish back doors that are later exploited by attackers.

4.12- SQL injection Attacks

These cyber-threats have been a widespread problem for websites that are driven by databases. They come into effect when the malefactor executes an SQL query to a website database with data input from the client to the server. Various SQL commands are then inserted into the data-plane input so that predefined SQL commands can be run. A successful SQL injection attack can gain access to sensitive data in a database and insert, update, or delete data. [1]

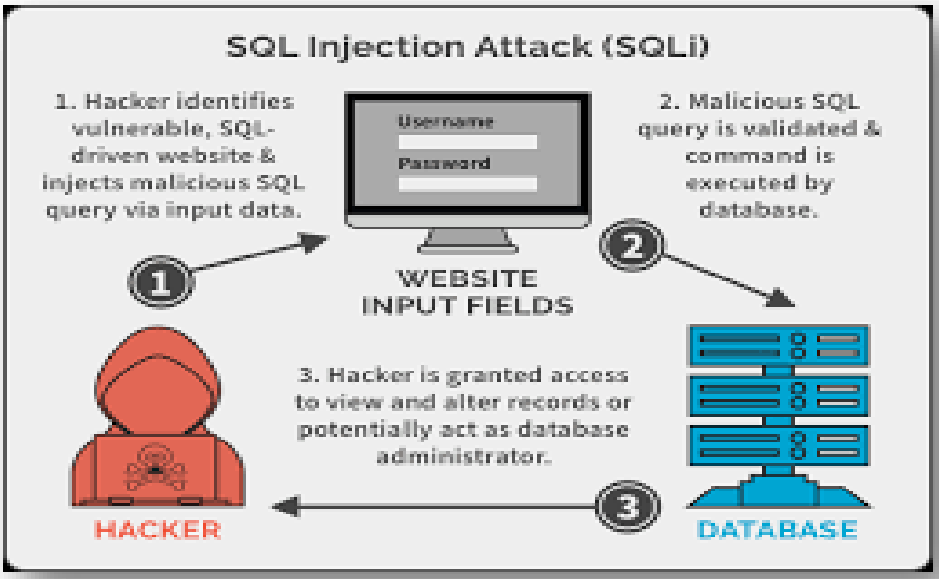

Figure 1.10:SQL injection Attack (SQLI).

4.13- Click jacking

This attack tricks the user into clicking on a link or button, which looks genuine but has malicious scripts embedded in it. This book has only covered the tip of the iceberg for cybersecurity terminology, but this will serve as a Launchpad to better understanding.

5- Information Security Governance

Information security governance is a part of an overall IT governance strategy that is focused on reducing risk and providing value back to the organization in the form of protecting assets and aligning to business needs. Governance is about assigning responsibility for the use and protection of corporate assets to the managers and employees entrusted with their care in addition to

measuring the results of their actions. Data has become one of the most valuable assets a company possesses, and its proper protection is no longer best effort but required by law and shareholders.

Scanning through laws like HIPAA, GLBA, and SOX, you will find that each one identifies either the board of directors or executive management as being ultimately responsible for securing corporate data. To adhere to the law and to provide due care in managing the business data, organizations must build a sustainable security program that addresses these requirements.

A security governance strategy provides the blueprint to direct and control the security program with clearly defined goals and objectives. Figure 1-12 shows how information security governance provides the glue to coordinate the efforts of people, processes, and technologies to better secure the key assets of a company. [2]

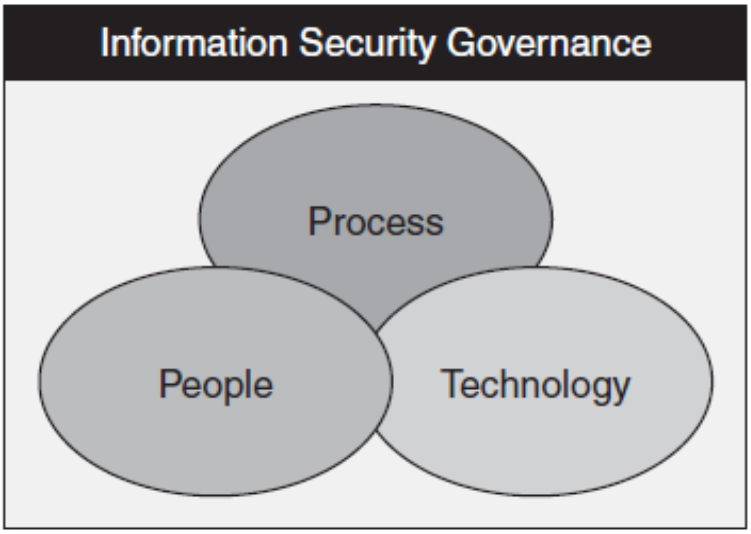

Figure 1.11: Information Security Governance Infrastructure

6- Information Audit

The investigation and evidence gathering activity that supports this overall control process is called “information audit.” Generally speaking, the audit function is well known to most IT managers. It can be conducted both externally (audit) or internally (internal audit). Audits are normally held at preplanned times or at predetermined milestones as specified in a project plan.

Audits are performed based on audit criteria, and the company’s management has to agree on all outcomes and responsibilities for any action item and closure criteria. The purpose of the

audit process is to gather sufficient reliable, pertinent, and practical evidence to demonstrate that the defined security and performance control objectives have been satisfied. [3]

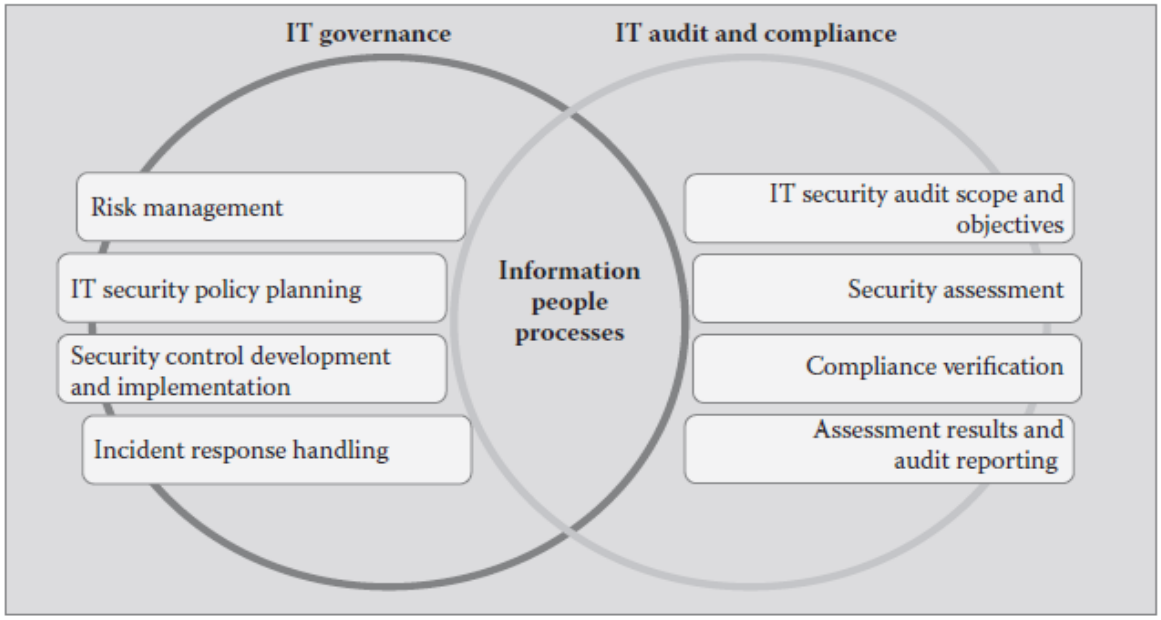

Figure 1.12: IT governance and IT audit.

6.1- Risk Assessment

There are many techniques used to assess risk. Some measure risk through numerical models, and others use experience and professional opinion to measure risk. No matter how good your models are or how extensive your research is, a portion of the equation always has a level of uncertainty. Risk by its very nature does not lend itself to a numeric value that is 100 percent accurate. [2]

- **Asset:** an asset is anything of value to your organization.
- **Threat:** A threat is any type of event that can cause loss and is usually measured in terms of probability or likelihood of occurrence. In the case of viruses, companies can see hundreds of infection attempts a month, so the threat of being exposed (though not necessarily infected) to a virus is close to 100 percent.
- **Vulnerability:** A vulnerability is a weakness that can result in a threat being able to compromise an asset. Hardware and software vulnerabilities are discovered on a daily basis, but the greatest number of vulnerabilities still comes from default configurations. One other thing to note is that just because a system is "vulnerable" does not mean that the vulnerability can be exploited.

Hardening systems by removing unneeded services can help to reduce the potential vectors of attack by preventing access to vulnerable services.

- **Cost of exposure:** Cost of exposure refers to the total tangible and intangible cost associated with an asset being compromised. Many times the actual monetary value of an asset is a small portion of its total value to the organization. A $10,000 database server might have 1 million dollars' worth of data on it. You have to understand the business processes and interconnections to assess the true cost of exposure. Interdependent systems can grind to a halt if a 25-cent part breaks.

6.2- Types of Audits

Audits can be broken down into a number of types, from the simple analysis of security architecture based on opinion, to a full-blown, end-to-end audit against a security framework such as ISO27001. The difference between types of audits is in what the auditor based the findings on and how detailed the audit's scope is. [2]

6.2.1- Security Review

A security review is when you examine the security posture of an organization based on professional experience and opinion. Think of a security review as a site survey. In this type of examination, you look for issues that stand out as a way to help define the starting point for further activities. Examples include:

- Penetration test
- Vulnerability scan
- Architecture review
- Policy review
- Compliance review
- Risk analysis

6.2.2- Security Assessment

Security assessments utilize professional opinion and expertise, but they also analyze the output for relevancy and criticality to the organization. The analysis aspect of an assessment attempts to quantify the risk associated with the items discovered to determine the extent of the problem. If you have two servers with the same vulnerability, but one is your financial server, and the other operates as a print server a security assessment would rank the financial server as a high risk and the print server as a lower risk based on the severity and damage potential. The biggest

differentiator between an assessment and a review is the depth to which the auditor examines the system and analyzes the results. Examples include:

- Vulnerability assessment
- Risk assessment
- Architecture assessment
- Policy assessment

6.2.3- Security Audit

A security Audit examines the organization's security posture against an industry standard (ISO27001 or COBIT) and/or regulatory compliance such as HIPAA or PCI. An audit includes review and assessment; it also conducts a gap analysis against standards to measure how well the organization complies. Audits take into account people, processes, and technologies, and it compares them to a benchmark in a standardized and repeatable way. Examples include:

- Compliance audit
- Policy audit
- Procedure audit
- Risk audit.

6.3- Security Auditing Tools

One thing is certain about security auditing tools: The power and sophistication of tools that auditors have at their disposal increase exponentially every year. Not only are the authors of these tools truly brilliant individuals (and some scary ones, too), they have also helped the security community significantly through the automation of advanced testing techniques. [2]

6.3.1- Service Mapping Tools

Service mapping tools are used to identify systems, remote services, and open ports.
These types of tools can be used to test a firewall rule base or response given different real or crafted IP packets.

- **Nmap**

Nmap is the network and service scanning tool of choice for most security professionals. It is a free, open source application available on all UNIX and Windows operating systems.
The tool is command-line based, but there are a number of graphical frontends for those who want a point-and-click experience. [2]

6.3.2- Vulnerability Assessment Tools

There are many vulnerability assessment tools available today, from commercial applications to well-known open source tools. A vulnerability scanner's purpose is to map known vulnerabilities in products and present a report of potential vulnerabilities. This type of tool is great for automating the assessment of multiple hosts and usually provides nice severity categorization and output for reports. Obviously, you need to be careful when performing vulnerability tests on business systems because some of the assessment mechanisms these tools use to find vulnerabilities can crash services or cause an outage. Auditors should have a plan in place for restoring service in the event of a problem and perform testing outside of peak utilization times.

- **Nessus**

Nessus is a popular vulnerability scanner that looks for known vulnerabilities in operating systems, networking gear, and applications. Currently at version 4, Nesus has expanded its functionality significantly since it was introduced as an open source project more than 10 years ago. With the release of Version 4, Nessus has become a closed source product owned by Tenable Network Security. While the scanner is still free for home use to scan your personal devices, if you use it in any other capacity outside of the home, a professional feed license is required. The professional feed provides access to the latest updates and advanced features such as compliance checks (PCI NIST or CIS), SCAP protocol support, the ability to load it as virtual appliance, and product support from Tenable. The yearly professional license fee for Nessus is around $1,200.

6.3.3- Packet Capture Tools

Validation and testing of security controls are the most important aspects of conducting an audit. Auditors shouldn't just assume a firewall or IPS will enforce policy; they must test it and gather evidence about how well those controls do their jobs.

Packet capture tools are familiar to anyone who has had to troubleshoot a challenging network redesign or configuration. Packet capture tools are also extremely valuable when testing firewall rules, IPS signatures, and practically any other scenario where you need to see exactly what is going across the wire. Tcpdump and Wireshark are two free tools that should be in every auditor's repertoire.

- **Tcpdump**

Tcpdump is a free packet capture program that operates as a simple command-line based "sniffer". It has been compiled for practically every operating system and leverages the UNIX Libpcap

library (Winpcap on Windows) to copy traffic from the wire and display it on the screen or save it to a file. This simple packet sniffer provides a detailed view into the actual bits and bytes flowing on a network. Tcpdump is a simple application that doesn't have a graphical interface that abstracts the details of the packet capture process to automatically detect problems. It is left to the auditor to use his knowledge and experience to identify anomalies or issues. That doesn't mean that Tcpdump doesn't decode traffic; it just doesn't perform higher-level interpretation like Wireshark.

6.3.4- Penetration Testing Tools

Auditors can leverage high-quality penetration testing tools to make auditing security controls significantly easier. Most professional penetration testers use a combination of general purpose exploit frameworks such as Core Impact and Metasploit in addition to their own custom scripts and applications. Not everyone in security is an uber hacker or has the time to build their own tools to test for exploitable services. These two applications are powerful and represent the best of the commercial and open source penetration testing tools available.

- **Core Impact**

In the world of penetration tools, Core Impact is widely considered the best commercial product available. Developed by Core Security Technologies, this software package is a comprehensive penetration testing suite with the latest commercial grade exploits and a drag-and-drop graphical interface that can make anyone look like a security penetration testing pro. Writing exploit code and delivering it to a remote system is not a trivial task, but Core Impact makes it look easy.

7- Technical control

Technical control in turn provides the appropriate level of protection automatically. The following list provides a few examples.

■ **Encryption:** Encryption is a strong technical control used to protect the confidentiality of data. This includes data transferred over a network, and data stored on devices such as servers, desktop computers, and mobile devices. Cryptography uses the combination of algorithms and public key infrastructures to convert data into an undecipherable format during transfer.

■ **Antivirus software:** Once installed, the antivirus software provides protection against malware and spyware infection.

■ **IDSs:** An IDS is installed and configured by an organization in order to monitor network or host for intrusions and provide ongoing protection against identified threats. [4]

■ **Firewalls:** Network firewalls restrict network traffic going in and out of a network.

■ **Least privilege:** The principle of least privilege stipulates that an individual or process is granted only the privileges needed to perform an intended task or function. Based on this principle, privileges are a combination of rights and permissions.

8- Difference between Firewall and Intrusion Detection System

A firewall is a hardware and/or software which functions in a networked environment to block unauthorized access while permitting authorized communications. Firewall is a device and/or a software that stands between a local network and the Internet, and filters traffic that might be harmful. An Intrusion Detection System (IDS) is a software or hardware device installed on the network (NIDS) or host (HIDS) to detect and report intrusion attempts to the network.

We can think a firewall as security personnel at the gate and an IDS device is a security camera after the gate. A firewall can block connection, while a Intrusion Detection System (IDS) cannot block connection. An Intrusion Detection System (IDS) alert any intrusion attempts to the security administrator.

Firewall	IDS
A firewall is a hardware and/or software which functions in a networked environment to block unauthorized access while permitting authorized communications	An Intrusion Detection System (IDS) is a software or hardware device installed on the network (NIDS) or host (HIDS) to detect and report intrusion attempts to the network.
A firewall can block an unauthorized access to network (E.g. A watchman standing at gate can block a thief)	An IDS can only report an intrusion; it cannot block it (E.g. A CCTV camera which can alert about a thief but cannot stop it)
A firewall cannot detect security breaches for traffic that does not pass through it (E.g. a gateman can watch only at front gate. He is not aware of wall-jumpers)	IDS is fully capable of internal security by collecting information from a variety of system and network resources and analyzing the symptoms of security problems
Firewall doesn't inspect content of permitted traffic. (A gateman will never suspect an employee of the company)	IDS keeps a check of overall network
No man-power is required to manage a firewall.	An administrator (man-power) is required to respond to threats issued by IDS
Firewalls are most visible part of a network to an outsider. Hence, more vulnerable to be attacked first. (A gateman will be the first person attacked by a thief!!)	IDS are very difficult to be spotted in a network (especially stealth mode of IDS)

Table 1.1: Difference between Firewall and Intrusion Detection System

Conclusion

In this chapter, we presented an overview of computer security in a network and the importance of setting up a security policy by tracing the needs and desired objectives in order to remedy the constant threats that a computer network undergoes. These threats generally manifest themselves in the form of computer attacks which we have illustrated in order to show the intensity of the danger. Finally, we have proposed some existing solutions in order to protect themselves and reduce the risks.

CHAPTER 2

INTRUSION DETECTION SYSTEMS

1- Introduction

With the increasing importance of information systems in today's complex and global economy, it has become mission and business critical to defend those information systems from attack and compromise by any number of adversaries. Intrusion prevention and detection systems are critical components in the defender's arsenal and take on a number of different forms.

Formally, intrusion detection systems (IDSs) can be defined as "software or hardware systems that automate the process of monitoring the events occurring in a computer system or network, analyzing them for signs of security problems." Intrusion prevention systems (IPSs) are systems that attempt to actually stop an active attack or security problem. Though there are many IDS and IPS products on the market today, often sold as self-contained, network attached computer appliances, truly effective intrusion detection and prevention are achieved when viewed as a process coupled with layers of appropriate technologies and products.

In this chapter, we will discuss the nature of computer system intrusions, those who commit these attacks, and the various technologies that can be utilized to detect and prevent them and Anti-IDS Techniques; The limits of the ids.

2- Definition of Intrusion

Information security concerns itself with the confidentiality, integrity, and availability of information systems and the information or data they contain and process. An intrusion, then, is any action taken by an adversary that has a negative impact on the confidentiality, integrity, or availability of that information. Given such a broad definition of "intrusion" it is instructive to examine a number of commonly occurring classes of information system (IS) intrusions.

3-Definition of an intrusion Detection system

An Intrusion Detection System (IDS) is a set of software and/or hardware components whose main function is to detect and analyze abnormal or suspicious activities on the analyzed target (a network or host). It thus allows to have a knowledge on successful attempts as failed intrusions.

Some terms are often used when talking about SDI:

• **False Positive:** an alert from an IDS, but not an actual attack.

• **False negative:** an actual intrusion that has not been detected by the IDS. [9]

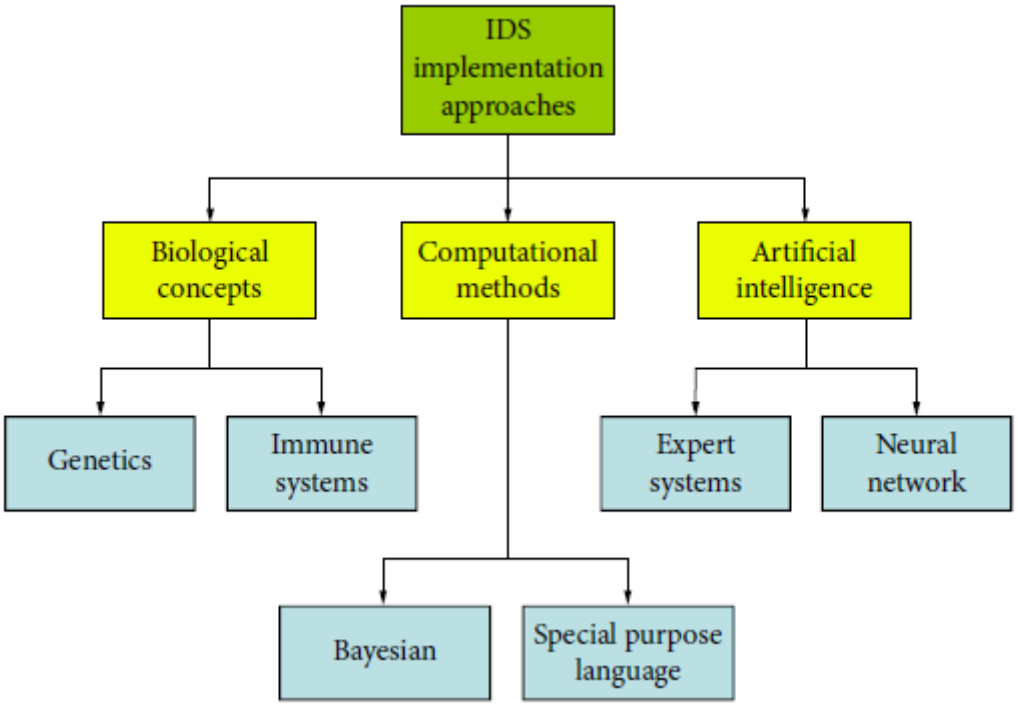

Figure 2.1: Some IDS implementation approaches

4- How does an IDS work?

To better understand the work of an IDS, consider that if firewalls are security guards, intrusion detection systems are security cameras. According to Barracuda, an IDS monitors traffic and spots patterns of activity, alerting you if it concludes that your network is under attack.

Signature detection compares network or system information to attacks already listed in the IDS database. Anomaly detection compares current network traffic to the normal levels of packet size or activity and analyzes the result statistically. If network traffic suddenly shoots up to a high level, for instance, that could indicate a hacking attack. These monitoring features offer one of the advantages of an intrusion detection system over firewall options. [10]

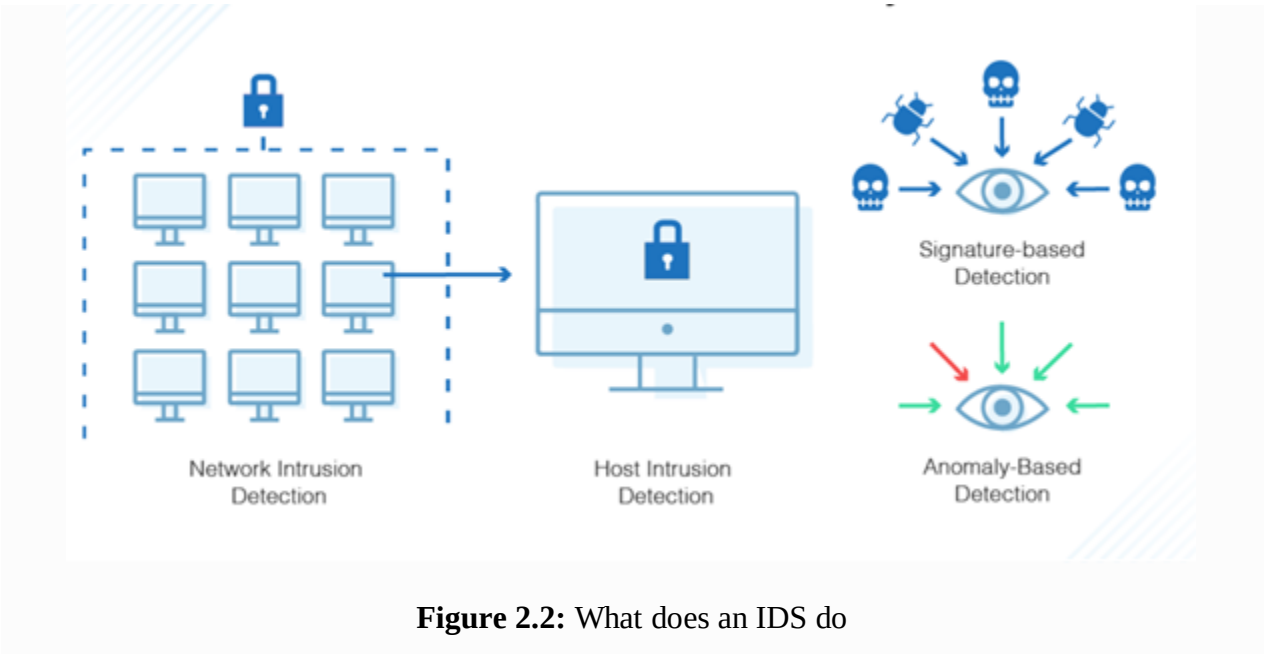

Figure 2.2: What does an IDS do

5- Types of intrusion detection systems

Because of the diversity of attacks carried out by hackers, intrusion detection must take place at several levels. There are therefore different types of SDI:

5.1- Network-based Intrusion Detection System (NIDS)

NIDS are SDIs dedicated to networks. They usually include a machine that listens on the network segment to be monitored, a sensor and a motor that performs traffic analysis to detect intrusions in real time. A NIDS listens to all network traffic, then analyzes it and generates alerts if packets seem dangerous. (See figure 2.3).

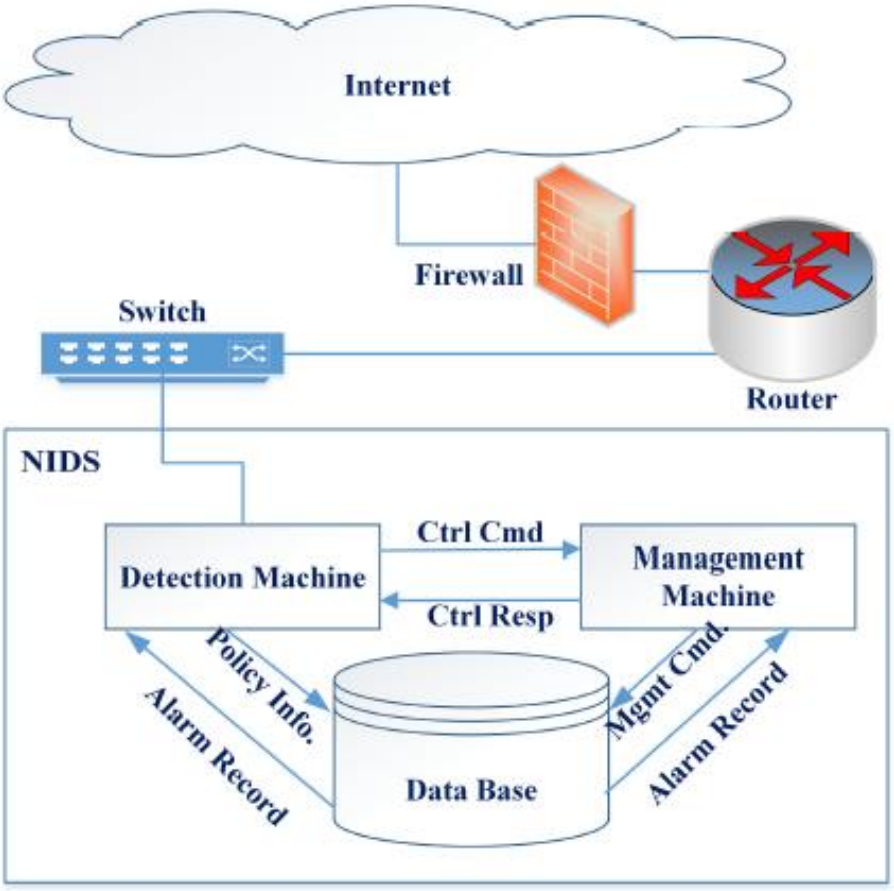

Figure2.3: NIDS with its components.

The implantation of a NIDS on a network is done as follows: sensors (often simple hosts) are placed in strategic locations on the network and generate alerts if they detect an attack. These alerts are sent to a secure console that analyzes them and possibly handles them. This console is generally located on an isolated network that connects only the sensors and the console.

• Inside the firewall

If the sensors are inside the firewall, it will be easier to tell if the firewall has been improperly configured so that we can know if an attack has come from that firewall.

• Outside the firewall

Sensors located outside the firewall are used to detect and analyze attacks. It offers the advantage of writing to logs, so the administrator sees what he needs to change in the firewall configuration.

Some examples of NIDS:

- NetRanger [http ://www.cisco.com].
- Dragon [http ://www.securitywizards.com].

- NFR [http ://www.nfr.net].
- Snort [http://www.snort.org].
- DTK [http://all.net/dtk/dtk.html].

5.2- HIDS (Host-based Intrusion Detection System)

Host-based intrusion detection systems only analyze information about that host. As they do not have to control network traffic but only the activities of a host they usually show more accurate about the types of attacks. In addition, we immediately notice the impact on the machine concerned, for example if a user successfully attacked it. These IDSs use two types of sources to provide information about the activity: logs and audit traces of the operating system. Each has its advantages: audit trails are more accurate and detailed and provide better information while logs that only provide essential information are smaller. (See figure)

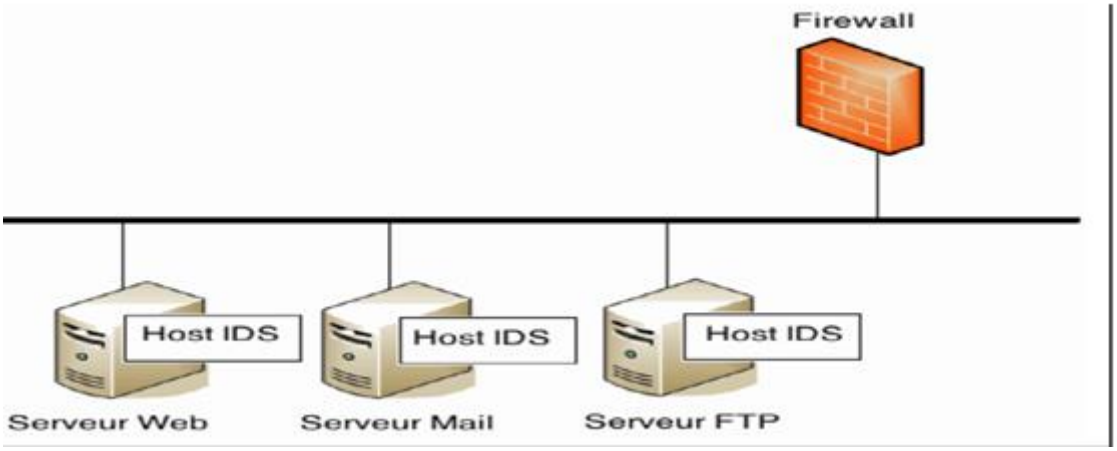

Figure2.4: HIDS (Host Intrusion Detection System).

Some examples of HIDS:

- Tripwire [http://www.tripwire.com/products/index.cfml].
- SWATCH [http://freshmeat.net/redir/swatch/10125/url_homepage/swatch].
- Dragon Squire [http://www.enterasys.com/ids/squire/].
- Tiger [http ://freshmeat.net/redir/tiger-audit/30581/url_homepage/tiger].

5.3- Hybrid IDS

Hybrid intrusion detection systems bring together the characteristics of several different detection systems. In practice, there is a combination of NIDS and HIDS to monitor the network and the host.

The probes act as a NIDS or a HIDS It allows to gather the information of various probes placed on the network. The best known example in the Open-Source world is Prelude.

-This IDS allows alerts from a variety of systems to be stored in a database.

Using Snort as NIDS, and other software such as Samhain as HIDS, it allows to combine powerful tools all together to allow centralized visualization of attacks.

6- Characteristics of IDS

Some of the desirable features found in an intrusion detection system include:

- Resist attempts at corruption, that is to say, it must be able to detect if it has undergone an undesirable modification itself.
- Use minimal system resources under supervision.
- Adapt over time to changes in the monitored system and user behavior.
- Be easily configurable to implement a specific network security policy.

7- The architecture of an IDS

This section describes the three components that typically make up an intrusion detection system. Figure 2.5 illustrates the interactions between these three components

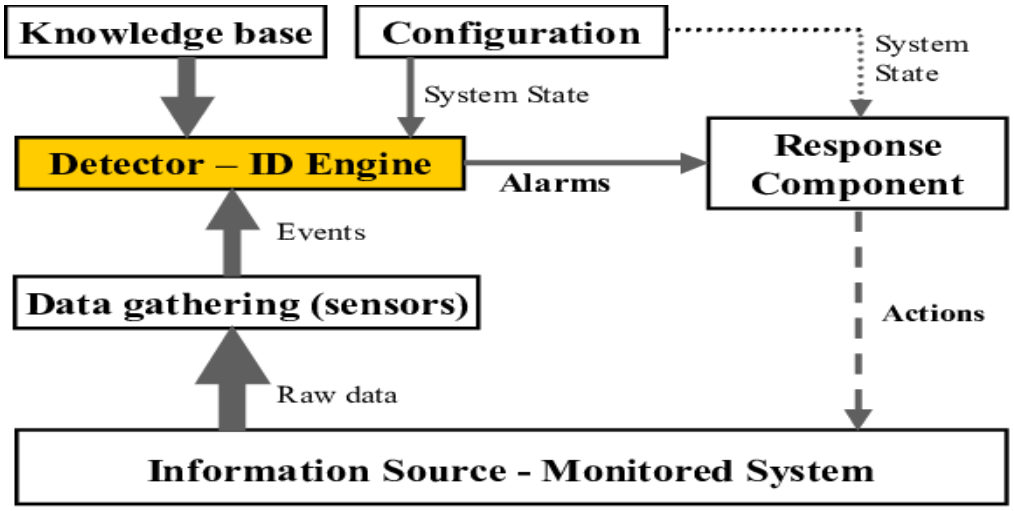

Figure2.5: IDS architecture

7.1- Sensor

The sensor observes system activity through a data source and provides the analyzer with a sequence of events that provide information on the changing state of the system. The sensor can content itself with directly transmitting this raw data, but in general a pre-processing is carried out.

There are typically three types of sensors depending on the data sources used to observe the activity of the system: system sensors, network sensors and application sensors.

7.2- Analyzer

The objective of the analyzer is to determine if the event stream provided by the sensor contains elements characteristic of malicious activity.

7.3- Manager

The manager collects the alerts produced by the sensor, formats them and presents them to the operator. Eventually, the manager is responsible for the reaction to adopt, which can be:

• Containment of the attack, which aims to limit the effects of the attack.

• Attack eradication, which tries to stop the attack.

• Recovery, which is the stage of restoring the system to a healthy state.

• Diagnosis, which is the problem identification phase.

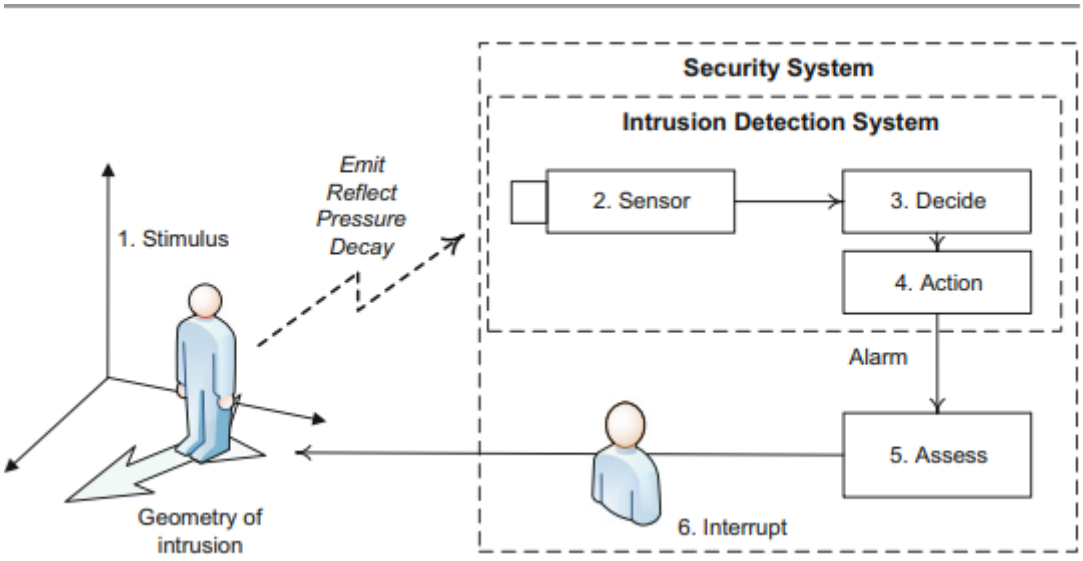

Figure 2.6: The principle of detection.

8- Setting up an IDS

8.1- The positioning of the IDS

There are several strategic locations where an SDI should be placed. The following diagram shows a local network and the three positions a IDS can take:

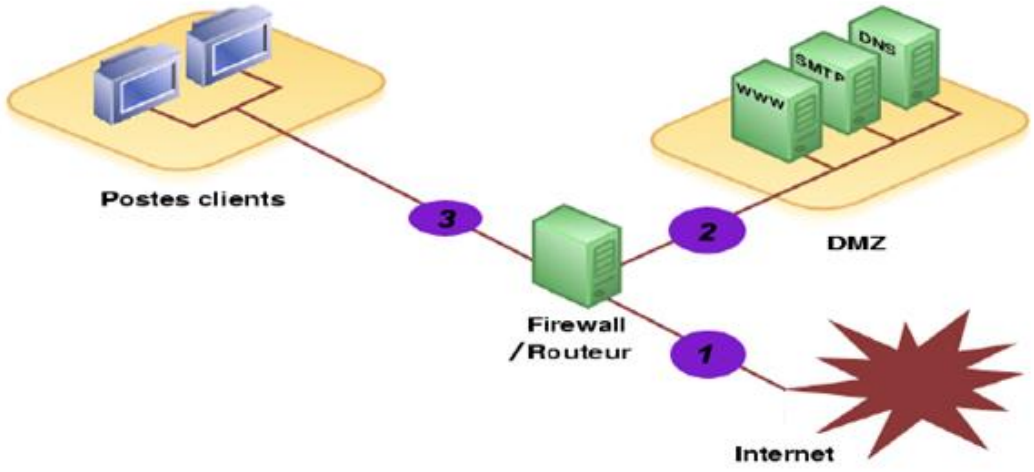

Figure2.7: The position of IDS.

• **Position (1):** On this position, the IDS will be able to detect all frontal attacks, coming from the outside, upstream of the firewall. Thus, many alerts will be reported which will make the logs difficult to consult.

• **Position (2):** If the IDS is placed on the DMZ, it will detect attacks that have not been filtered by the firewall and that fall within a certain skill level. The logs will be clearer to consult since the attacks will not be recorded.

• **Position (3):** The IDS can here account for internal attacks, coming from the company's local network. It may be wise to place one at this location given that 80% of attacks come from within. In addition, if trojans have contaminated the computer park (unsuspecting browsing on the internet) they can be here easily identified and then eradicated.

9-Criteria for testing an IDS

When setting up an IDS, it is necessary to take into consideration several criteria that will make it possible to choose the best IDS. Testing an IDS with vulnerability scanners is a necessary measure to assess an IDS, but is far from sufficient. Other criteria must be taken into account:

Speed, Detection methods and capabilities, Software architecture, Data usability

10- Mode of operation

Two aspects must be distinguished in the operation of an IDS: the detection mode used and the response provided by the IDS when an intrusion is detected.

There are two modes of detection, anomaly detection and signature recognition. On their own, there are two types of responses, passive and active.

We will first study the modes of detection of an IDS, before presenting the possible responses to an attack.

Detection modes:

We note two detection modes that are:

- The detection of anomalies
- The recognition of signature

It should be noted that signature recognition is the mode of operation most implemented by the market IDS. However, new products tend to combine both methods to refine intrusion detection. [11]

10.1- Detection of anomalies

It consists in detecting anomalies in relation to a "usual traffic" profile. The implementation always includes a learning phase during which the SDI will discover the normal functioning of the monitored elements. They are thus in a position to point out deviations from the reference operation.

Behavioural models can be developed from statistical analyses. They have the advantage of detecting new types of attacks.

However, frequent adjustments are needed to evolve the reference model to reflect normal user activity and reduce the number of false alerts generated.

In the case of HIDS, this type of detection can be based on information such as the CPU usage rate, the activity on the disk, the times of connection or use of certain files (office hours…).

10.2- Recognition of Signatures

This approach involves searching the activity of the monitored element for fingerprints (or signatures) of known attacks. This type of IDS is purely reactive; it can only detect attacks of which it has the signature.

As a result, it requires frequent updates. In addition, the efficiency of this detection system depends heavily on the accuracy of its signature base. This is why these systems are circumvented

by hackers who use so-called “escape” techniques that consist of masking the attacks used. These techniques tend to vary the signatures of attacks that are no longer recognized by the IDS.

It is possible to develop more generic signatures, which can detect variants of the same attack, but this requires a good knowledge of the attacks and the network, so as to stop the variants of an attack and not hinder the normal network traffic.

A signature allows defining the characteristics of an attack, at the packet level (up to TCP or UDP) or at the protocol level (HTTP, FTP…).

• At the packet level, the IDS will analyze the different parameters of all the packets passing through and compare them with known attack signatures.

• At the protocol level, the IDS will check at the protocol level whether the commands sent are correct or do not contain an attack. This feature has mostly been developed for HTTP at the moment.

It should be noted that signatures are updated to reflect new attacks identified.

However, the more different signatures to test, the longer the processing time. The use of more elaborate signatures can therefore provide a saving of time appreciable.

However, a poorly crafted signature may ignore real or identified attacks from normal traffic as an attack. The development of signatures should therefore be handled carefully and with a good knowledge of the monitored network and the existing attacks.

Once an attack is detected, an IDS has the choice between several types of responses, which we will now detail. [8]

Type	Advantages	Disadvantages
Anomaly-based	1. Ability to detect zero-day attack attempts. 2. Low false negative rate.	1. Slow to work when placed in a new environment. 2. High false positive rate. 3. Low detection rate for known attacks.
Signature-based	1. High response time for known attacks. 2. Low false positive rate.	1. Limited capability to detect zero-day attacks. 2. Signature database must be updated frequently.

Table 2.1: Advantages and Disadvantages between IDS types

11- Active and passive responses

There are two types of responses, depending on the SDIs used. Passive response is available for all SDIs while active response is more or less implemented. [13]

11.1- Passive Response

The passive response of an IDS is to record the detected intrusions in a log file that will be analyzed by the security manager.

Some IDSs allow you to log all of a connection identified as malicious. This helps to address security vulnerabilities to prevent recorded attacks from happening again, but it does not directly prevent an attack from happening.

11.2- Active Response

The active response, on the contrary, aims to stop an attack when it is detected. For this we have a reconfiguration of the firewall.

The reconfiguration of the firewall makes it possible to block malicious traffic at the firewall level, by closing the port used or by prohibiting the address of the attacker. This functionality depends on the firewall model used, not all models allow reconfiguration by an IDS.

In the case of an active response, it is necessary to be sure that the traffic detected as malicious is actually, under penalty of disconnecting normal users.

In general, SDIs do not actively respond to all alerts. They only respond to alerts when they are positively certified as attacks. The analysis of the generated alert files is therefore an obligation to analyze all the detected attacks.

	Positive	Negative
TRUE	Alert whene there is malicious traffic	Silent when traffic is bengin
FALSE	Alert when traffic is benign	Silent when malicious traffic occurs

Table 2.2: True positives and negatives.

12- Methods of detections

To properly manage an intrusion detection system, it is important to understand how it works. A simple question arises on the detection of such a system and on the criterion of differentiating a flow containing an attack from a normal flow.

These questions led us to study the internal workings of an SDI. From this, we deduced two techniques put in place in the detection of attacks. The first is to detect known attack signatures in packets circulating over the network. The second is to detect suspicious activity in the user's behavior. These two techniques, no matter how different, can be combined within a single system to increase safety. [12]

12.1- The behavioral approach (Anomaly Detection)

This technique consists of detecting an intrusion based on the past behavior of the user. To do this, you must first create a user profile based on your habits and trigger an alert when off-profile events occur.

This technique can be applied not only to users, but also to applications and services. Several metrics are possible: the CPU load, the volume of data exchanged, the connection time on resources, the statistical distribution of the protocols and applications used, the connection times... However, it has some disadvantages:

Unreliable: any change in user habits triggers an alert;

- requires a non-functioning period to implement self-learning mechanisms: if a hacker attacks during this time, his actions will be assimilated to a user profile, and therefore will go unnoticed when the detection system is fully implemented.
- profiling must be flexible so that there are not too many false alerts: the hacker can discreetly intervene to modify the user's profile in order to get after several days or weeks, a profile that will allow him to set up his attack without it being detected.

13- Anti-IDS Techniques

Like almost any computer system, there are flaws in IDS, or rather techniques that allow to bypass these systems without being detected. If a hacker detects the presence of an IDS, he can disable it, or better yet, generate fake attacks while he quietly commits his plan. [*] There are three categories of attacks against SDIs:

• Denial of service attack: make the SDI inoperative by saturating it.

• insertion attack: the hacker, to avoid being spotted, injects packages of decoys that will be ignored by the target's operating system, but taken into account by the IDS: the IDS detects nothing abnormal, while on the target system, The attack takes place because the superfluous packets are ignored.

• Escape attack: this is the opposite technique to insertion attack. Here, superfluous data is ignored by the IDS, but taken into account by the operating system.

14-The limits of the ids

Like any computer system, SDIs have limitations. These include:

• Pollution/overload: SDI can be polluted or overloaded, for example by generating large volumes of traffic (as difficult and heavy as possible to analyze). A significant amount of attacks can also be sent to overload IDS alerts. Possible consequences of this overload may be the saturation of resources (disk, CPU, memory), loss of packets, partial or total denial of service, etc.

• Resource consumption: in addition to the size of the log files (in the order of
Go), intrusion detection is extremely resource intensive. Indeed, a NIDS system must generate logs of abnormal or doubtful activity on the network.

• Packet loss (performance limitation): transmission speeds are
sometimes such as they far exceed the writing speed of hard drives, or even the
processor processing speed.

It is therefore not uncommon for packets not to be processed by the IDS, and for some packets to be received by the receiving machine.

• Vulnerability to Denial of Service: an attacker may try to cause denial
service at the level of the intrusion detection system, or worse at the level of the operating system of the machine supporting the IDS.

Once the IDS is deactivated ("off"), the attacker can try anything that suits him or her.

15- Security Experts

With a network IDS, the biggest challenge—aside from false negatives and false positives—can be the sheer volume of alerts. One of the most important elements of using a network intrusion detection system effectively is ensuring you have IT security personnel with the knowledge and skills to necessary weed out false alarms and identify suspicious or malicious traffic the network IDS might have missed.

Attacks don’t have work hours—they occur around the clock every day of the year. You should have a security operations center (SOC) with security experts who can monitor alerts and analyze log data to identify and prioritize potential attacks and take the appropriate action to block the traffic or thwart the attack.[18]

Conclusion

In this chapter we talked about ids defining how its types work its features and all its accessories from its boundary examples and techniques as well.

In recent years, intrusion detection systems have gained an important place in the security design of information systems. They are widely deployed in companies for various reasons such as: documenting attacks, evaluating security, and more generally monitoring information systems to stop, even prevent attacks in order to limit damage.

CHAPTER 3

MACHINE LEARNING TECHNIQUES

1- Introduction

One of the best features of today's era of technology is its flexibility and adaptableness. A new scientific innovation comes out almost every day. This ever-changing nature of scientific and technological world changes the trajectory of the world every day. Things that were considered dreams and fiction once are now rapidly turning into reality. Human beings are slowly but steadily trying to defeat nature at its own game. However, one field remains to be conquered. We still have not managed to conquer the world of machine learning or AI. However, it has become a buzzword now, and the whole of the world is talking about it. Not everyone is excited about it though. Most people are worried or scared of it. However, there is no need to be afraid of machine learning or AI, as it will help humanity to achieve things that we cannot even currently imagine.

2- Machine Learning

If you check the search results for the most popular keywords of 2016, you will find that machine learning and AI are leading the figures by a large margin. This steady rise in the fame of machine learning is because of its rising use in our daily lives. It is nowadays being used in various devices and machines as well as gadgets.

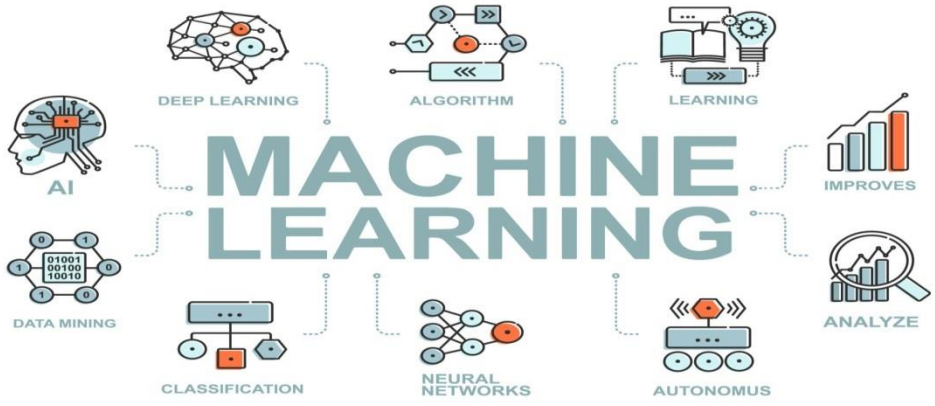

Figure3.1: machine learning environment

However, the general population is still wary of it. Machine learning can be defined as a process of inputting data to the computer systems in a way that the computer will learn the ability to process and perform the activity in the future without being explicitly programmed or being fed with similar or extra data. What this means in simple words is that it will allow computers to develop a 'mind' of their own and allow them to "think." Sounds scary but it isn't.

If computers are provided with the ability to think, they become smarter and thus easier to use. Their functionality will increase by a large margin, and they become an integral asset for humanity. Machine learning can be used in almost all the fields of epistemology. Right now, it is being used in areas such as cheminformatics, computational anatomy, gaming, adaptive websites, natural language processing, robot movement and locomotion, medical diagnosis, sequence mining, behavior analysis, linguistics, translation, fraud detection, *etc*. The list goes on. [14]

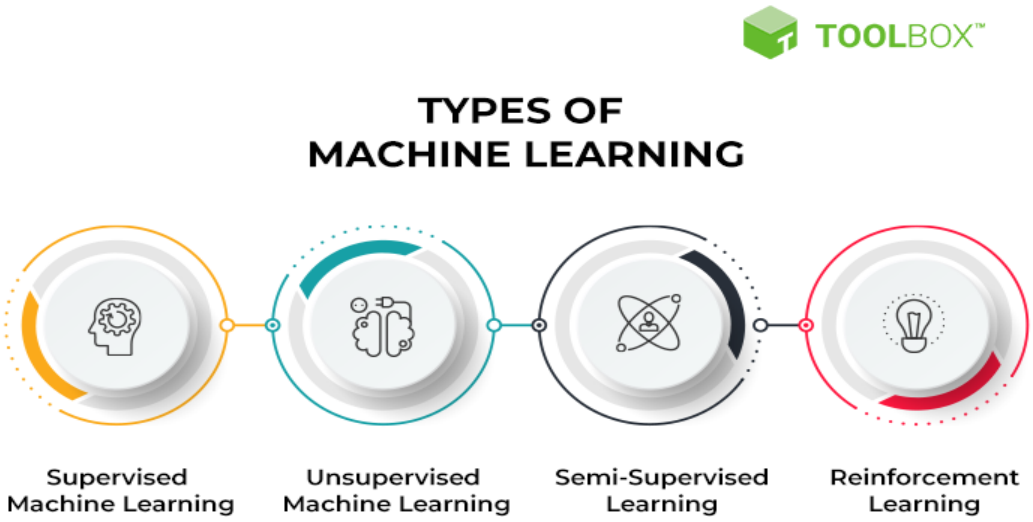

Figure 3.2: Types of machine learning

3- Supervised Learning

In this, the input data is known as training data. It features a known result or a label, for instance, spam/not-spam or stock price, *etc*. A proper prediction model is constructed using the training process. It is needed to make predictions, and these predictions get corrected if they are wrong. The training process continues to repeat itself until perfection, or the desired level of accuracy is achieved.

Examples: Logistic Regression and the Back Propagation Neural Network.

4- WEKA Environment

Weka (Waikato Environment for Knowledge Analysis) is a data mining environment developed by the "machine learning" research group of the Department of Computer Science at the University of Waikato in New Zealand [WIT 00]. It is used in research, education and industry. It is written in the Java language and tested on several platforms such as Linux and Windows. This

environment is an "open source" software and is available on the Web. It is a collection of learning algorithms that can be applied directly to a data set or called via a Java program. [19]

Weka contains tools for data preprocessing, classification, regression, clustering, association rules and visualization. Weka allows you to pre-process a data set, apply a learning algorithm, and analyze the results and performance of a classifier. It is also well suited to integrate new learning algorithms. [15]

Figure 3.3: Weka environment.

5- Classifiers

5.1- Introduction

In recent years, research has placed a great deal of importance on the processing of textual data. This is for several reasons: an increasing number of collections networked and distributed internationally, the development of communication infrastructure and the Internet. Manual processing of this data is very costly in terms of time and personnel, it is not very flexible and its generalization to other fields is almost impossible; that is why we are trying to develop automatic methods.

At present, many text classification software are available, they have been the subject of publications and their fields of application are expanding from day to day. In general, these systems

are based on machine learning algorithms, so we present learning methods that, from documents already filed, allow to classify new documents.

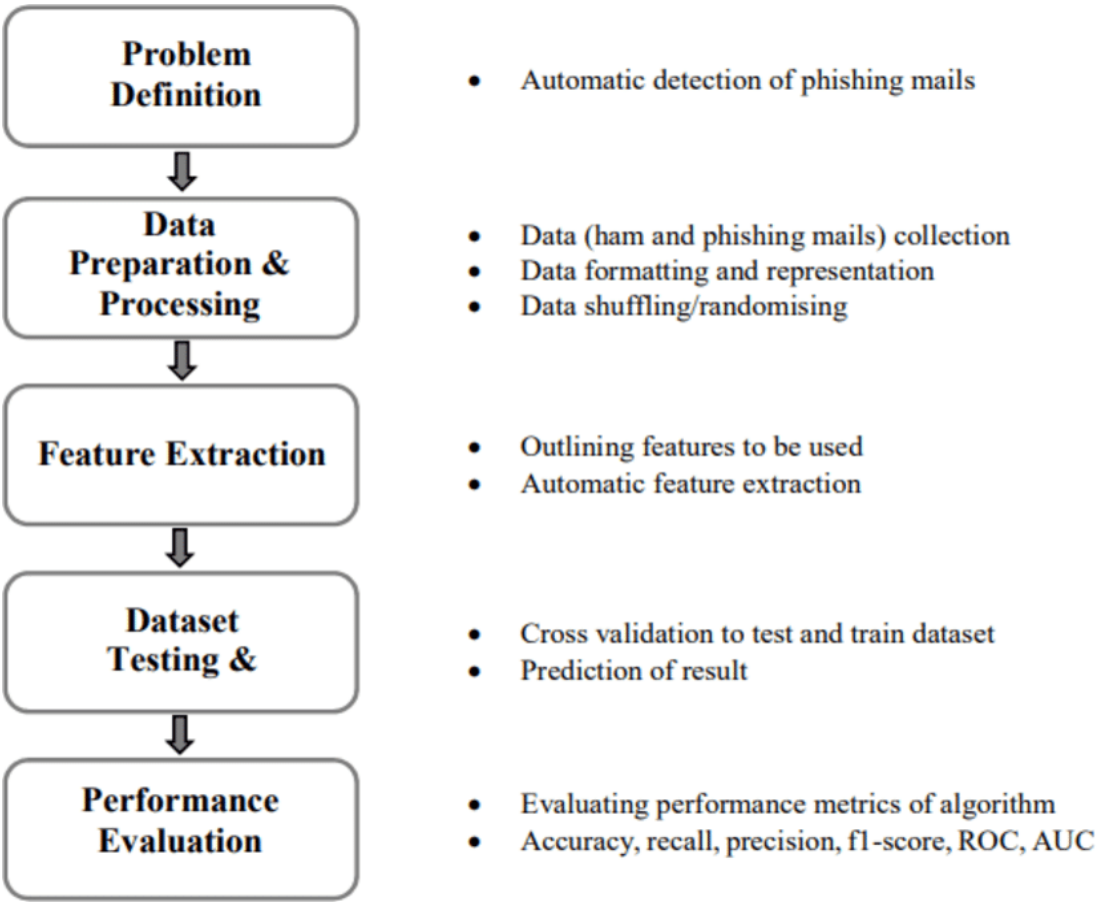

Figure 3.4: Steps-carried-out-in-the-machine-learning-process

5.2- Learning Algorithms

In machine learning, different types of classifiers have been developed, with the aim of achieving a maximum degree of accuracy and efficiency, each having its advantages and disadvantages. But they do share common characteristics.

There are many supervised learning algorithms, including:

- The decision trees.
- Vector support machines (or VSM) or SMO in WEKA.

6- Decision trees

6.1- Definition

Decision trees are the most popular learning methods. The known algorithms are ID3 (Quinlan 1986) and C4.5 (Quinlan 1993) called in WEKA under the name J48. They are also popular for document classification [20]. Like any supervised learning method, decision trees use examples. If documents are to be categorized, a decision tree by category must be constructed. To determine which category(s) a new document belongs to, the decision tree for each category to which the document to be filed is used. Each tree answers Yes or No (it makes a decision).

Concretely, each node of a decision tree contains a test (an IF...THEN) and the sheets have the values Yes or No. Each test looks at the value of an attribute of each example. Indeed, we assume that an example is a set of attributes/values. For documents, each attribute can be a word and the value will be for example 0 or 1 depending on whether this word belongs to the document or not.[16]

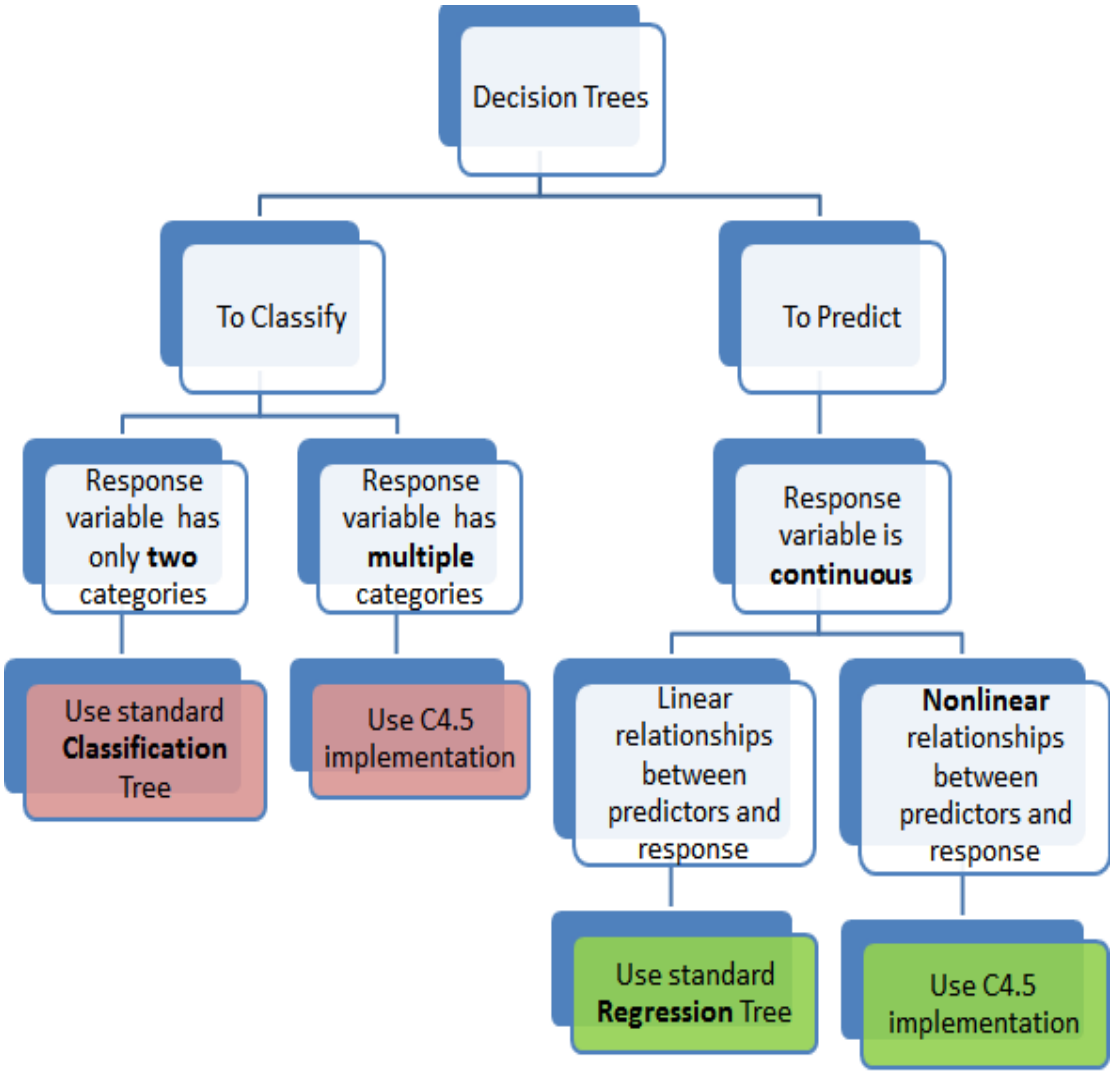

Figure 3.5: The decision tree.

To build the decision tree, you have to find out which attribute to test at each node. This is a recursive process. To determine which attribute to test at each step, a statistical calculation is used that determines how well this attribute separates the Yes/No examples.

We then create a node containing this test, and we create as many descendants as possible values for this test.

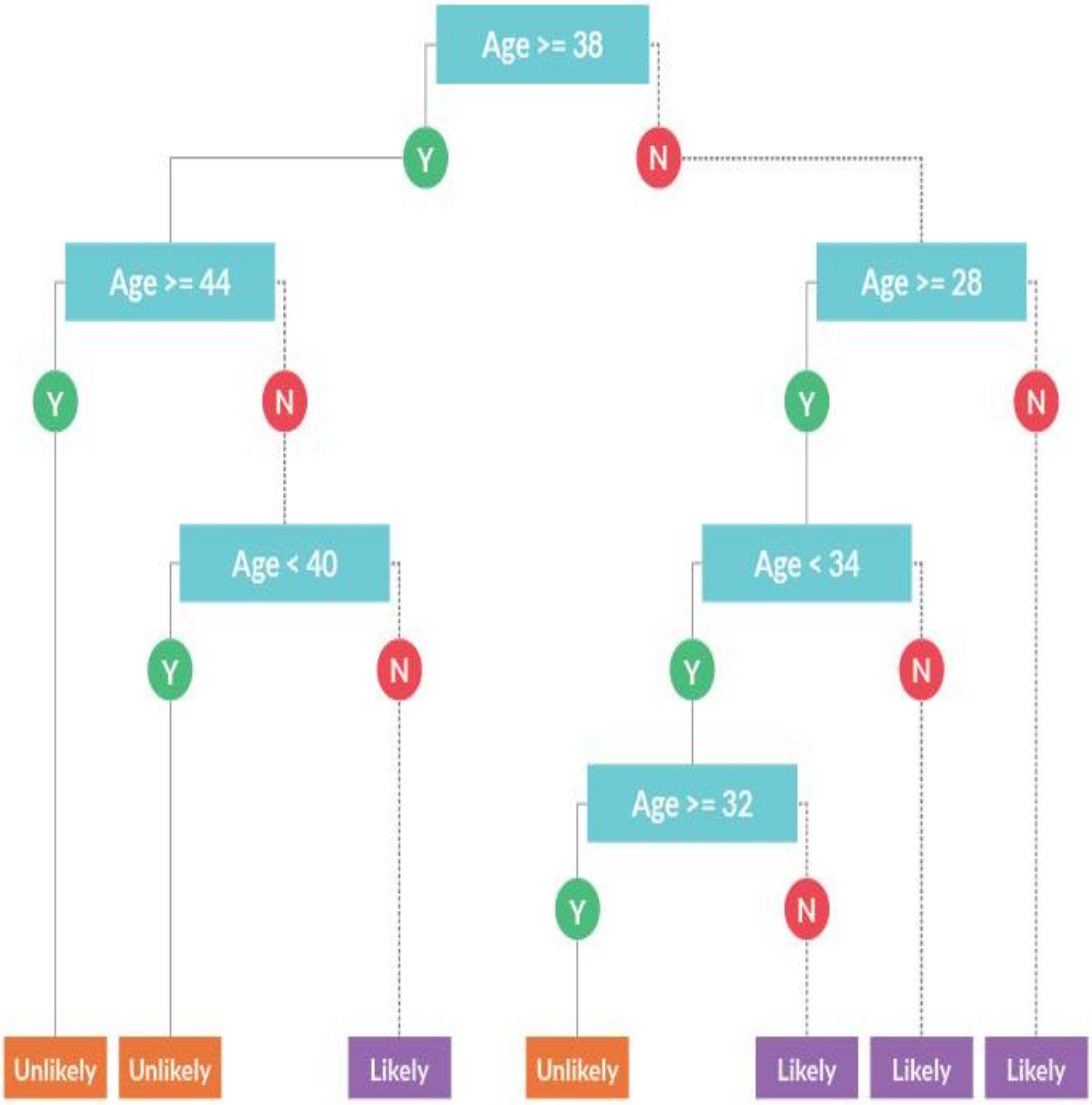

Figure 3.6: Example of the decision tree

6.2- Algorithm

In general, the decision tree algorithm looks like this:

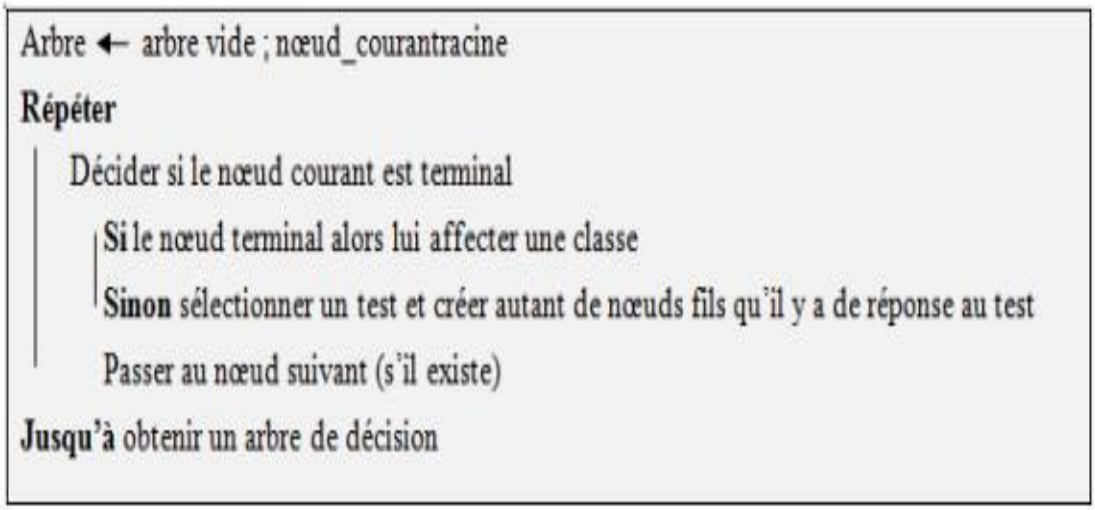

Figure3.7: the decision tree algorithm.

6.3- Criticisms of the method

The decision tree is a method widely used for reasons of efficiency and simplicity compared to other existing methods; indeed, it is very understandable for all users since its rules are of type If... It is based on the simultaneous use of qualitative and quantitative variables (discrete or continuous). Its classification is fast: to classify a new object, we go through a single path from the root tree to the leaf that corresponds to its class. On the other hand, its performance is less good when classes are numerous, the trees can be very complex and are not necessarily optimal. Building decision trees usually takes a long time because you have to find the right choice of attributes. If the data changes over time, it is necessary to relaunch the learning phase on a complete sample that contains new and old examples. [16]

6.4- The fields of application

This method can be used in several fields such as: Studies (to understand the main criteria in the purchase of a product, the impact of advertising expenses), sales (to analyze performance by region, by brand, by vendor), risk analysis (to detect predictive factors for non-payment behavior), The medical field (to study the relationships between certain diseases and physiological or sociological peculiarities).

7- algorithm/J48

The C4.5 algorithm is a classification algorithm which produces decision trees based on information theory. It is an extension of Ross Quinlan's earlier ID3 algorithm also known in Weka as J48, J standing for Java. The decision trees generated by C4.5 are used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

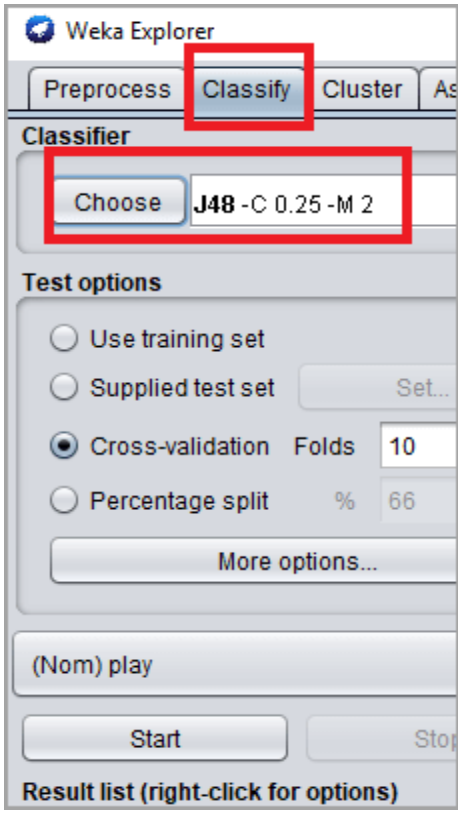

Figure 3.8: J48 (classifier).

The J48 implementation of the C4.5 algorithm has many additional features including accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open-source Java implementation of the C4.5 algorithm. J48 allows classification via either decision trees or rules generated from them. [17]

8- Random Forests

The random forest can be defined as a versatile and smart machine learning method that can perform both classifications as well as regression tasks. It can also perform dimensional reduction methods, outlier values, treat missing values and other steps of data exploration as well. It is an expert solution for most of the problems. It is known as an ensemble way of learning as a group of weak models are combined to form this, powerful model.

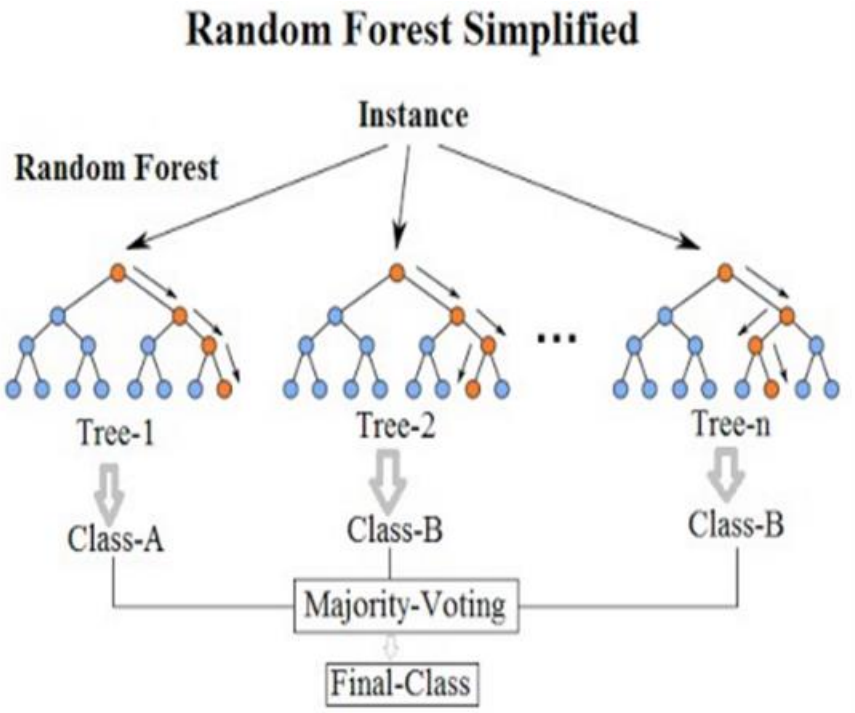

Figure 3.9: Random forest simplified

8.1- Advantages of Random Forest

● Random forest algorithm can be used in both sorts of problems. It can be used in regression as well as classification.

● It can handle a large amount of data set in high dimensionality. It can handle more than a few thousands of variables and can very well identify the significant ones among them. It is therefore considered to be an important dimensionality reduction method.

● It can effectively estimate the missing data and can easily maintain accuracy even if it is fed a large amount of data.

● It has various methods that can be used to balance errors in the data set.

● The above features can be used with unlabeled data as well. Thus, it can work unsupervised.

● It samples input data with replacement. This process is known as bootstrap sampling.

8.2- Disadvantages of Random Forest

● Like advantages, Random Forest has certain disadvantages too. However, as compared to the disadvantages, the number of advantages is large, thus making Random forests a far better option than other options.

● It is not as good at regression as it is with classification. It often does not come out with precise, continuous nature predictions. It cannot make predictions beyond the range of the provided training data in the case of regression.

● The data may become over-fit if the sample data is too noisy.

● It can act as a black box approach for statistical modelers as you cannot control the performance of the model. You can only try random seeds and different parameters.

9- Voting:

It is an ensemble method that combines the performances of multiple models to make predictions.

Voting only serves to benefit when the machine learning classifiers perform at similar levels. A voting estimator built from models with contrasting levels of efficiency may perform erratically.

9.1- Benefits of Voting

Incorporating voting comes with many advantages.

Firstly, since voting relies on the performance of many models, they will not be hindered by large errors or misclassifications from one model. A poor performance from one model can be offset by a strong performance from other models. To better understand this concept, let's use investing as an analogy.

When you invest, you are often advised to allocate your wealth in a variety of shares. If you put all of your money in one stock, your entire portfolio relies on the performance of that one stock, subjecting you to high risk. On the other hand, if you put your money in a variety of shares, that risk is mitigated; one poor-performing stock will not foil your success.

Voting works in a similar manner. By combining models to make a prediction, you mitigate the risk of one model making an inaccurate prediction by having other models that can make the correct prediction. Such an approach enables the estimator to be more robust and prone to overfitting.

In classification problems, there are two types of voting: **hard voting** and **soft voting**. Hard voting entails picking the prediction with the highest number of votes, whereas soft voting entails combining the probabilities of each prediction in each model and picking the prediction with the highest total probability.

Voting in regression problems is somewhat different. Instead of finding the prediction with the highest frequency, regression models built with voting take the predictions of each model and compute their average value to derive a final prediction.

In either classification or regression problems, voting serves as a means to enhance predictive performance.

9.2- Drawbacks of voting

As you learn about the benefits of combining models, you might be tempted to just use voting in all of your future machine learning projects. After all, why go through the trouble of considering the pros and cons of each model when you can just chug them all into one estimator and create the "perfect model"? Unfortunately, this line of thinking is flawed. Understand that models built with voting should not be treated as a one-size-fits-all approach in machine learning. After all, the voting ensemble method also has its limitations.

Firstly, there are cases where an individual model can outperform a group of models. For instance, in a regression problem, if the predictive features and the target variable have a strong linear relationship, a single linear regression model can no doubt perform very well. However, a voting estimator made with other regression models will nullify the accurate predictions of the linear regression model. Secondly, since voting requires the use of multiple models, they are naturally more computationally intensive. Thus, creating, training, and deploying such models will be much more costly.

10-Vector Support Machines (or VSM)

Vector support machines (VSM, SMO in WEKA) are at the origin of new methods of categorization, although the first publications on the subject date back to the 1960s. [18] The VSM principle consists of a structural risk minimisation strategy but the problem is to find a decision boundary that separates the space into two regions, to find the hyperplan that classifies the data correctly and which is the most far from all examples. We say we want to maximize the margin, which means the distance from the closest point to the hyperplane. [18] In the case of text categorization, entries are documents and exits are categories.

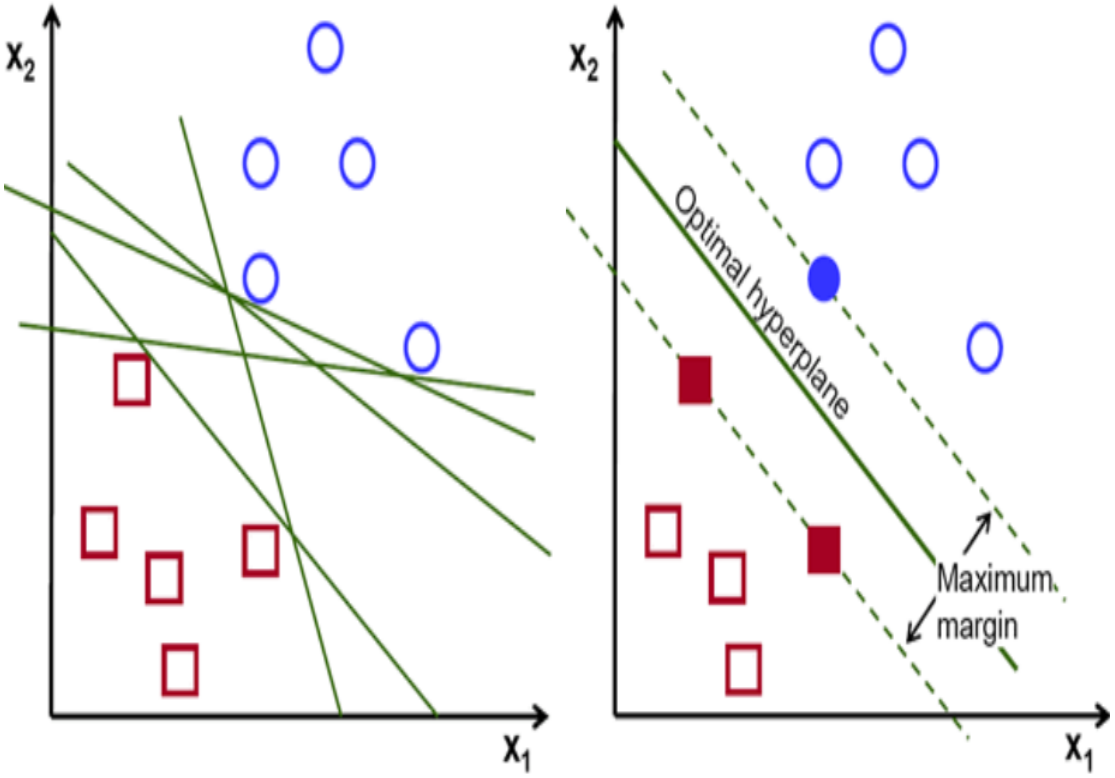

Figure 3.10: Case of text categorization with VSM.

VSM is well suited for the classification of texts because a high dimension does not affect them since they protect against over-learning. In other words, he argues that few attributes are completely useless to the classification task and that VSM can avoid aggressive selection that would result in loss of information. We can afford to keep more attributes. Also, a characteristic of textual documents is that when they are represented by vectors, a majority of entries are null.[21] However, VSMs are well suited to so-called sparse vectors. Another positive aspect of VSM is that no manual parameter adjustments are required, as they have the ability to automatically find suitable parameters. VSM was introduced by Vapnik (1995) and developed dazzling on the one hand because of the mathematical beauty of their concept and on the other hand because of their excellent performance in terms of error rates in a large number of problems such as the recognition of shapes, genomic sequences, spam, etc.

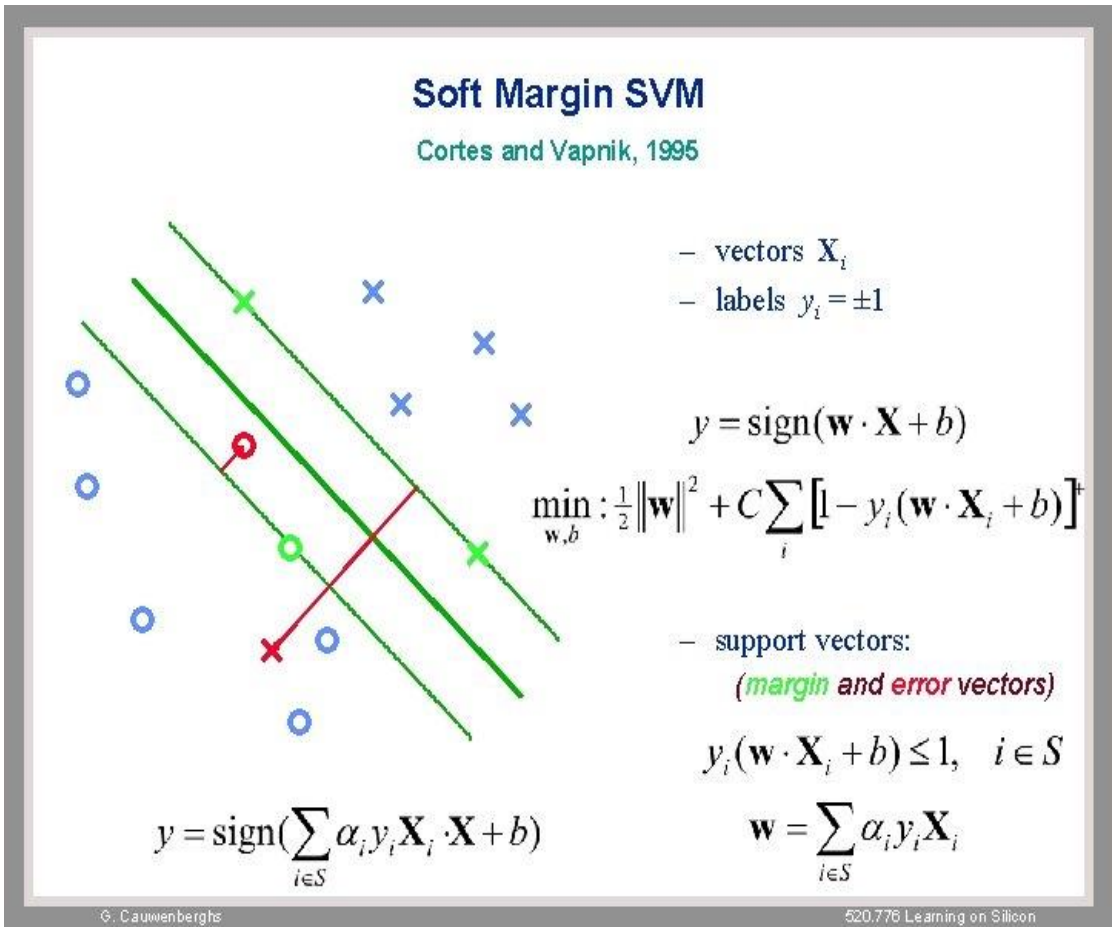

Figure 3.11: Soft Margin VSM (cortes and Vapnik 1995).

They were originally conceived as a classifier for binary target, but were extended to the categorical case (at the price of the same contortions as the RLog) and to the case of regression at the price of a mathematical complexity that may be repulsive. On the other hand, a judicious selection of hyper parameters (there are several to choose from, especially the VSM nuclei) makes it possible to treat mixed characteristics and many non-standard cases, such as text classification. But this selection implies, in the adjustment phase, a complex engineering of nuclei that is risky to perform without expert intervention since these nuclei are functions. To date, there does not appear to be any diagnostic tools indicating whether the selection is well or poorly suited.

The integration of cognitive elements of C, the treatment of missing characteristics and the study of the importance of variables are not yet well developed. For these problems, ad ho c solutions exist, but they also require human supervision and their reliability is not yet well established. It is possible (through the Platt algorithm) to calibrate their scores for prescriptive analytical purposes. Finally, the construction of an VSM classifier requires an optimization which, without being too

dangerous, is no less demanding from an IT point of view, since the problem becomes more complex than the simple binary classification.

Conclusion

In addition to the available tools (Explorer, Experiment, Knowledge Flow), Weka provides a user with a collection of machine learning algorithms. Among these algorithms, we quote in this chapter the most commonly used. These algorithms can be used for either regression or classification. Regression focuses on finding a correlation relationship between inputs (the different independent variables) and outputs (the dependent variable or variables).

The purpose of classification is to predict the class of a test instance. These algorithms are applied in supervised mode, guided by the output (or target class).

CHAPTER 4

LEARNING AND TESTING

1- Introduction

After having presented in the previous chapter our method and the tools we goanna use for the classification. In this chapter we present the application and description of the data base used and explain it, and the experimentation and the result which resulted and discussion of the result.

2- Dataset used NSL-KDD

NSL-KDD is an updated version of KDD cup99 data set which suggested to solve some problems of previous version. This data set is an effective benchmark for researchers to compare different types of Intrusion detection system (IDS) methods, build an Intrusion detection system (Host based or Network based), doing for some experiments in Cyber security area likewise there is so many advantages.

He competition task was to build a network intrusion detector, a predictive model capable of distinguishing between "bad'' connections, called intrusions or attacks, and "good'' normal connections. [5] The class variable has two categories:

• Normal

• Anomaly

2.1- Basic Features of Each Network Connection Vector

Network Connection Vector	Description
Duration	Length of time duration of the connection
Protocol type	Protocol used in the connection
Service	Destination network service used
Flag	Status of the connection – Normal or Error
Src bytes	Number of data bytes transferred from source to destination in single connection
Dst bytes	Number of data bytes transferred from source to destination in single connection
Land	if source and destination IP addresses and port numbers are equal then, this variable takes value 1 else 0
Wrong fragment	Total number of wrong fragments in this connection

urgent	Number of urgent packets in this connection. Urgent packets are packets with the urgent bit Activated
Hot	Number of „hot“ indicators in the content such as: entering a system directory, creating programs
num failed logins	Count of failed login attempts
logged in	1 if successfully logged in; 0 otherwise
num compromised	Number of compromised conditions
Root shell	1 if root shell is obtained; 0 otherwise
Su attempted	1 if “ su root ” command attempted or used; 0 otherwise
Num root	Number of ``root'' accesses or number of operations performed as a root in the connection
num file creations	Number of file creation operations in the connection
Num shells	Number of shell prompts
num access files	Number of operations on access control files
num outbound cmds	Number of outbound commands in an ftp session
is host login	1 if the login belongs to the ``hot'' list i.e., root or admin; else 0
is guest login	1 if the login is a ``guest'' login; 0 otherwise
count	Number of connections to the same destination host as the current connection in the past two
srv count	Number of connections to the same service (port number) as the current connection in the past two seconds
serror rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in count (23)
srv serror rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in srv_count (24)
Rerror_rate	The percentage of connections that have activated the flag (4) REJ, among the connections
srv_rerror_rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in srv_count (24)

same_srv_rate	The percentage of connections that were to the same service, among the connections aggregated in count (23)
diff_srv_rate	The percentage of connections that were to different services, among the connections aggregated in count (23)
srv_diff_host_rate	The percentage of connections that were to different destination machines among the connections aggregated in srv_count (24)
dst_host_count	Number of connections having the same destination host IP address
dst_host_srv_count	Number of connections having the same port number
dst_host_same_srv_rate	The percentage of connections that were to the same service, among the connections aggregated in dst_host_count (32)
dst_host_diff_srv_rate	The percentage of connections that were to different services, among the connections aggregated in dst_host_count (32)
dst_host_same_src_port_rate	The percentage of connections that were to the same source port, among the connections aggregated in dst_host_srv_count (33)
dst_host_srv_diff_host_rate	The percentage of connections that were to different destination machines, among the connections aggregated in dst_host_srv_count (33)
dst_host_serror_rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_count (32)
dst_host_srv_serror_rate	The percentage of connections that were to different destination machines, among the connections aggregated in dst_host_srv_count (33)
dst_host_rerror_rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_count (32)
dst_host_srv_rerror_rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_srv_count (33)

Table 4.1: NSL-KDD basic features

3- Software

Smart Network Intrusion Detector System (SNIDS) is a software developed from existing open source code to learn how to detect attacks. SNIDS uses WEKA API to run machine learning algorithms. Specifically, J48 classifier is used on "J-48 OPTION", and Vote meta-classifier is used combining VSM and RF classifiers on "COMBINE OPTION". Both "J-48 OPTION" and "COMBINE OPTION" are software components. SNIDS was implemented in Java using Eclipse IDE. A GUI was also developed using WindowBuilder.

The program contains the following classes:

Class MainIDS: This is the main class of SNIDS. It shows the main screen of the software. User can run use the J-48 and the Meta-voting by importing a training and a testing dataset.

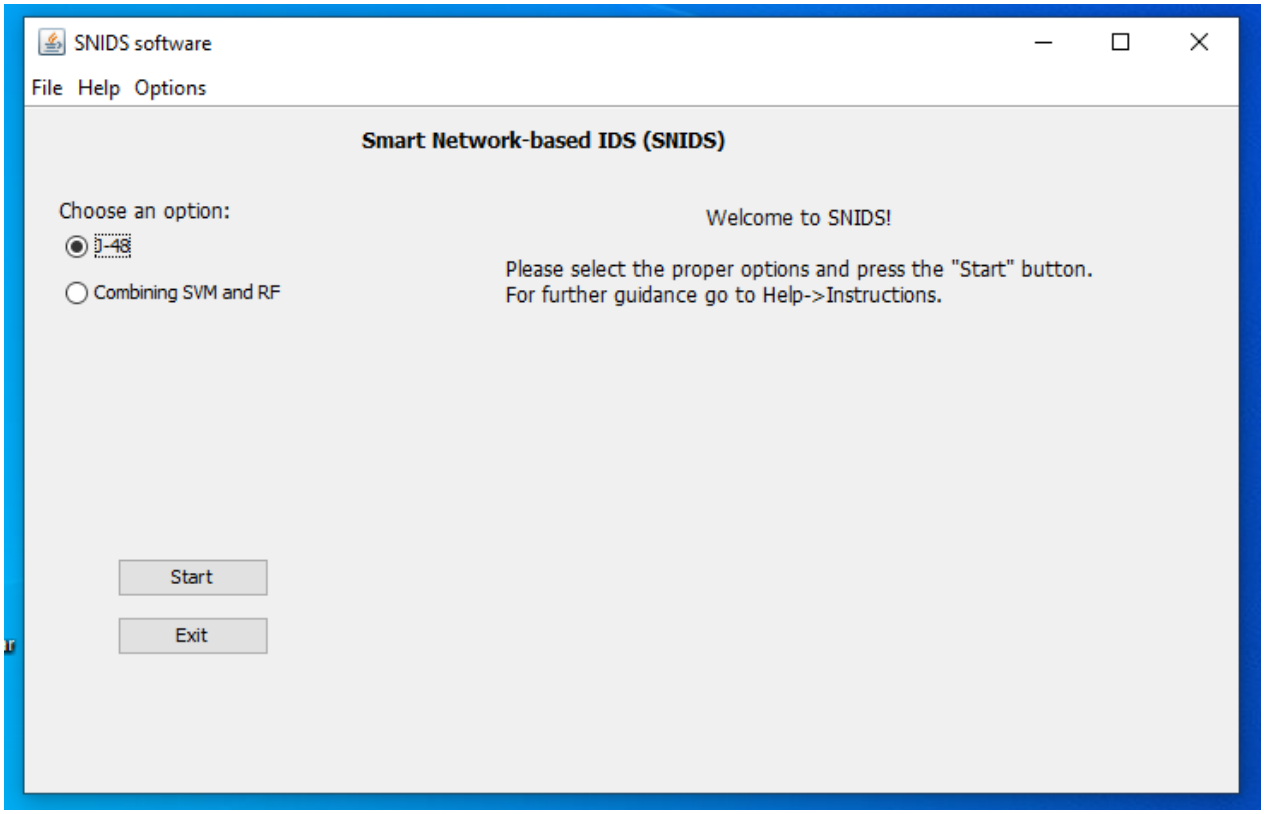

Figure 4.1: The Main window interface

Class MethodeIDS : the class contain the method necessary to use example read dataset from extern file, Method for executing VSM and Random Forest and J-48.

Class OptionHAN : This class is a JFrame. It implements the HAN-let options where user can select parameters for J-48 classifier. All the selections are saved in HANarrayopt.

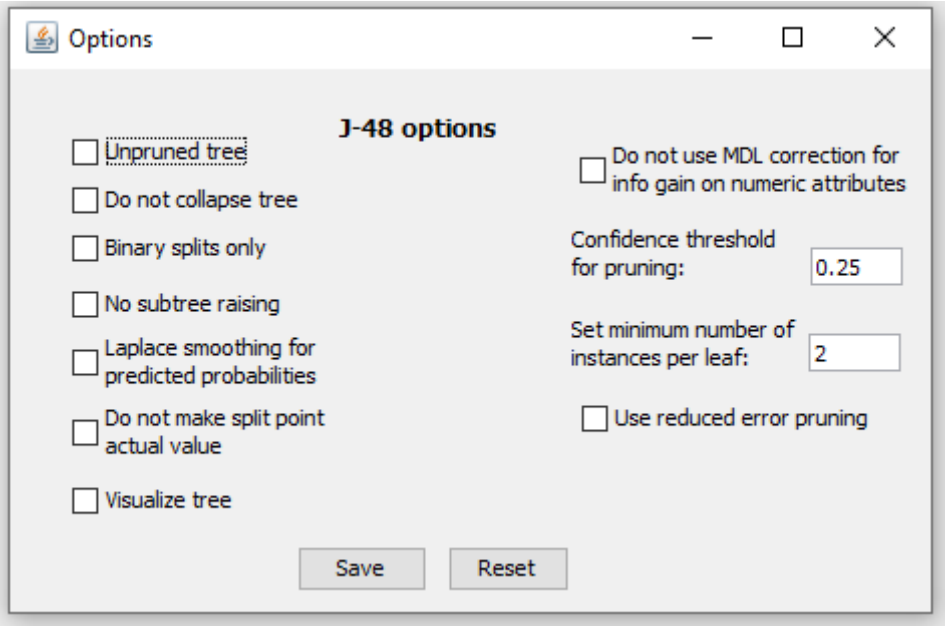

Figure 4.2: The OptionHAN interface

Class OptionNAN : This class is a JFrame. It shows the NAN-let options. The user can select the parameters for VSM and Random Forest.

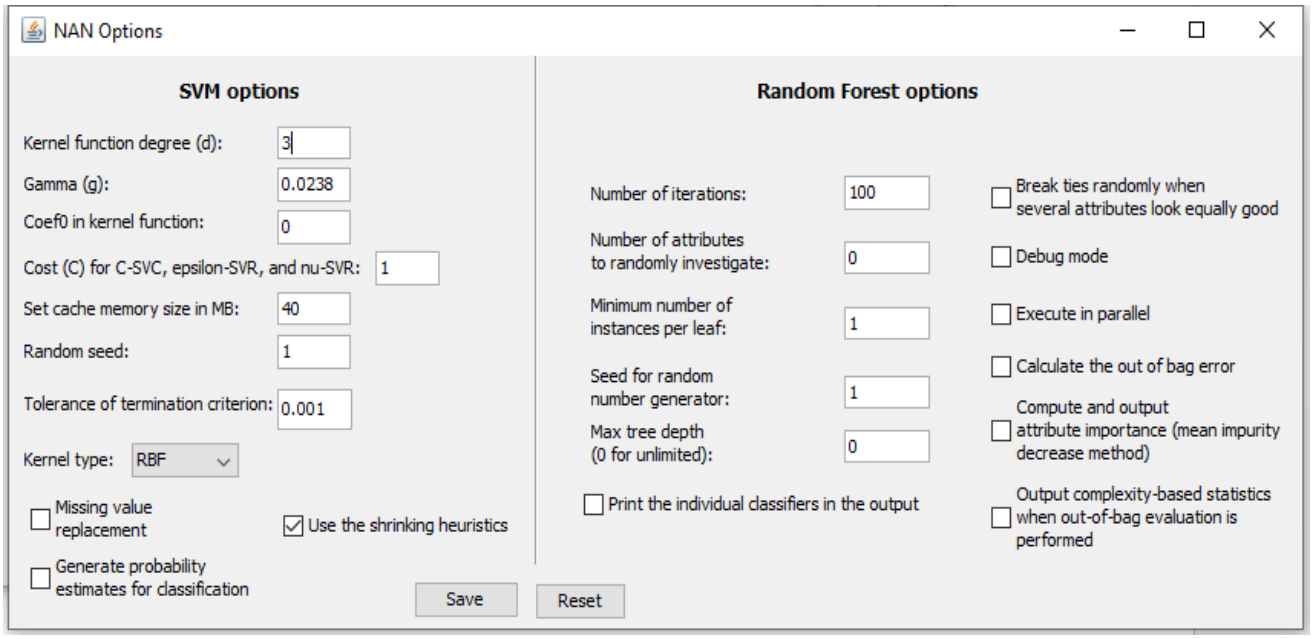

Figure 4.3: The OptionNAN interface

Class OptionGEN : The class implements JFrame. It shows the available options of SNIDS. User can select various parameters such as Load/Save model, choose Vote method, use filter or not and choose test method.

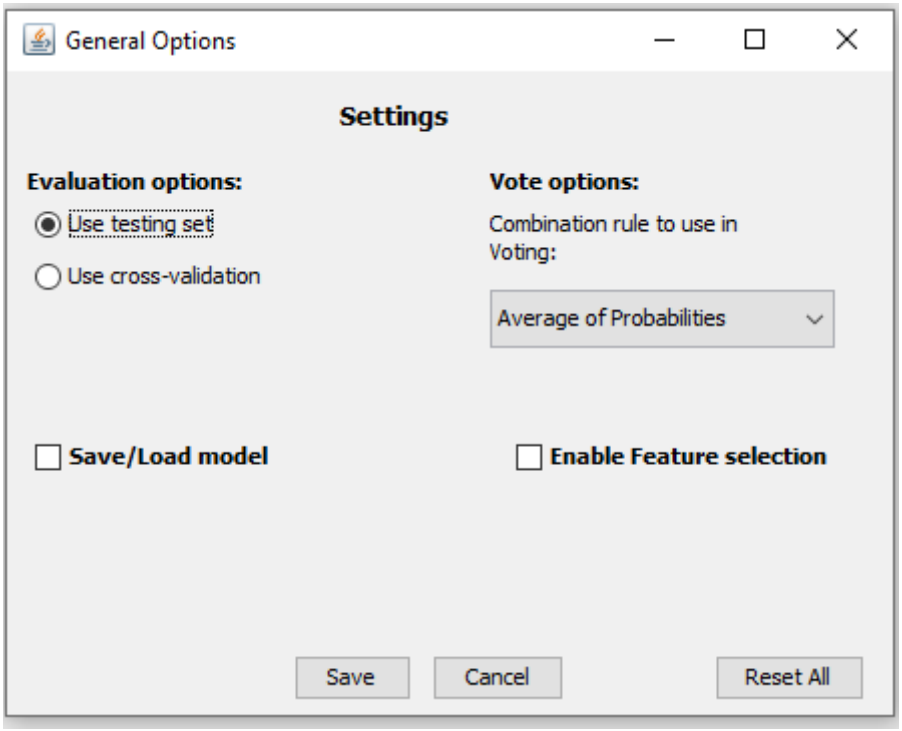

Figure 4.4: The OptionGEN interface

4- Experimentation

In this task to do the classification we used three algorithms J-48 and Vote meta-classifier is used combining VSM and RF classifiers.

With tow approach to evaluate the performance of the model by using train data for training and other file contain data test or using cross-validation.

4.1- Use testing set

Split the dataset manually using another program for example. Prepare the model on the entire training dataset and use the separate test set to evaluate the performance of the model. This is a good approach when we have a large dataset (many tens of thousands of instances). [6]

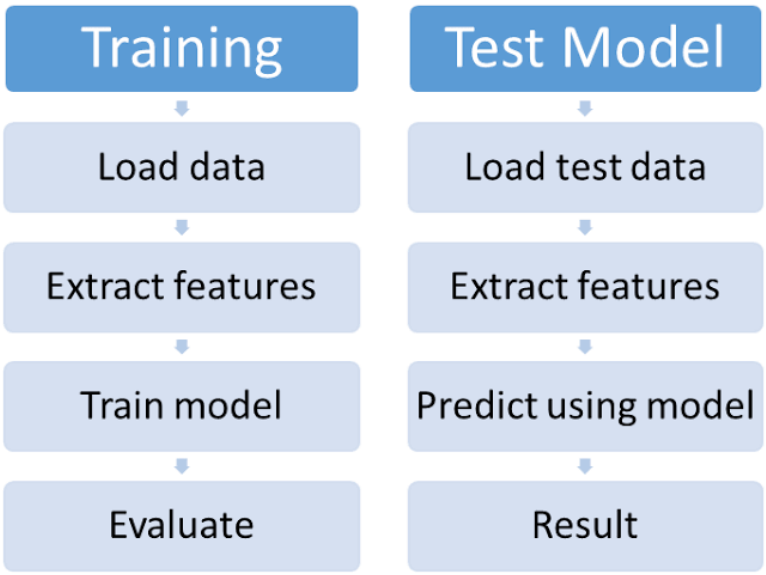

Figure 4.5: Testing set approach

4.2- Cross validation

Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations.

Cross-Validation has two main steps: splitting the data into subsets (called folds) and rotating the training and validation among them. The splitting technique commonly has the following properties:

- Each fold has approximately the same size.
- Data can be randomly selected in each fold or stratified.
- All folds are used to train the model except one, which is used for validation. That validation fold should be rotated until all folds have become a validation fold once and only once.

Each example is recommended to be contained in one and only one fold. [7]

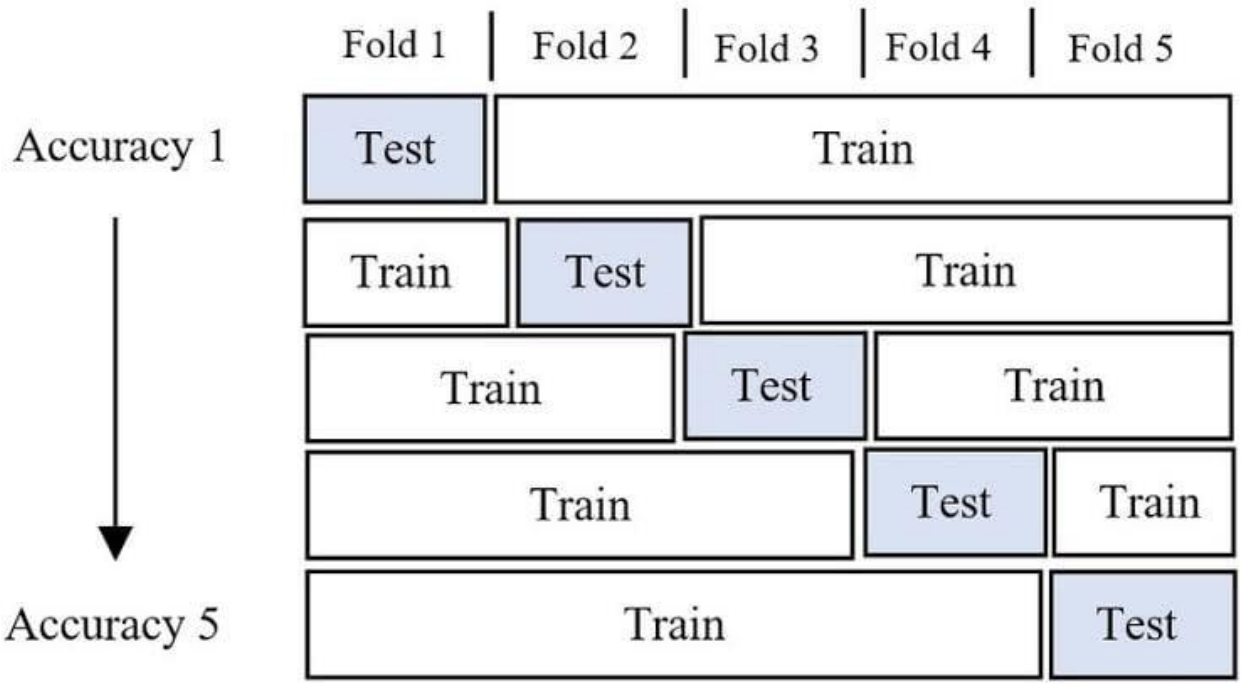

Figure 4.6: Cross-validation approach

5-Testing with j-48

The processing of the dataset generates a decision tree. Figure 4.7 shows the display interface of data classification results.

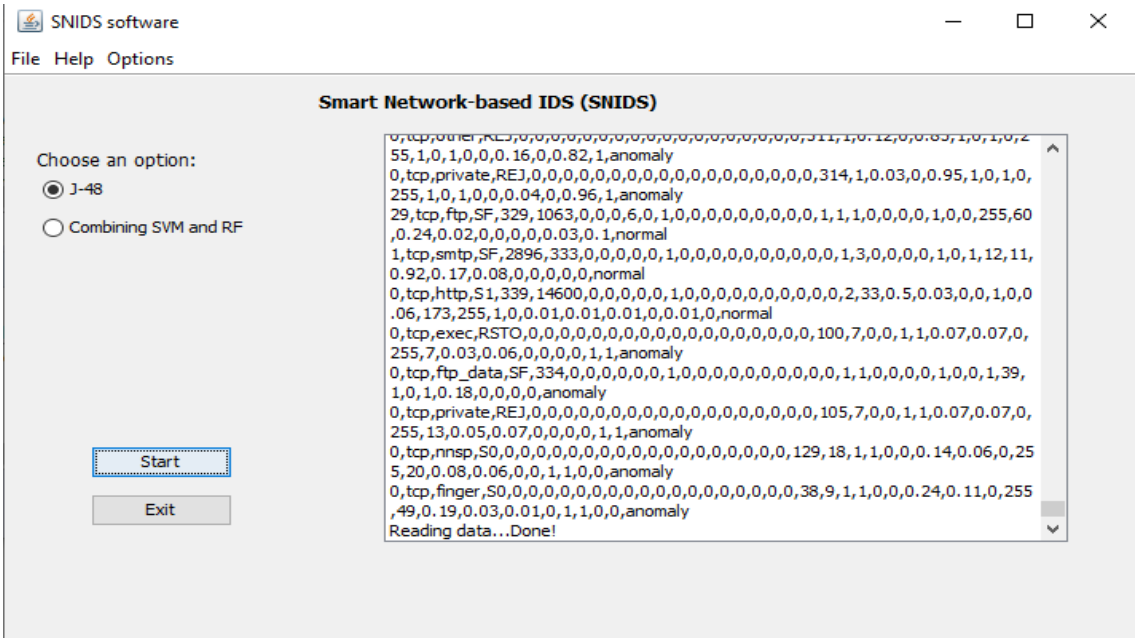

Figure 4.7: Classification J-48 results

5.1- Result j-48

Abstract:

Metrics	Value	Parcentage
Correctly Classified Instances	18273	81.0548 %
Incorrectly Classified Instances	4271	18.9452 %
Kappa statistic	0.6309	
Mean absolute error	0.1792	
Root mean squared error	0.413	
Relative absolute error		35.5072 %
Root relative squared error		81.6584 %
Total Number of Instances	22544	

Table 4.2: the result of classification with J-48

Detailed Accuracy By Class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Normal	0.972	0.311	0.703	0.972	0.815	0.666	0.874	0.777
Anomaly	0.689	0.028	0.970	0.689	0.995	0.666	0.874	0.906
Weighted Avg	0.811	0.150	0.855	0.811	0.810	0.666	0.874	0.850

Table 4.3: The result by class of classification with J-48

Confusion Matrix :

a	b	classified as
9436	275	a = normal
3996	8837	b = anomaly

Visualize tree :

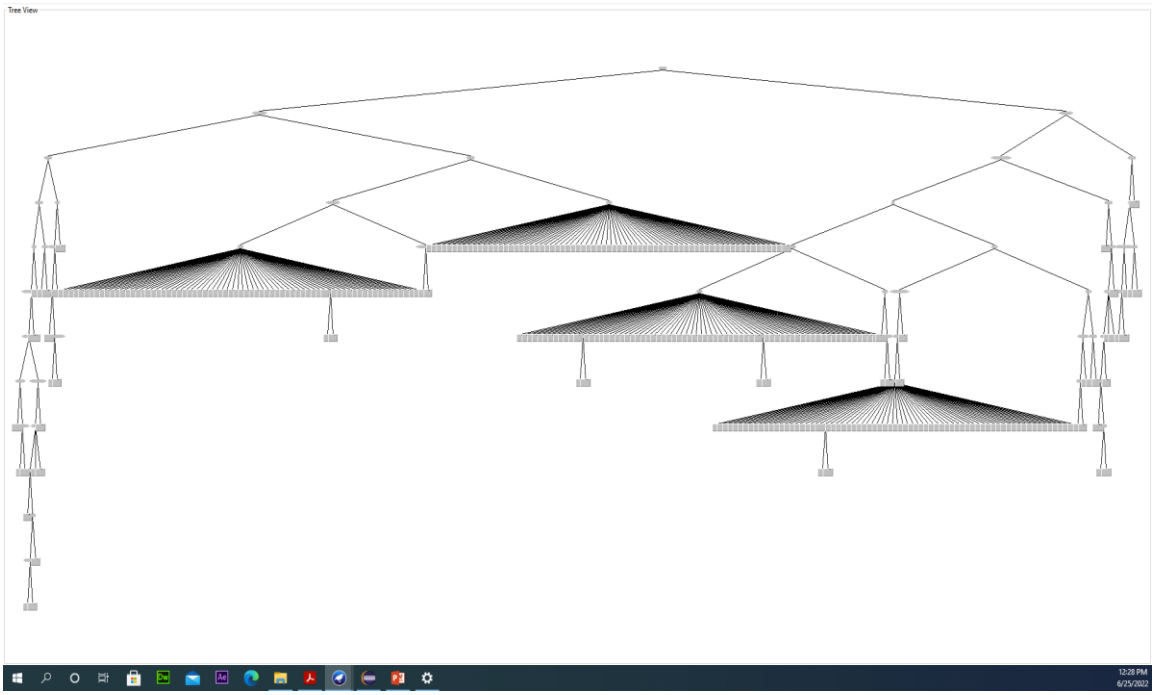

Figure 4.8: Shape of the decision tree for J-48

Note: this is just the shape of the decision tree on a 42 inches' screen, in order to see details that need a bigger screen.

We keep the high rate of the Kappa metric (Dice) which is 0.6309 which shows good choice from the training data considering we use only 20% of the NSL-KDD dataset. Also, a detection system based on this classifier is considered reliable and efficient, because there are only 0.3691 according to Dice which represents the false classification of intrusion data.

6- Combining VSM and Random Forest

we use voting to Combining VSM and RF is technique that used to improve model performance. It predicts the class with the largest summed probability from models. There are different parameters of voting that can be used:

- average probabilities

- product of probabilities
- majority voting
- maximum probability
- minimum probability

in this test we average probability.

6.1- Result of combining VSM and RF

Abstract :

Metrics	Value	Parcentage
Correctly Classified Instances	17252	76.5259 %
Incorrectly Classified Instances	5292	23.4741 %
Kappa statistic	0.5483	
Mean absolute error	0.2215	
Root mean squared error	0.4074	
Relative absolute error		43.8794 %
Root relative squared error		80.539 %
Total Number of Instances	22544	

Table 4.4: The results of classification using voting between VSM and RF

Precision details per classes :

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Normal	0.973	0.392	0.653	0.973	0.781	0.600	0.956	0.945
Anomaly	0.608	0.027	0.968	0.608	0.747	0.600	0.956	0.958
Weighted Avg	0.765	0.184	0.832	0.765	0.762	0.600	0.956	0.953

Table 4.5: the result by class of classification using voting between VSM and RF

Confusion Matrix :

a	b	classified
9452	259	a = normal
5033	7800	b = anomaly

7- Discussion of results

we applied various classification algorithms for intrusion detection in NSL-KDD dataset software platform developed in java, According to the results obtained with the J-48 classifiers and combining VSM and random forest classifiers with the J-48 classifier can get a high accuracy rate when comparing with the voting between VSM and random forest, the correctly classified instance of J-48 algorithm is 81.0548 % and the correctly classified instance of combining VSM and random forest is 76.5259 % from this accuracy rate, we can decide that the J-48 classifiers algorithm is the best classification algorithm for NSL–KDD datasets comparing to the voting between VSM and random forests.

Conclusion

In this final chapter, we presented the experimental part of our work, which consists of testing the snooping data of the NSL-KDD database, in a Java developer program we tested the Voting and C 4.5 classifiers by introducing various performance metrics including the dice index (Kappa) and the correctly classified instance. In both cases, the two classifiers performed well. However, we found that classifier J-48 scored slightly ahead of VSM and random forests.

General Conclusion

Today, computer attacks represent a real risk that threatens computer systems and corporate networks, which has led us to try in this work to develop a security model capable of confronting this threat by detecting any previously known or recent malicious attempt. To achieve this goal, we carried out a comparative study of different types of classifiers, and this to classify the connections in two categories: normal or attack based on the NSL-KDD dataset.

Finally, since scientific research has no limits and there are possible improvements to perfect this model such as the use of combinations of classifiers to choose the other parameters of the network and also the realization of a multi-class classification which also gives the type of attack detected.

BIBLIOGRAPHY

[1]: John Snowden, Cyber Security: An Ultimate Guide to Cybersecurity, Cyberattacks, and Everything You Should Know About Being Safe on The Internet, 2021.

[2]: Chris Jackson, Network Security Auditing, Cisco Systems, Inc. Cisco Press, 2010.

[3]: Dan Shoemaker, Anne Kohnke, and Ken Sigle, Guide to the National Initiative for Cybersecurity, Internal Audit and IT Audit Series, 2016.

[4]: John Vacca, Computer and Information Security Handbook, Morgan Kaufmann Publishers Inc, 2009.

[5]: https://www.kaggle.com/datasets/sampadab17/network-intrusion-detection .2022/05/10

[6]: What is data splitting and why is it important? (techtarget.com). 06/06/2022

[7]: Jason Brownlee, Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-to-end, Jason Brownlee, 2016

[8]: Yeo, L.H., Che, X., & Lakkaraju, S, Understanding Modern Intrusion Detection Systems: A Survey, arXiv: Cryptography and Security, 2017.

[9]: Ben Brahim Embarka, Amiche Selyna, Mise en place d'une solution de détection d'intrusion, Mémoire de Master Académique, Spécialité : Réseaux & Télécommunications, University of Tizi Ouzou, 2017.

[10]: Fraser Sherman, The Differences Between a Firewall and an Intrusion Detection System, https://smallbusiness.chron.com/differences-between-firewall-intrusion-detection-system-62856.html. 06/06/2022

[11]: Le Grand Livre de la Sécurité Informatique, SecuriteInfo, Edition 2006.

[12]: Laurent Bloch-Christophe Wolfhugel, Sécurité informatique, Eyrolles, 2ème édition, 2005.

[13]: https://www.alertlogic.com. 06/06/2022

[14]: William Sullivan, Machine Learning for Beginners - Algorithms, Decision Trees & Random Forests - An Introduction, 2017.

[15]: Eibe Frank, Mark A. Hall, and Ian H. Witten, Weka workbench, 2016.

[16]: Bourass kamel, Boulmaiz walid, Etude Comparative des Méthodes de Détection d'intrusions, Mémoire de Master Académique, Option : Réseaux et Systèmes Distribués, Université 20 Aout 1955 de Skikda, 2020/2021.

[17]: https://medium.com/@nilimakhanna1. 26/02/2022

[18]: Touina Hanane, Classification automatique de textes, Mémoire de Master Académique en Informatique, Option : Systèmes d'Information Avancées, Université Mohamed Boudiaf, M'Sila, 2018.

[19]: Ralf Mikut, Markus Reischl, Data mining tools, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011.

[20]: J.R. Quinlan, Induction of Decision Trees, Machine Learning, (1), 81-106, 1989.

[21]: Bilel Bahloul, Méthode pour l'analyse automatique d'opinion de la langue arabe, Mémoire de Master en Informatique, Option : Traitement Automatique de la Langue, Université Saad Dahleb, Blida, 2019.