PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
Ministry of Higher Education and Scientific Research
Mohamed Khider University – BISKRA
Faculty of Exact Sciences, Natural sciences and Life
Department of Computer Science

Order N°: RTIC29/M2/2022

THESIS

Presented for the Academic Master's degree in

Computer Science

Option: Information and communication networks and technologies

---

Cyber-Attacks prediction using Data Mining technics

By:

Sedrata Mohamed

Presented in 26/06/2022 Board of Examiners:

:

Mrs. Aloui Imene	**President**
Mrs. Berima Salima	**Supervisor**
Mr. Hamida Ammar	**Examiner**

Academic year 2021-2022

Summary

List of Figures

List of Tables

Dedication

First,I give thanks to Allah who helped me and gave me the strength and patience to endure all the difficulties to complete the work and who taught us the purpose of life .

First,I dedicate this work to my deceased father Bennadji, he was a good father who gave me everything, I hope he is in heaven and thank so mush to my great mother Miloudi Nassima, who gave me courage and financial and psychological support and always by my side.I would like to sincerely thank my deep gratitude to Mrs. Berima Salima, as supervisor of the dissertation. She has always been careful and advised, and for the effort .

I also dedicate this dissertation to all my family ,my brother and my sisters,and all my friends who always have supported me throughout the process. I will always appreciate all they have done.

Abstract

Cyberattacks are the most common concern right now, which is very much about diversion. If a person does not have a suitable security system, linked information can be hacked easily. One of the most frequent causes of cyberattacks is because of intruders. Therefore, it has enhanced the security process by using machine learning algorithms and prediction and with the help of artificial intelligence (AI) to avoid cyber attacks.

In this study, we will focus on the precision and approximation of the correct result in case of attack or not through the use of machine learning algorithms, so there are ways to detect and prevent attacks and protect them from attackers like IDS, IPS, Firewall and They just reduce and don't get the job done.

We propose protective methods such as machine learning, we gonna use Linear regression to predict data that allows the machine to learn and predict using the data we study, and we specifically propose a Linear regression algorithm for analyzing those inputs, where we propose a dataset imported from the Kaggle Internet with the addition of data with a suggestion a template for including normal and unusual transactions.

Keywords: Cyber attacks, Linear regression, prediction , data processing, training, testing

Résumé

Les cyberattaques sont la préoccupation la plus courante à l'heure actuelle, et il s'agit surtout de déjudiciarisation. Si une personne ne dispose pas d'un système de sécurité approprié, l'information liée peut être piratée facilement. Une des causes les plus fréquentes de cyberattaques est à cause des intrus. Par conséquent, il a amélioré le processus de sécurité en utilisant des algorithmes d'apprentissage automatique et de prédiction et à l'aide de l'intelligence artificielle (IA) pour éviter les cyberattaques.

Dans cette étude, nous nous concentrerons sur la précision et l'approximation du résultat correct en cas d'attaque ou non grâce à l'utilisation d'algorithmes d'apprentissage automatique, afin qu'il y ait des moyens de détecter et de prévenir les attaques et de les protéger contre les attaquants comme IDS, IPS, Firewall et ils ne font que réduire et ne font pas le travail.

Nous proposons des méthodes de protection telles que l'apprentissage automatique, nous allons utiliser la régression linéaire pour prédire les données qui permettent à la machine d'apprendre et de prédire en utilisant les données que nous étudions, et nous proposons spécifiquement un algorithme de régression linéaire pour analyser ces entrées, où nous proposons un ensemble de données importé de Kaggle Internet avec l'ajout de données avec une suggestion un modèle pour inclure les transactions normales et inhabituelles.

Mots-clés : Cyberattaques, Régression linéaire, prédiction, traitement des données, formation,tester

ملخص

الهجمات الإلكترونية هي الشاغل الأكثر شيوعًا في الوقت الحالي، وهو يتعلق إلى حد كبير بالتحويل. إذا لم يكن لدى الشخص نظام أمان مناسب، فيمكن اختراق المعلومات المرتبطة بسهولة. أحد أكثر أسباب الهجمات الإلكترونية شيوعًا هو الدخلاء. لذلك، فقد عززت عملية الأمان باستخدام خوارزميات التعلم الآلي والتنبؤ وبمساعدة الذكاء الاصطناعي (AI) لتجنب الهجمات الإلكترونية.

في هذه الدراسة، سنركز على دقة وتقريب النتيجة الصحيحة في حالة الهجوم أم لا من خلال استخدام خوارزميات التعلم الآلي، لذلك هناك طرق لاكتشاف ومنع الهجمات وحمايتها من المهاجمين مثل IDS و IPS و جدار الحماية وهم فقط يقللون ولا ينجزون المهمة.

نقترح طرق وقائية مثل التعلم الآلي، سنستخدم الانحدار الخطي للتنبؤ بالبيانات التي تسمح للآلة بالتعلم والتنبؤ باستخدام البيانات التي ندرسها، ونقترح على وجه التحديد خوارزمية الانحدار الخطي لتحليل تلك المدخلات، حيث نقترح مجموعة بيانات مستوردة من إنترنت كاغل مع إضافة بيانات مع اقتراح نموذج لإدراج المعاملات العادية وغير العادية.

الكلمات المفتاحية : الهجمات الإلكترونية، الانحدار الخطي ، والتنبؤ، ومعالجة البيانات، والتدريب، والاختبار
.

General introduction

Cyberattacks have become one of the world's greatest challenges. Each day they cause serious financial damage to countries and peoples. Key factors in the fight against crime and criminals are the identification of those responsible for cybercrime and the understanding of attack methods.
The detection and avoidance of cyberattacks are challenging tasks. Researchers have recently solved these problems by developing safety models and making predictions using AI methods. A lot of crime prediction methods are available in the literature. Also, they suffer from a disability in the prediction of cybercrime and cyberattack methods. This problem can be solved by identifying the attacker, using real data. The data includes the type of crime, the sex of the perpetrator, the damage and the methods of attack.

Data can be obtained from the applications of individuals who have been exposed to cyberattacks against judicial units, we analyze cyberattack with machine learning methods and predict the effect of the defined characteristics on the detection of the cyberattack method and the perpetrator.
In the first model, we could predict with great precision the types of attacks to which victims might be exposed.
Logistic regression was the most used method to detect attackers with an accuracy ratio of 65.42/100. In the second model, we predicted whether authors were identifiable by comparing their characteristics.

This thesis is divided into four chapters: In chapter 1, dedicated to study everything related to cybersecurity like definition,importance,causes and damage of cyberattacks.
In chapter 2, describe the field of Data Mining and explain its efficacity to solve the problem of cyberattack.
In chapter 3, introduce the design of our system to secure the cyber from attack.
Finally, in the last chapter, we will proceed to the development of a system which allows to train the system based on regression on different incidents (database) occurred and their classification then predict if it generate an attack or not, which is the main objective of this project.

Chapter 1

Cybersecurity

1.1 Introduction

secure information infrastructure in cyberspace develop the capacity to prevent and respond to cyber threats, reduce vulnerabilities and minimize damage caused by cyber incidents through a combination of institutional structures, people, processes, technology and co-operation.

This chapter describes cybersecurity, where we will try to explain the most important aspects of this field to give an overview of the topic. [32]

1.2 Objectives

Cyber security is a fundamental task to protect sensitive information on the Internet and the devices that protect it from attack, destruction or unauthorized access. [21]

The goal of cyber security is to ensure a secure and risk-free environment to keep data, network and devices protected from cyber threats. It also protects:

- Protect your bank balance from being rolled over to a random hacker.
- Protect your on-line accounts and devices from compromise.
- Protect your privacy from being violated like switching on your laptop's camera to surprise you.
- Protect your information from interception in the process of transfer to a trusted entity.
- Protect your intel from being exposed by a spy.
- Protection of an organization from loss of internal data.
- Protect some software from cracking.
- Check the source of a data package or communication generally. [13]

1.2.1 Why is cybersecurity important ?

It protected data and systems against cyberattacks from infrastructure connected to the Internet, including hardware, software and data.
In a computer context, security includes cybersecurity and physical security, both of which are used by companies to prevent unauthorized access to data centers and other computerized systems. Security, which is intended to protect the confidentiality, integrity and availability of data, is a subset of cybersecurity.[27]

1.2.2 What Are the fundamentals of Information Security ?

The main purpose of information security is to protect information resources from threats and vulnerabilities to which the organization's attack surface can be exposed. Threats and vulnerabilities combined represent an information risk. Ensuring that safety objectives are met and risks are mitigated will be beneficial to an organization by contributing to:

- Business continuity .
- Operational Efficiency .
- Cost Efficiency .

A proper cyber security program should not just protect internal data that a company deems confidential and/or proprietary,It should also protect the personally identifiable information (PII) of its customers. An example of PII is a consumer's social security number, driver's licence number, even his or her email address.

• **Confidentiality :**
Exposure of sensitive or privately held system information.
• **Integrity :**
System failure, characterized by a different outcome according to the order in which the system components and customers act.
• **Availability :**
Is protect the functionality of systems and ensures data is available at the point in time, ensuring that data is available when it is needed. [17]

Figure 1.1: 3 Fundamentals of information security .[8]

1.3 Definition of cyber attack

A cyberattack is basically nothing more than a means to compromise the computer functionality of a victim's networkor to gain unauthorized digital access to a victim's computer by removing barricades.
The Institute for Security Technology Studies at Dartmouth College has defined, A cyberattack is considered to be an attack against a computer system that compromises the confidentiality, integrity or availability of the information contained in that system. Cyber attacks can be classified in different categories from different perspectives, cyber attacks are clustered.[6]

1.4 Causes and damage

These are computers that built to handle many tasks. This includes home computers, most servers, tablets and smart phones.They host most of our financial, organizational and personal data along with our intellectual capital.
They are based on standard commercial components such as Windows, iOS or Linux operating systems. Being in general use gives these computers great flexibility, but also creates many opportunities for hostile players to exploit.
Being built from commercial components lowers costs, but also means that the same hostile actors can realize economies of scale when writing malicious software.

- Computer systems can be disrupted by human error, intentional cyberattacks, physical damage caused by secondary hazards, and electromagnetic impulses (EMP).
- Cyberattacks can take a variety of forms, including amateur piracy, hacktivism, ransomware attacks, cyberespionage, or sophisticated state-sponsored attacks. They can cause Internet or utility failures, leaks

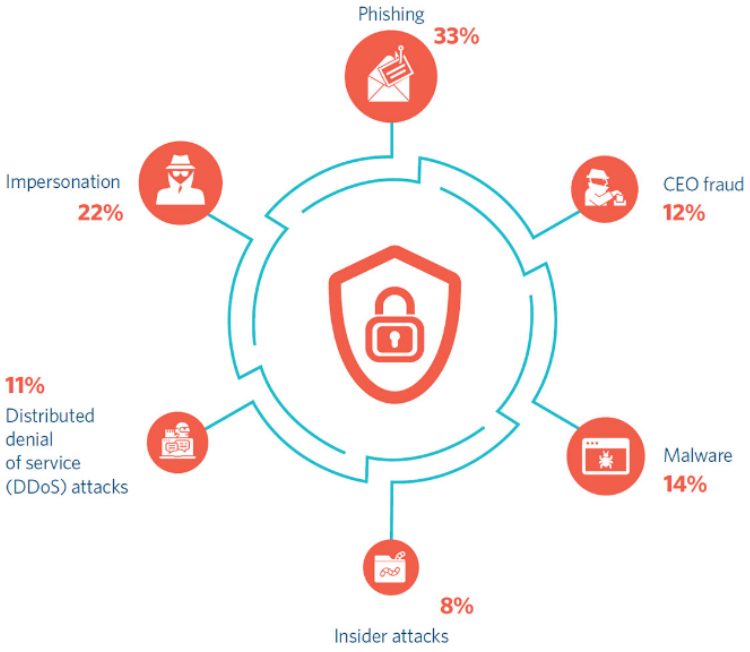


Figure 1.2: Types of cybercsecurity threats .[6]

or deletion of sensitive data and information, compromise critical infrastructure or services, or cause physical destruction.

- Cyberattacks are increasingly common and sophisticated around the world.Despite improved security, the United States is still falling behind in mitigating the threat of cyber-attacks. Many experts believe that a major cyber-attack will cause widespread harm to the security of a country and its ability to defend itself and its population by 2025.
- Though a cyberattack or catastrophic disruption has not yet occurred in our world, the consequences of such an attack in Seattle could seriously harm the public and degrade or interrupt the town's core functions and services. [17]

1.5 Types of cyber attack

Cybercrime increases extremely every year, while attackers become more well-organized and sophisticated. Cyberattacks occur for various reasons and in different ways. However, one thing in common is that cyber-criminals will seek to exploit vulnerabilities in an organization's security policies, practices or technologies. Therefore, there are many types of attack. [19]

Type	Impact
Malware (ransomware, spyware, viruses, worms) Malicious software used by attackers to breach a network through a vulnerability, such as clicking a link, that automatically downloads the software to the computer.[389]	• Blocks legitimate access to components of the network • Installs additional harmful software • Obtains information by transmitting data from the hard drive • Disrupts components and makes the system inoperable
Phishing Fake communications (typically through email) appearing to be from a trustworthy source that allow hackers to obtain login information or install malware on a computer when someone interacts with their message.[390]	• Obtains a person's confidential information for financial gain • Obtains employee log-in credentials to attack a specific company • Installs malware onto a computer
Man-in-the-middle attack (MitM) Attackers insert themselves into a two-party transaction. Common points of entry include unsecure public Wi-Fi networks and computers affected with malware.[391]	• Interrupts a transaction to steal personal data
Denial-of-service attack (DoS) Attackers flood a site host or network with digital traffic until the target site/service cannot respond or crashes completely. A distributed denial of service attack (DDoS) is when multiple machines are used to attack a single target. Botnets, which are networks of devices that are infected with malware, are often used in DDoS attacks.[392]	• Legitimate users cannot access websites, online services, or devices • Slows down network performance
Structured Query Language (SQL) injection Attackers use malicious code on vulnerable servers to force the server to reveal	• Obtains contents of an entire database, including sensitive information • Allows attackers to modify and delete records in a database

Figure 1.3: Common Cyber Attacks and their Impacts[17]

1.6 Related works

The importance of fighting cyberattacks, cybercrimes and cybersecurity is highlighted in various studies.
Cybersecurity is about protecting physical and digital data, networks and technology systems from cyberattacks, unauthorized access, interruptions, changes. destruction and damage by various methods, applications and applied technologies .
Cyber attacks, such as denial of service attacks distributed through the transmission of malicious packets , phishing attacks on banking and shopping sites that mislead the user have increased dramatically.
In addition, attackers have used malware attack software (viruses, worms, Trojans, spyware and ransomware) which is installed in the user's computer without the user's permission more and more.
Again, the most common of these attacks and one of the most difficult ones to prevent are the applied sociology attacks. They are based on technical skills, ruse and persuasion, made while taking advantage of the victim's weakness.
Kevin Mitnick, one of the world-renowned pirates in social engineering attacks, penetrated most systems he attacked with this method.
This attack is referred to as one of the greatest security vulnerabilities in the system, regardless of the security of a technical system.
Similarly, attacks on IoT devices, which have grown rapidly in recent years, have significantly affected the company.

The more significant part is that it currently provides analysis to the press. Information on cybercrime incidents in India has been classified through machine learning techniques.[15]

1.7 Defense mechanisms

In order to defend against such attacks or to limit these attacks to minimal damage, a number of defense mechanisms specific to specific attack zones have been developed. This is where the defensive mechanisms are put in place.
As a general rule, depending on the attack zones and their defensive types, the intrusion detection system is subdivided into two categories.
Before discussing a particular defensive mechanism for multiple attack areas, a brief overview of the intrusion detection system is provided below. [33]

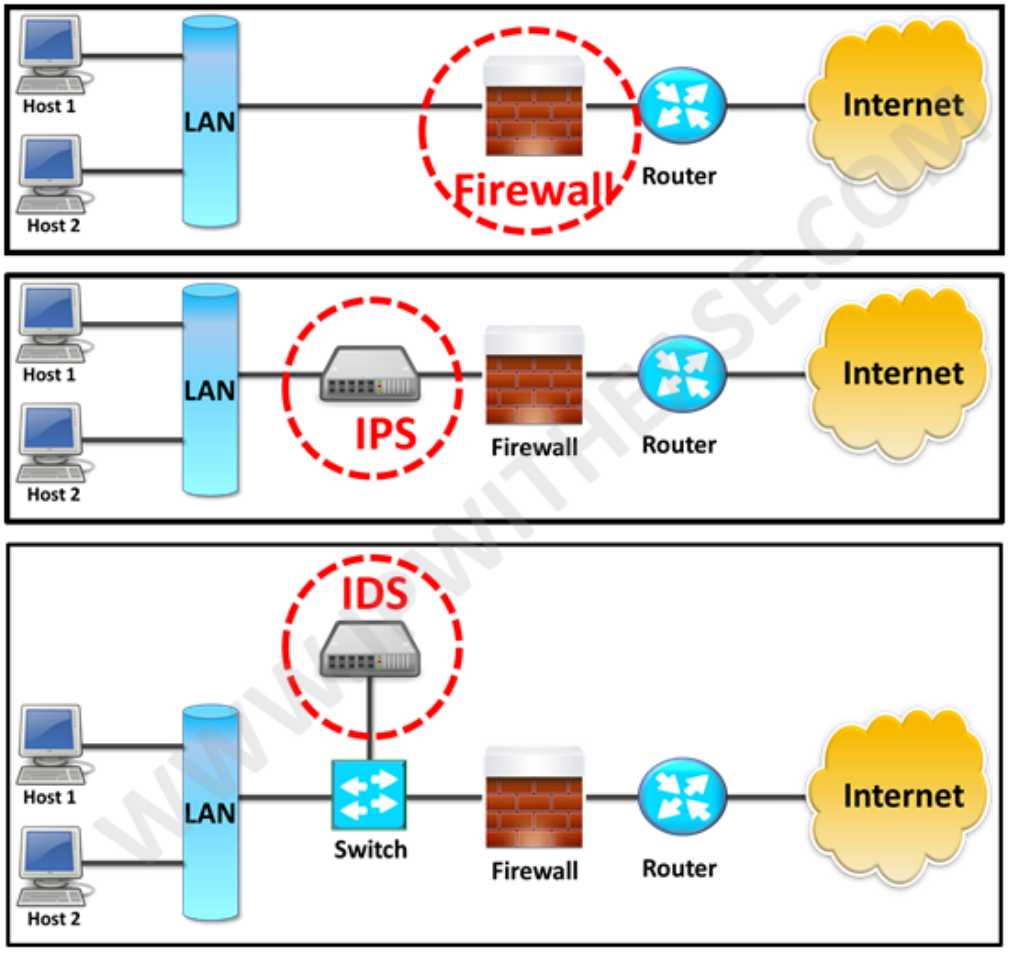


Figure 1.4: IDS vs IPS vs Firewall [12]

1.7.1 Definition Intrusion Detection System

Intruder detection device. As a broader class of technology, IDS includes intrusion detection as well as a mechanism to prevent intrusion.

The IDS is typically constructed with a combination of software and hardware to observe and control network activities within a network.

Depending on the target and detection mechanism, the IDS may be classified into two different groups. There are two types of IDS classification methods: the detection-based approach and the source-based approach.

Methods based on detection are also under-classified in misuse and anomaly-based methods, while methods based on data sources are under-classified in both network and host-based methods.

1.7.2 Definition Firewall

Firewalls can be software-based or hardware-based and are used to ensure the security of a network.

Its primary objective is to control inbound and outbound network traffic by analysing data packets and decide whether it should be authorised or not, on the basis of a predetermined set of druls.

The firewall of a network constructs a bridge between an external network which is supposed to be secure and

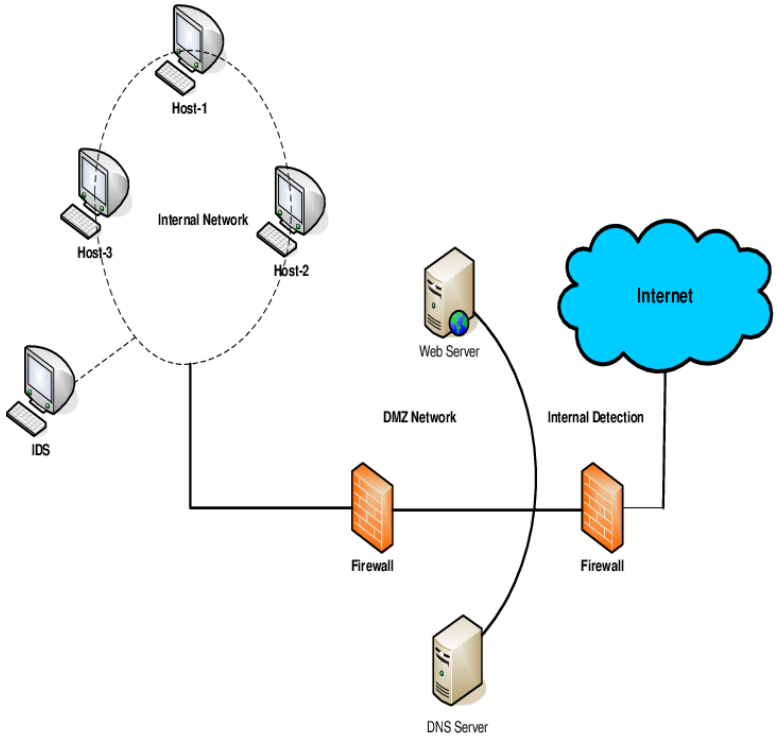


Figure 1.5: Intrusion Detection system (IDS)[10]

reliable, and is not a network, usually an external (inter) network, like the Internet, which is not expected to be secure and trustworthy.Many personal computers include firewalls based on software that protect against threats coming from public Internet.

Many data passes between networks contain firewall components and, conversely, many firewalls can perform core routing functions.[26]

1.8 conclusion

In this chapter, we have tried to cover ways to describe the field of cyber security, first, we have explained its importance, and the fundamentals of information security are: Confidentiality, Integrity, Availability. We have also given examples of causes and damage in chapter 2, we will introduce the technics of Data Mining. [20]

Chapter 2

Data Mining

2.1 Introduction

Increase in digital data acquisition and storage technology are driven us to the growth of huge databases. This happened in all domains of human activity (such as transaction data supermarket, credit card usage records, and government statistics)

As interest has grown in the possibility of exploiting this data, extracting from them knowledge that could be of value to the owner of the database. The discipline that deals with this task is known as data mining.

In this chapter, we will discuss this discipline by showing its power to study a large database in order to extract knowledge, as well as its different techniques [22]

2.2 What is Data Mining

Data Mining is the computational process of discovering patterns in large datasets involving methods at the intersection of learning automation, statistics and database systems. It's a process essential where intelligent methods are applied to extract patterns from data.

Data Mining is an essential component of Big Data technologies and big data analysis techniques.

This is the source of big data analysis, predictive analysis and data exploration. [9]

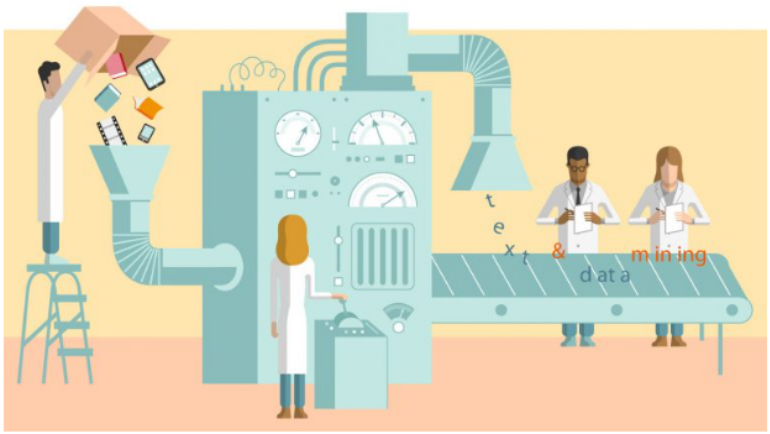


Figure 2.1: Data Mining Process [4]

2.2.1 Data Mining Objectives

The principal objective of data mining is the automatic analysis of large amounts of data. This serves to extract exciting patterns previously unknown. We talk about the groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rules mining).

This generally involves the use of database techniques such as spatial indexes. Thus, these patterns can be seen as a kind of summary of the input data. In addition to being able to be used in additional analysis or, for example, in machine learning and predictive analysis.

One of the examples we can give is data mining. This could identify several groups in the data, which can then be used to obtain more accurate results being able to predict problems through a decision support system. [28]

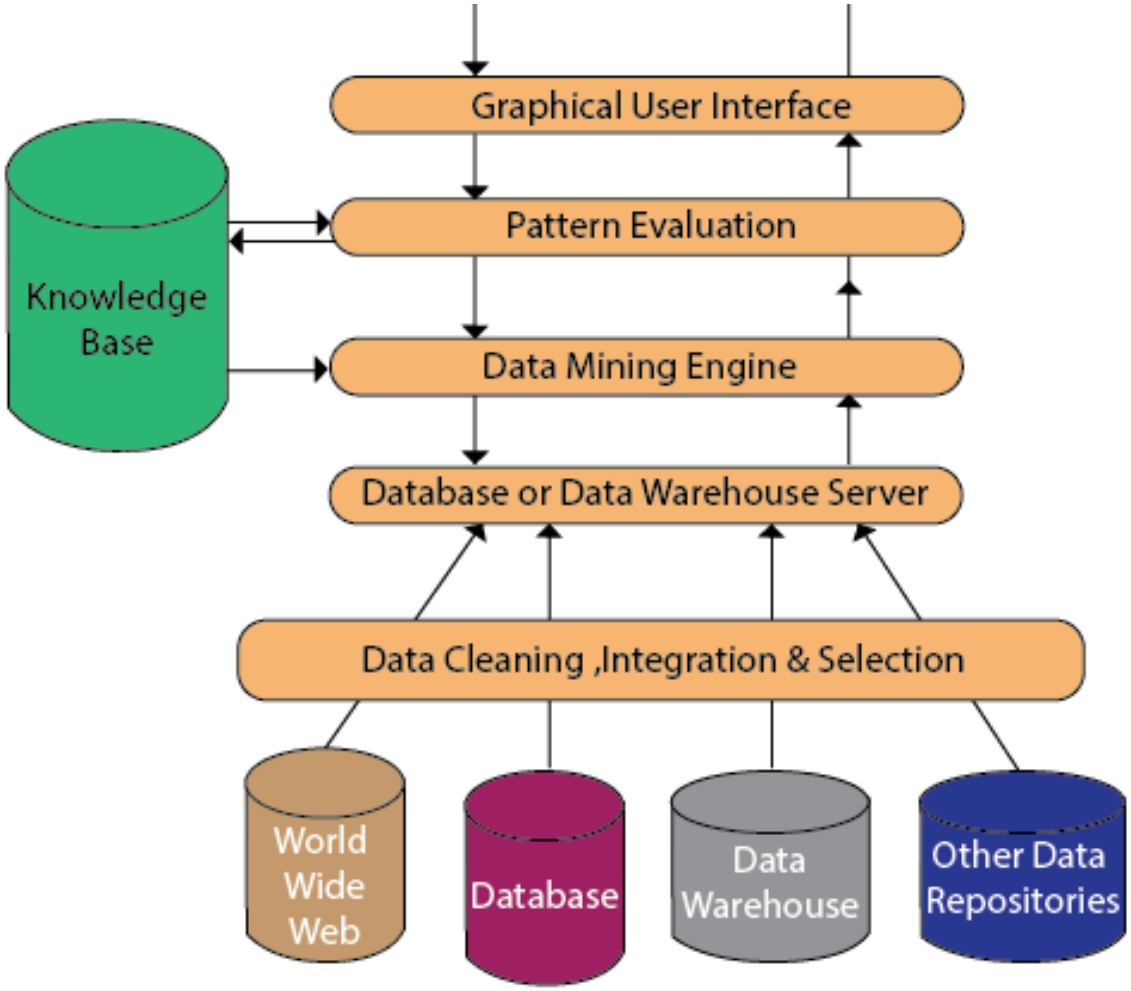


Figure 2.2: Data Mining Architecture [9]

The first operation is data understanding : Determine what information will be needed to achieve the defined objective, list and collect resources that contain relevant data.

Then, Data preparation : Repair the data in the appropriate format to meet the requirements, ensure their quality and correct duplication or lack problems.[9]

The Important Components of Data Mining Systems are look at figure:

A. Data source.

B. Motor for data mining.

C. Data warehouse server.

D. The template evaluation module.

E. Graphical user interface and knowledgebase.

2.2.2 What future for Data Mining?

The future is bright for this field and data science as the volume of data continues to increase. And just as mining techniques have evolved and improved through technological improvements, technologies that extract valuable information from data improve the quality of mining operations. Today, artificial intelligence, machine learning and deep learning are becoming more widely available.[31]

2.2.3 Advantages of Data Mining

Data mining can bring significant benefits to businesses by uncovering models and relationships in the data that the company already collects and combining that data with external sources. Here are some of the potential benefits of exploring data for a company.

Optimal product/service pricing : Using data mining to analyze the interaction of price variables like distribution and brand perception can help a company set prices that maximize profits.

Better marketing : Data mining can help a company gain greater value from its marketing campaigns by segmenting customers with different behaviors, Optimize engagement by segment or provide information to facilitate the development of customized creative advertisements. Results from advertising campaigns can often be demonstrated through sales dashboards.

Improved customer retention : Understanding client behaviour can enhance client relationships and reduce turnover.

Increased cost efficiency : Manufacturing costs, for instance, could be reduced through extensive data mining analyses, from information on supplier price behavior to a better understanding of customers purchasing habits.[33]

2.3 Where We can apply Data Mining

Data mining is now primarily used by businesses whose retail, financial, communications and marketing organizations are highly consumer-focused.

Here is important areas where data mining is widely used: [11]

2.3.1 Future Healthcare :

There is big potential to improve healthcare systems through data mining. It uses data and analysis to determine best practices that improve care and lower costs. Scientists use data extraction methods such as multidimensional databases, machine learning, soft computing, data visualisation and statistics.

Mining is a way of predicting the volume of patients in each class. Processes are developed to ensure that patients are provided with appropriate care at the right place and at the right time.

2.3.2 Intrusion Detection :

Defensive measures to avoid intrusion include authentication of users, avoiding programming mistakes, and safeguarding of the information. Data exploration can help improve intrusion detection by focusing on detecting anomalies. It helps an analyst distinguish between an activity and a shared daily network activity. Data extraction also allows the retrieval of more relevant data for the problem.

2.3.3 Financial Banking :

With the computerized bank all over the place huge amount of data is supposed to be produced with new transactions. Data mining can contribute to solving business problems in the banking and financial sector by finding trends, causalities and correlations in business intelligence and market prices that are not immediately obvious to managers because data on volumes are too large or are generated too quickly to be reviewed by experts.

Managers can find this information to better segment, target, acquire, retain and maintain a cost-effective customer.

2.3.4 Criminal Investigation :

Criminology is a process that purposes at identifying the features of crime. In fact, crime analysis includes exploration and detection of crime and its relationship to criminals.

The large volume of crime datasets and the complexity of the relationships between these types of data have made criminology an appropriate domain for the application of data mining techniques. Text-based crime reports can be converted to text-processing files. This information can be used to carry out the crime data matching process.

2.3.5 Cyber security

We can apply data mining to any database and adjust it to any goal you want to achieve. In cybersecurity, mining algorithms often help to discover unusual data records and events that may indicate a security incident.

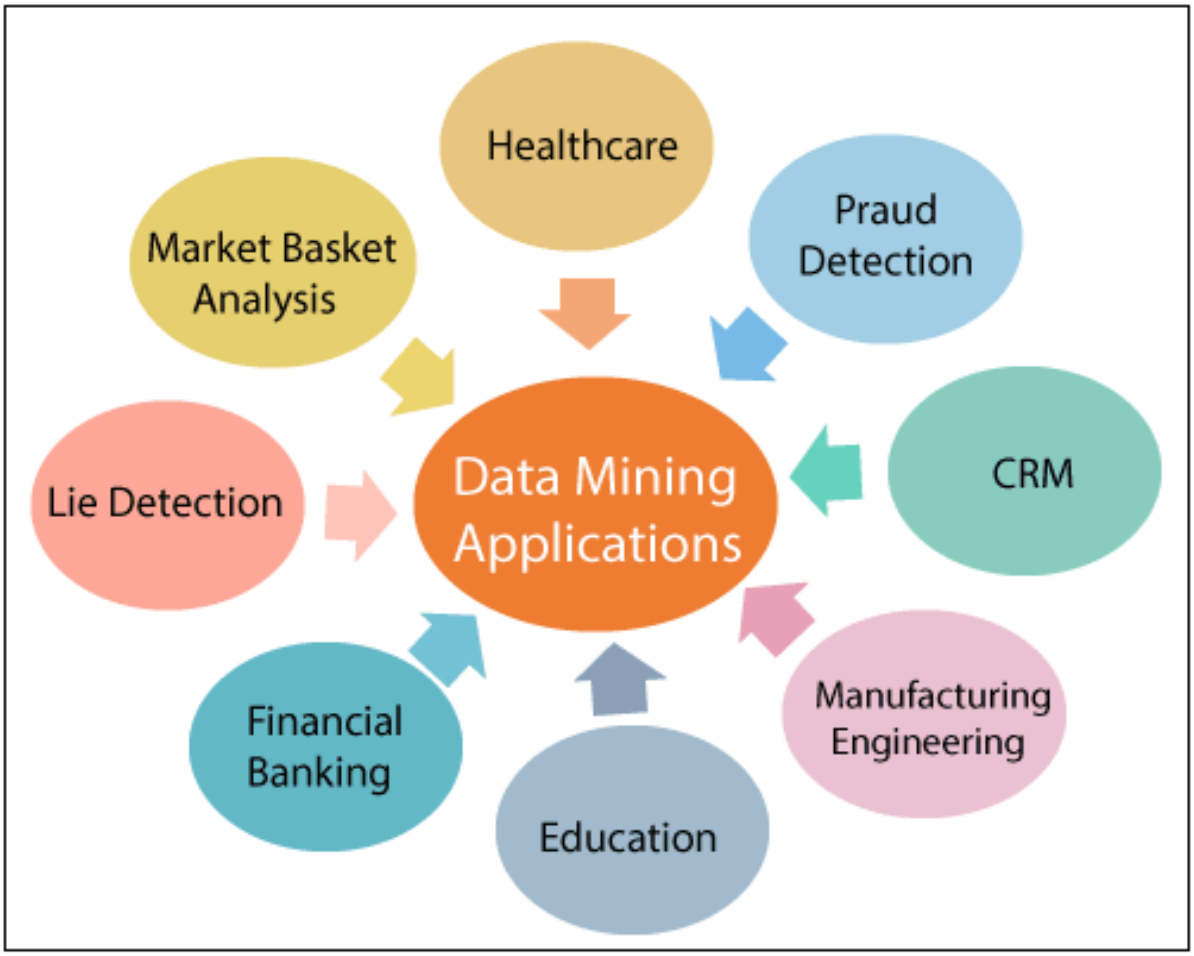


Figure 2.3: Data Mining application [5]

2.4 The process of Data Mining

Knowledge discovery(as shown in figure 2.3) is an iterative sequence of the following steps:

1.Data cleaning: to remove noise and inconsistent data.

2.Data integration: where multiple data sources can be combined.

3.Data selection: when the data concerning the analysis task is extracted from the database.

4. Data transformation: where data is transformed or aggregated into forms suitable for mining by performing synthesis or aggregation operations.

5. Extraction of information (Data mining): an essential process where Intelligent methods are applied to extract the data patterns.

6. Model evaluation: to identify really interesting models representing knowledge based on measures of interest.

7. Presentation of knowledge: where techniques of visualization and knowledge representation are used to present knowledge extracted to the user. [24] [23]

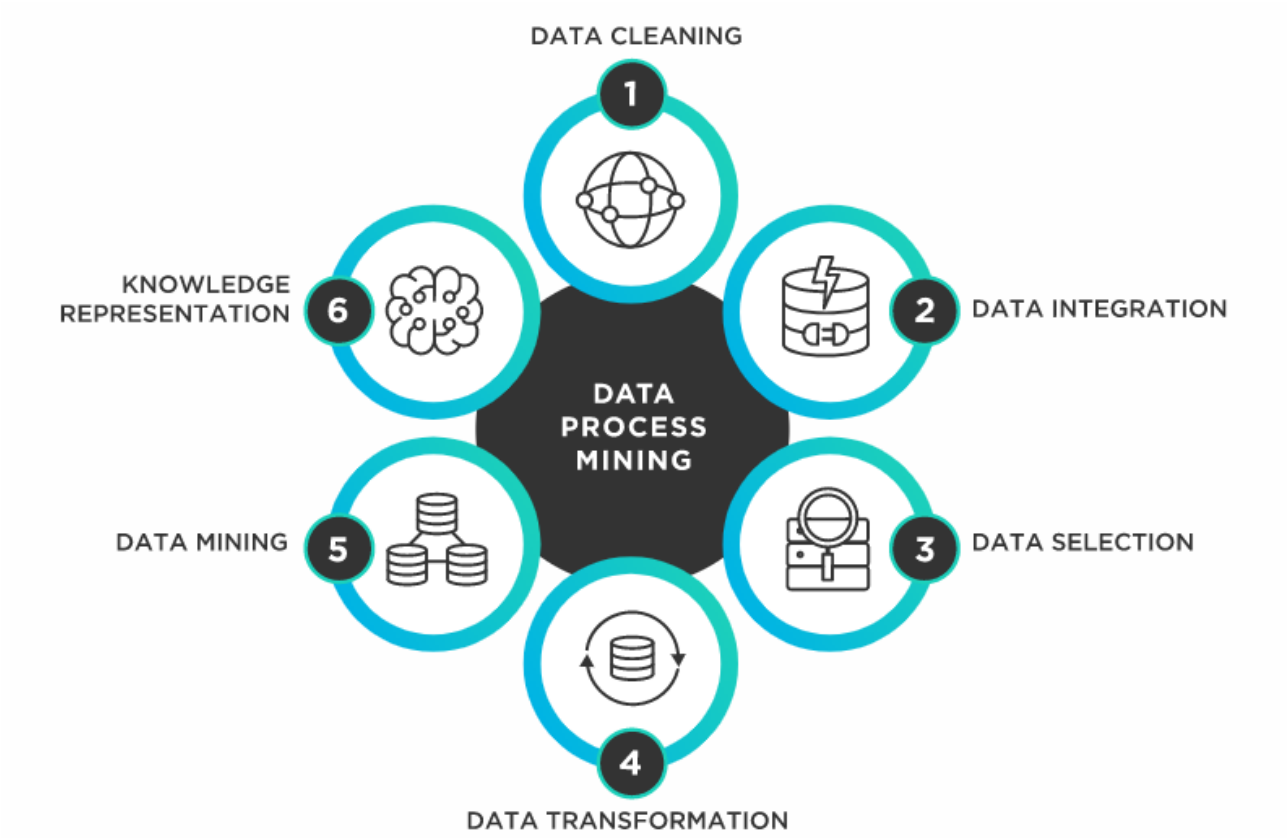


Figure 2.4: Data mining [3]

2.5 The Technics of Data Mining

According to various technologies and methods from the intersection of database management, machine learning, and statistics, experts in data mining have focused their works to better understanding how to process and make summarization from the huge amount of data, and give us these important techniques (figure 2.4):

1. Tracking patterns: One of the most basic techniques in data mining is learning to identify patterns or models in data sets. This is generally a recognition of some abnormality in data happening at regular intervals, or a flow of a certain variable over time. For example, you might see that your sales of a certain product seem to spike just before the holidays.

2. Classification: Classification used to collect various attributes together into distinct categories, which we can then use to draw further conclusions, or serve some function. For example, if we are evaluating data on individual customers and purchase histories, we might be able to classify them as "low," "medium," or "high" credit risks. we could then use these classifications to learn even more about those customers.

3. Association: It is related to following patterns, but is more specific to dependently linked variables. In this case, we look for specific events or attributes that are highly correlated with another event or attribute; for example, we might notice that when customers buy a specific item, they also often buy a second, related item.

4. Outlier detection: In many cases, we also need to be able to identify anomalies, or outliers in your data. For example, if your purchasers are almost exclusively male, but during one strange week in July, there's a huge

purchasers in female, we will want to investigate and see what drove it.

5. Clustering: Clustering is very similar to classification, but involves grouping amounts of data together based on their similarities: For example, you might choose to cluster different demographics of your audience into different packets based on how much income they have, or how often they tend to shop at your store. [29]

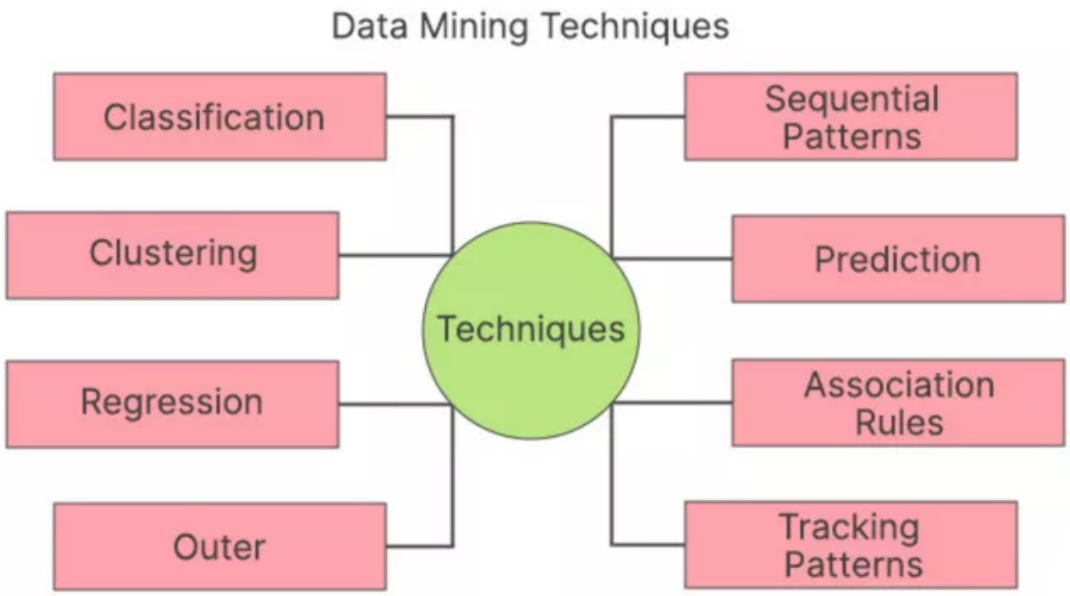


Figure 2.5: Data Mining Process

[1]

Data mining task	Example techniques
summarization	statistical analysis, histograms, plots
clustering	K-means, Kohonen, k-NN
association rules	A priori, FP Growth
regression	linear regression, neural networks,
classification	decision trees, neural networks, support vector machine (SVM), Bayes classifiers
time series analysis	ARMA, ARIMA, models ARCH models, HMM models, trend estimation, decomposition, spectral analysis

Figure 2.6: Example techniques used in DM tasks

[14]

6.Prediction: Prediction is one of the most valuable data mining techniques, since it's used to project the types of data you'll see in the future. In many cases, just recognizing and understanding historical trends is

enough to chart a somewhat accurate prediction of what will happen in the future. For example, you might review consumers' credit histories and past purchases to predict whether they'll be a credit risk in the future.[33] n below we present two method of prediction :

6.1 Neural networks : An artificial neural network is a computational model whose design is very schematically inspired by the functioning of biological neurons.

Neural networks are generally optimized by probabilistic-type learning methods. They are placed on the one hand in the family of statistical applications, and on the other hand in the family of methods of artificial intelligence [6].

Neural networks are composed of single elements (or neurons) operating in parallel. These elements were strongly inspired by the biological nervous system. As in nature, the functioning of the network is strongly influenced by

the connection of the elements between them. We can train a neural network for a

specific task (character recognition for example) by adjusting the values ..of the

connections (or weight) between the elements (neuron).[18]

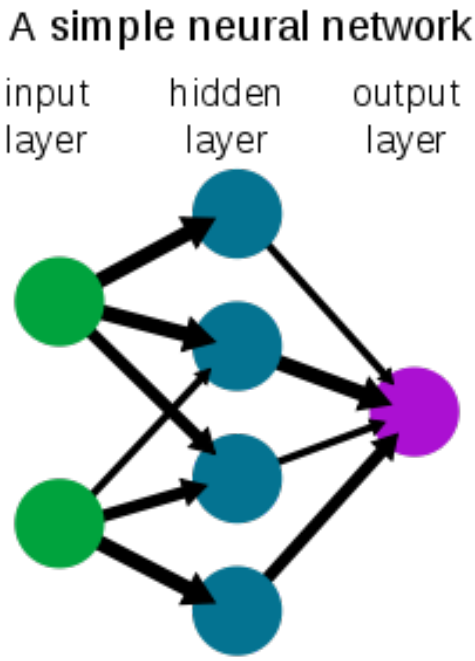


Figure 2.7: Simple Neural network [2]

2.6 Regression

Regression, used basically as a form of planning and modeling, is used to identify the probability of a certain variable, given the presence of other variables. For example, you could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression's main focus is to help you uncover the exact relationship between two (or more) variables in a given data set.

Regression analysis consists in essence of four different stages:

1. Identification of dependent and independent variables.
2. Identification of the shape of the relationship between variables like linear, parabolic, exponential, etc. by means of the dispersion diagram between dependent and independent variables.
3. Computation of regression equation for analysis.
4. Error analysis to understand the extent to which the estimated model matches the real dataset.[30]

2.6.1 Types of Regression Analysis Techniques

Many types of regression analysis techniques exist, and their use depends on the number of factors. These factors include the type of target variable, the shape of the regression curve and the number of independent variables.[34]

2.6.2 1. Linear Regression

Linear regression is one of the most elementary forms of regression within machine learning. The linear regression model consists of a predictive variable and a dependent variable that are linearly bound together. If the data implicates more than one independent variable,Next, linear regression is referred to as multiple linear regression models. [34]

This equation is used to refer to the linear regression model: y=mx+c+e where m is the slope of the line, c is an intersection and e is the fault of the model.

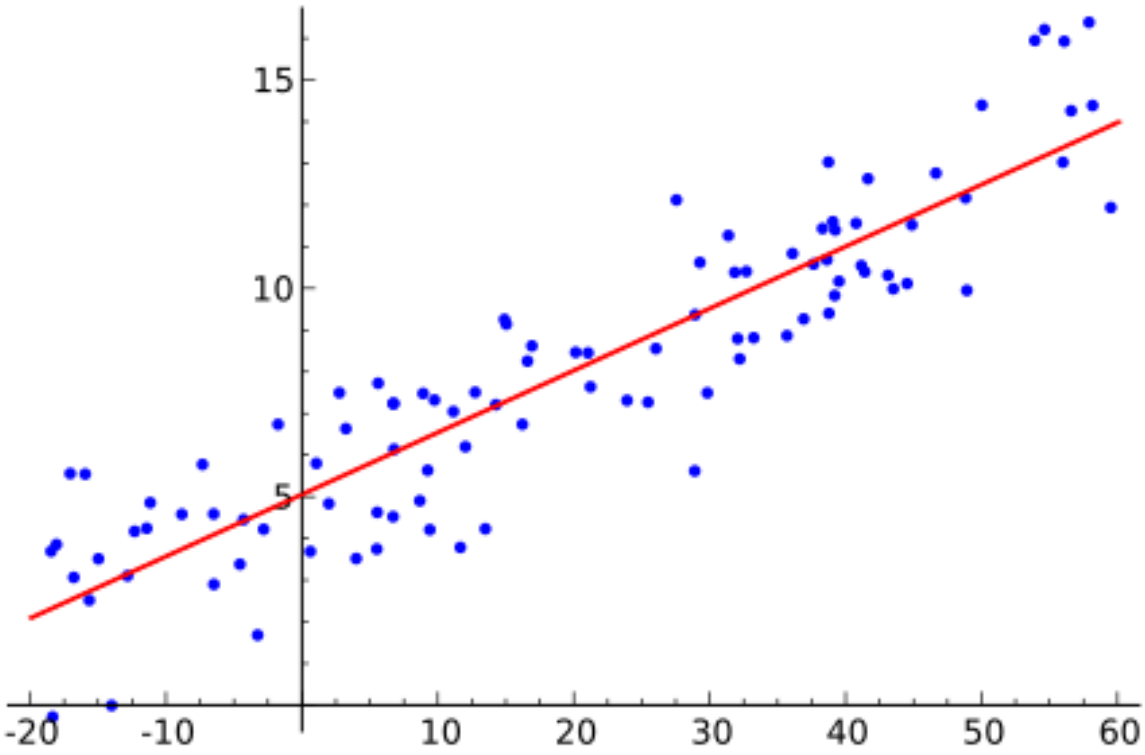


Figure 2.8: Linear Regression[34]

The best fit row is determined by varying the values of m and c. The prediction error is the difference between the observed values and the expected value. The m and c values are selected in order to give the minimum predictive error. It should be noted that a simple linear regression model may exhibit outliers. As such, it should not be used for large volumes of data.

2.6.3 Formulation of Linear Regression Technique

The linear regression model consists of a random variable Y (called the answer variable) as a linear function of another random variable X (called the predictive variable) that is represented by the equation:
Y= Alpha + Beta.X (eq 1)
Alpha and Beta are regression coefficients that specify Y interception and line slope respectively.
The regression coefficients Alpha and Beta are resolved using the least squares method, which minimise the error between actualdata values.
Shows the sample data ordata points of the form. (x1,y1),(x2,y2). . . . . .(xs, ys) that the regression coefficients Alpha and Beta are provided by. Beta =Somme(xi - x) (yi -y)/ Somme(xi -x) (eq 2) Alpha =y- Beta.x (eq 3)
These values of the regression coefficients Alpha and Beta of the computed inequality (2) and (3) are replaced by equation (1) to determine the relationship of the response variable X to the target variable Y. [16]

2.6.4 Algorithm of Linear Regression Technique

The linear regression technique uses the next algorithm. Step 1: Use the values for the variables Xi and Yi.
Step 2: Calculate the mean for Xi such as the mean is x= (X1 +X2 +.....+ Xi)/ Xi
Step 3:Calculate the mean for Yi such as the mean is y= (Y1 +Y2 +.....+ Yi)/ Yi
Step 4: Calculate the value of the regression coefficient Beta by replacing the values of Xi , Xi average Yi and Yi average in Equation 2.
Step 5: Calculate the value of another coefficients Alpha by replacing the values of Beta (calculated in Step 4), the mean of Xi and the mean of Yi in Equation 3.
Step 6: Lastly, replace the value of the Alpha and B regression coefficients in the equation Y= Alpha + BX [16]

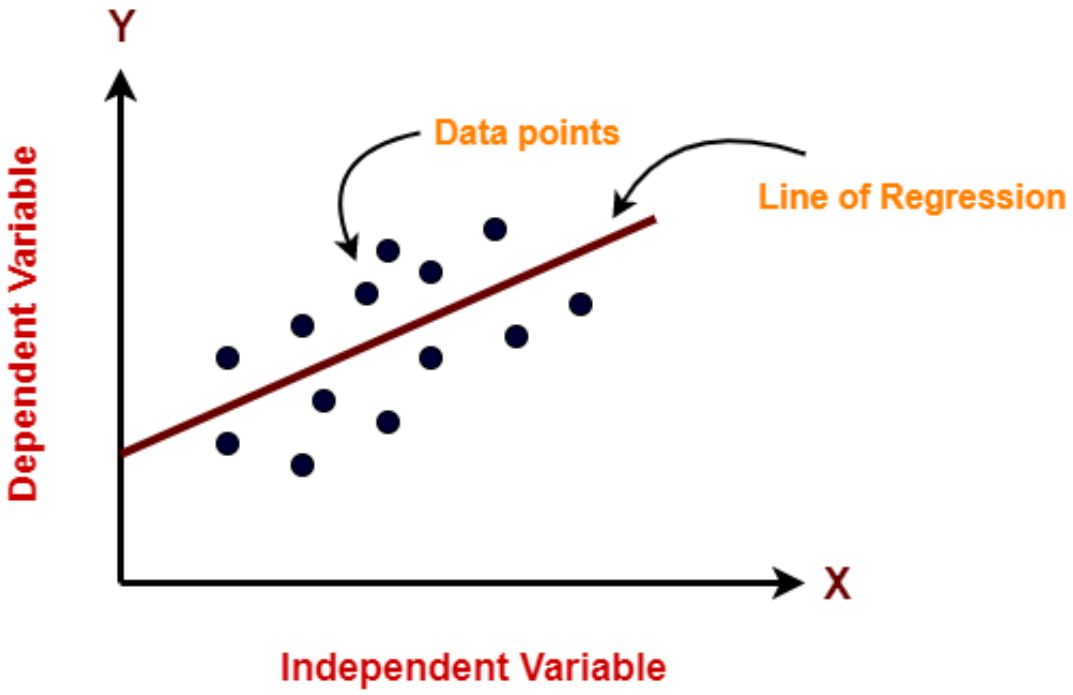


Figure 2.9: Linear Regression is an ML algorithm[7]

2.6.5 2. Logistic Regression

Logistic regression is one of the types of regression analysis technique that is used when the dependent variable is discreet. Example: 0 or 1, true or false, etc. The target variable can therefore only have 2 values, and a sigmoid curve indicates the relationship between the target variable and the independant variable.[34]

The Logit function is used in logistic regression to measure the relationship of the target variable to independent variables. The equation for logistic regression is as follows.

$\text{logit}(p) = \ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk$

where p is the probable appearance of the element.

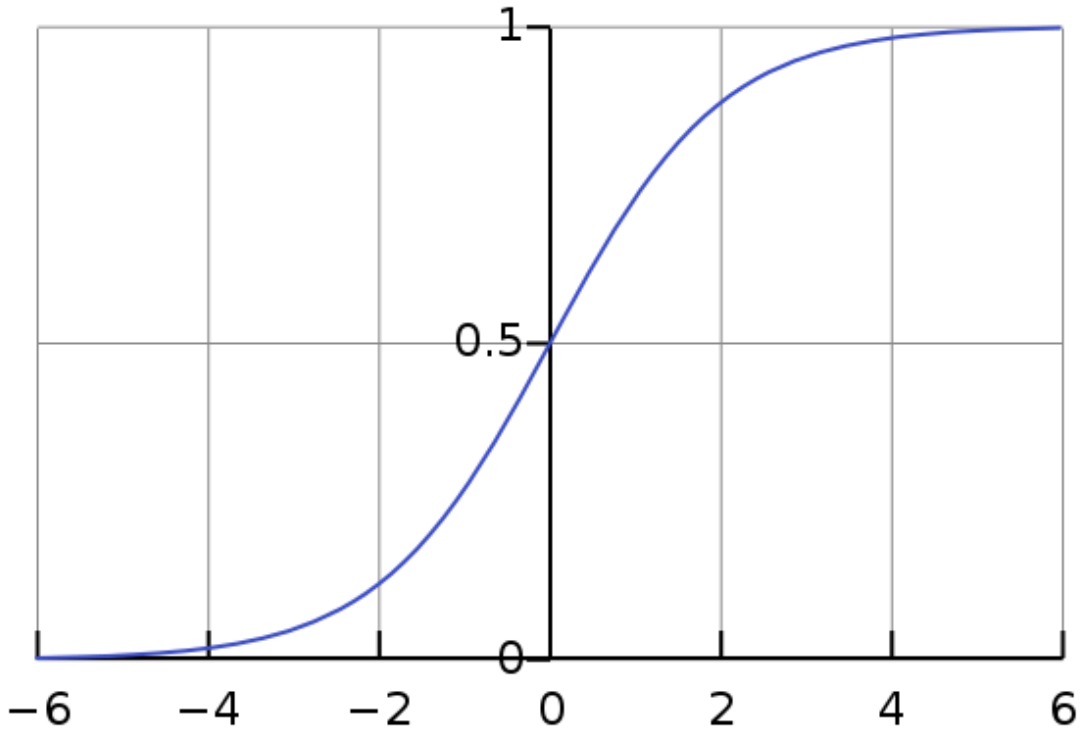


Figure 2.10: Logistic Regression[34]

To select logistic regression, as a regression analytical technique, Note that the data is large and that the future values in the target variables are almost identical. Also, there shouldn't be any multillinearity, which means there should not be any correlation between independent variables in the data set.

2.6.6 What is the purpose of logistic regression?

Logistic regression employs mathematics to measure the impact of multiple variables (e.g., age, gender, ad placement) with a given outcome. Logistic regression may also estimate probabilities of events, including the determination of a relationship between characteristics and probabilities of outcomes.[25]

2.6.7 Advantages and disadvantages of logistic regression

The main benefit of logistic regression is that it is much more easily installed and trained than other machine learning and AI applications.

Another advantage is that it is one of the most effective algorithms when the different results or distinctions represented by the data can be linearly separated. This means that you can draw a straight line between the results and a logistic regression computation.

One of the greatest advantages of logistic regression for statisticians is that it can help to highlight the inter-relationships between the different variables and their impact on the results. [25]

2.7 conclusion

Data mining is a very powerful technology for generating knowledge from a gigantic database for decision making. Future developments are expected to make data mining even more powerful and useful. That's why we find it in several domain like :marketing, banking, crime investigation and cyber security. In next chapter we will present the design of our system to solve the problem of cyber attack.

Chapter 3

Conception

3.1 Introduction

After seeing in detail, the essential notions and mechanisms of Data Mining, Cyber-attack, Machine learning, Intelligence artificial, we will present in this section 3 "Design" the development process of our system.
We will present the overall architecture of our system based on an internal view (structures and behaviors of the components).
We will also detail the features of this architecture before presenting its realization.
This section clarifies how we performed a regression algorithm through learning and detection. In addition, we seek to predict whether an event is considered an attack or not .

3.2 System presentation

we will describe our system globally and give the shape of its structure, such as its components and purpose.

3.2.1 System objectives

The objective of this thesis is to offer security agencies a system automatic attack detection, it will predict attack based on Data Mining techniques especially we talk about regression.

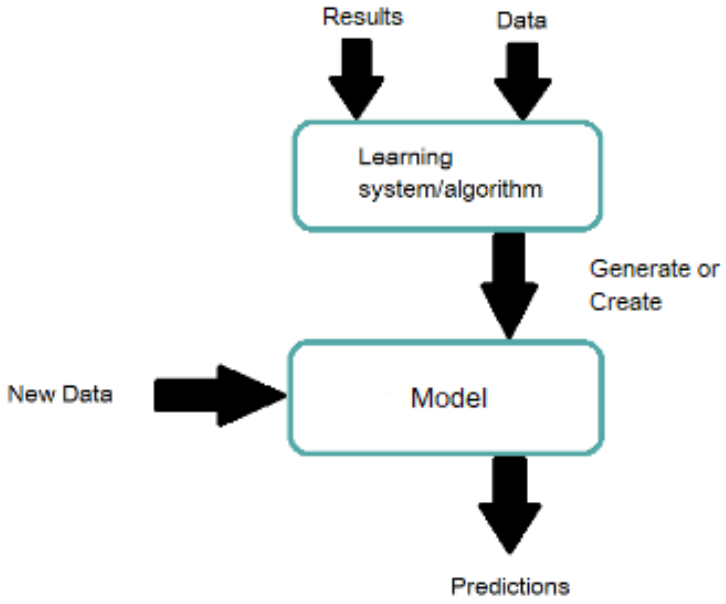


Figure 3.1: predictions system architecture

3.2.2 Global System Architecture

In general, we can represent the structure of decision model evaluation system as follows:

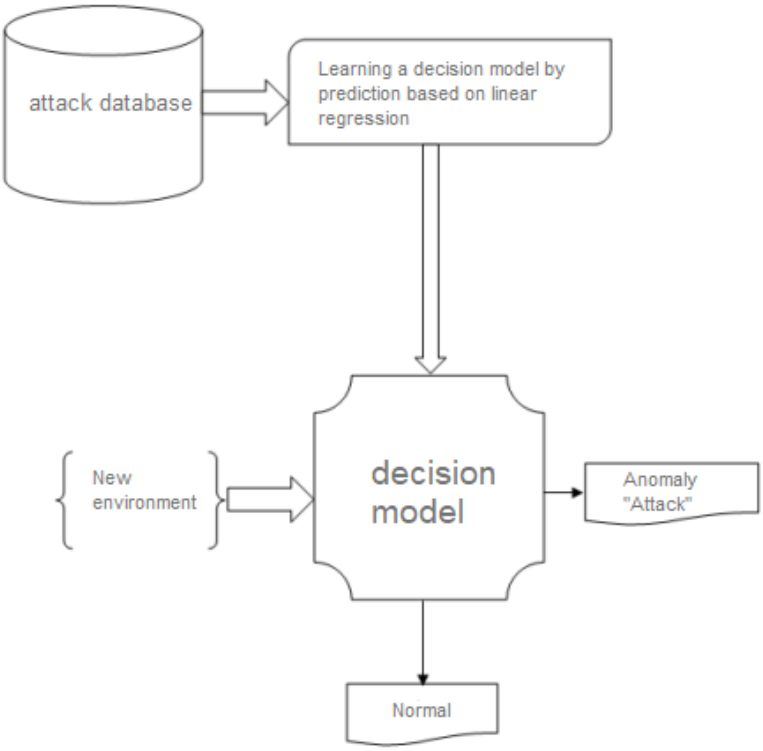


Figure 3.2: The General Architecture of the System

3.3 Detailed System Design

In cases of attack detection, based on the learning of individuals' behaviours, a data extraction phase is essential. It makes it possible to define, for each person, indices characterizing his behavior.
The following diagram will explain the detailed design of this module.

The system can be divided into 3 components:

- **Learning base :** It is a database used in the learning phase is that of features retrieved from a database of historical events.
- **Test basis :** This is a database used in the test stage is that of features extracted from a database of historical events.
- **Learning Parameters :** The model consists of a predictive variable and a dependent variable that are linearly linked to one another. In case the data requires more than one independent variable.

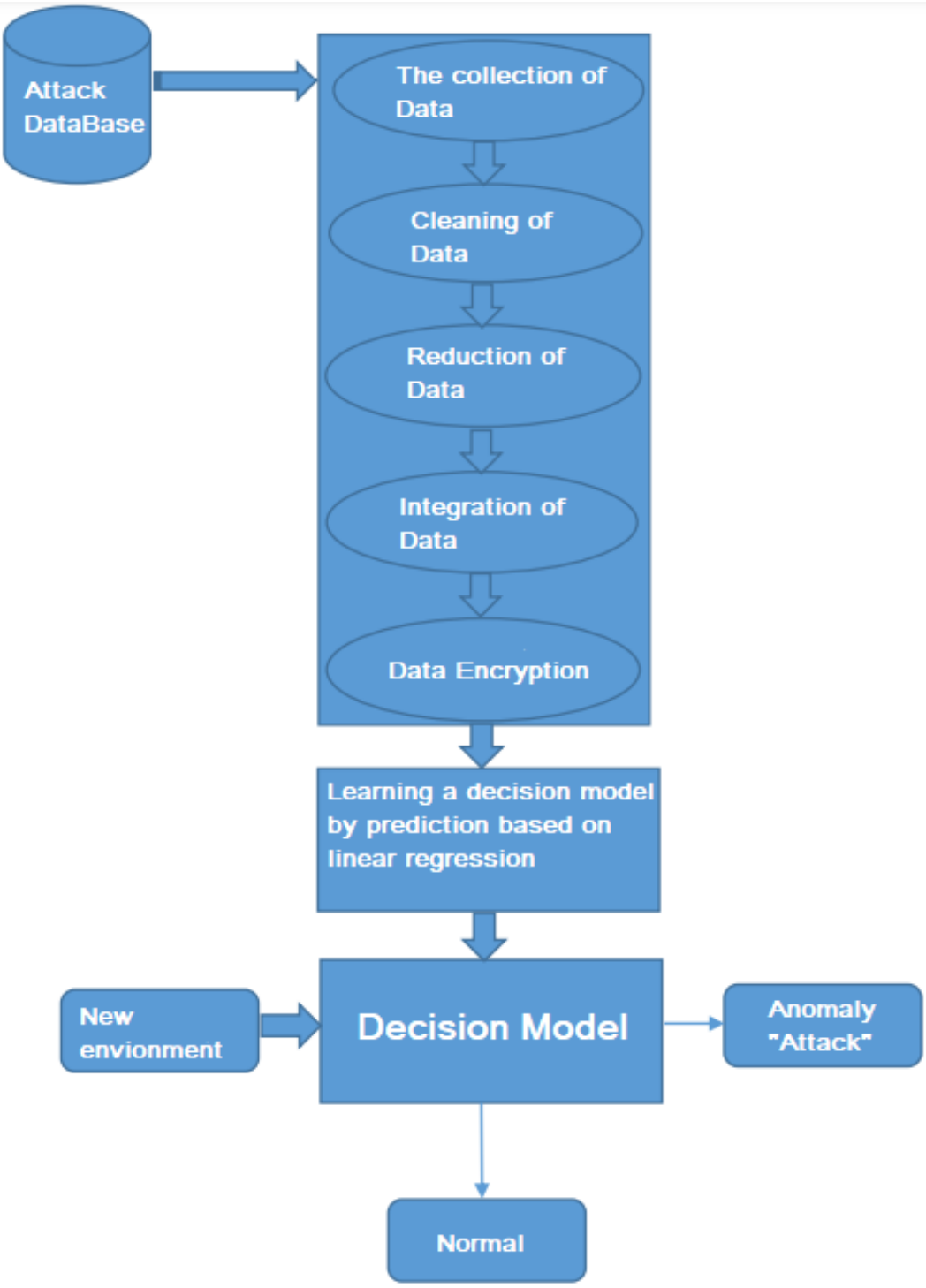


Figure 3.3: Detailed System Architecture

3.3.1 Data preprocessing

We performed the following pre-processing steps on a dataset:

1. The collection data.
2. Cleaning Data.
3. Reduction Data.

4. Integration Data.

5. Data encryption (for example: texts to numbers).

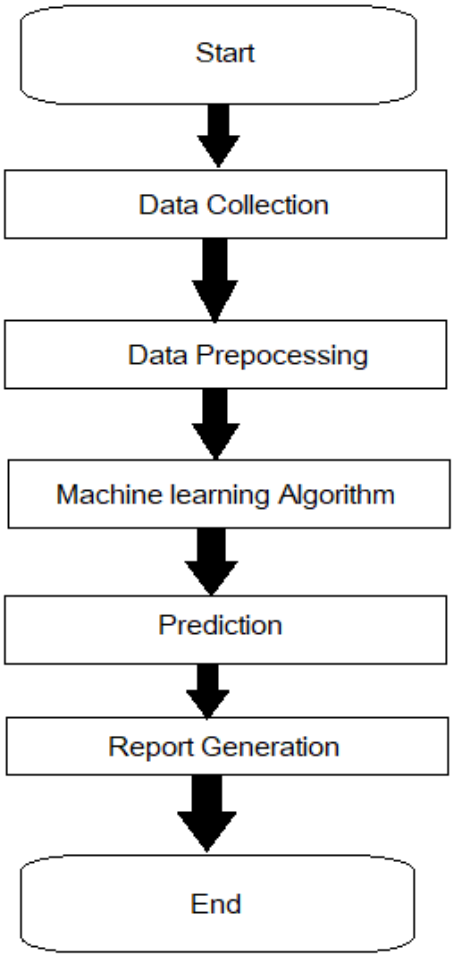


Figure 3.4: Processus de Classification

3.3.2 Data Collection

The process of collecting this dataset was difficult because most of the datasets were not free and did not have public access, but we succeeded to find one on the KAGGLE platform (kaggle.com), which did not contain much data, just a satisfactory amount.

this link from KAGGLE for dataset :

Dataset network-intrusion-detection : https://www.kaggle.com/datasets/sampadab17/network-intrusion-detection

Dataset name/label	Dataset size	Total (rows . columns)	Rows	Columns	Source File Name
Database Test-data	2420 KB	924304	22544	41	Test-data.csv
Database Train-data	2880 KB	1058064	25192	42	Train-data.csv
all dataset	5300 KB	1982368	47736	83	TotalData.csv

Table 3.1: Metadata of the Collected Data Set

3.3.3 Design by linear regression Algorithm

In linear regression, the model aims to obtain the most appropriate regression line to forecast the value of y as a function of the given input value (x). During model drive, the model computes the cost function which measures the average square error between the predicted value (pred) and the actual value (y). The model aims to minimize the cost function.

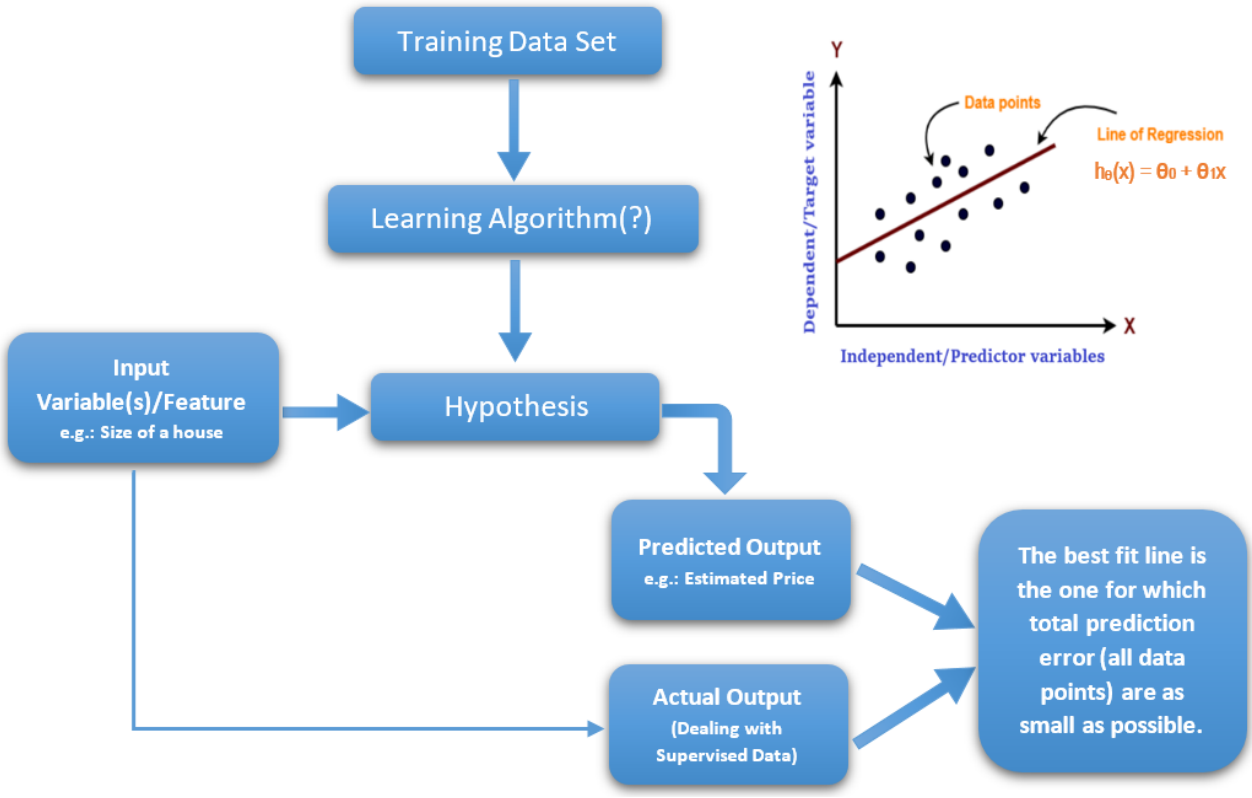


Figure 3.5: Machine Learning Algorithm Detailed View

Concepts and Formulas:

Linear regression uses the simple formula that we all learned in school:

$$\mathbf{Y = C + AX}$$

Just as a reminder, Y is the output or dependent variable, X is the input or the independent variable, A is the slope, and C is the intercept.

For the linear regression, we follow these notations for the same formula:

$$h = \theta_0 + \theta_1 X$$

If we have multiple independent variables, the formula for linear regression will look like:

$$h = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \cdots$$

Here, the assumption is referred to as "h". It is the expected output variable. Theta0 is the bias term, and any other theta values are coefficients. They are randomly initiated early on and then optimized with the algorithm so that this formula can predict the dependent variable closely.

Linear regression cost function:

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$minimize\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Cost function:

$$J\left(\Theta_0, \Theta_1\right) = \frac{1}{2m}\sum_{i=1}^{m}[h_\Theta(x_i) - y_i]^2$$

Predicted Value

True Value

Gradient descent:

$$\Theta_j = \Theta_j - \alpha \frac{\partial}{\partial \Theta_j} J(\Theta_0, \Theta_1)$$

Learning Rate

Now,

$$\frac{\partial}{\partial \Theta} J_\Theta = \frac{\partial}{\partial \Theta} \frac{1}{2m} \sum_{i=1}^{m} [h_\Theta(x_i) - y]^2$$

$$= \frac{1}{m} \sum_{i=1}^{m} (h_\Theta(x_i) - y) \frac{\partial}{\partial \Theta_j} (\Theta x_i - y)$$

$$= \frac{1}{m} (h_\Theta(x_i) - y) x_i$$

Therefore,

$$\Theta_j := \Theta_j - \frac{\alpha}{m} \sum_{i=1}^{m} [(h_\Theta(x_i) - y) x_i]$$

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

Now,

$$\frac{\partial}{\partial \theta} J_\theta = \frac{\partial}{\partial \theta} \frac{1}{2m} \sum_{i=1}^{m} \left[h_\theta(x_i) - y_i \right]^2$$

$$\frac{\partial}{\partial \theta} J_\theta = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x_i) - y_i \right) \cdot \frac{\partial}{\partial \theta_j} \left(\theta x_i - y_i \right)$$

$$\frac{\partial}{\partial \theta} J_\theta = \frac{1}{m} \sum_{i=1}^{m} \left[\left(h_\theta(x_i) - y \right) x_i \right]$$

Therefore,

$$\theta_j := \theta_j - \frac{\alpha}{m} \sum_{i=1}^{m} \left[\left(h_\theta(x_i) - y_i \right) x_i \right]$$

- j : weight of the hypothesis.
- h(x i): value y provided for the first entry.
- theta : for each feature of X, add one more column for theta 0.
- Alpha : Rate of learning gradient descent.
- y : is the true value of the dependent variable (y) for any given value of the independent variable (x).
- x : is the predicted variable.
- m : length number of variables.

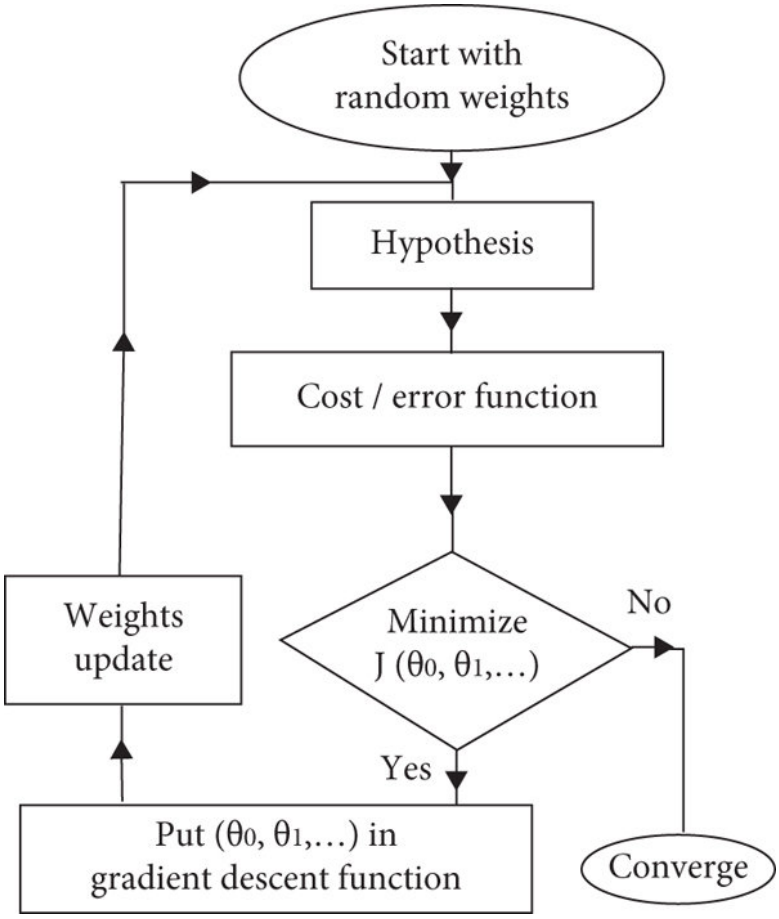


Figure 3.6: Flow chart of linear regression

3.4 Conclusion

In this chapter, we have presented how we designed our system, where we presented the overall design, and also detailed the steps we took to access our system, and we detailed the components used and the basic on all linear regression algorthim steps. We present, in the next chapter, the implementation of the system and the results obtained.

Chapter 4

Implementation

4.1 Introduction

In this chapter, we will introduce the work environment, the programming language, and the tools we used to build the system. Present the database used and some screenshot of the application interface. Then,we will explain to us all the experiments we have applied to the proposed method and the results obtained.

4.2 Development Environment

Our system is developed under the environment:

. G50-70 Laptop (Lenovo) - Type 80DY

. **Processor:** Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz, 2.4 GHz.
. **Hard Disk :** SSD 180 GB , HDD 1TB.
. **Memory:** SDRAM DDR3 8 GB.
. **Screen:** digital display lcd monitor 19.1" (48.5 cm)- 60p Hz.
. **Graphic card:** AMD Radeon-TM HD 8500M .

. Windows 10 Enterprise LTSC 64-bit.

4.2.1 development tools

Choosing the right programming environment is crucial for the development of the project. This is done in function of several factors the build power, user-friendliness, availability of multiple functions, communication with other environments, etc. In order to realize our system and its interfaces, we used the tools: Spyder IDE 5.2.2 / 21 January 2022, google Colab uses Python 3.6, we also choose the language of programming Python

3.9.5/ 3 may 2021.

4.2.2 Python

Figure 4.1: Python Logo

Python is the open source programming language most used by computer scientists. This language has propelled itself to the top of infrastructure management, data analysis or in the field of software development. Indeed, among its qualities, Python allows developers to focus on what they do rather than how they do it. It freed developers from the constraints of forms that occupied their time with older languages. Thus, developing code with Python is faster than with other languages.

4.2.3 Environment used for creating the model

Figure 4.2: google Colab Logo

Google Colab

was developed by Google to provide free access to GPU's and TPU's to anyone who needs them to build a machine learning or deep learning model. Google Colab can be defined as an improved version of Jupyter Notebook.

Another attractive feature that Google offers to the developers is the use of GPU. Colab supports GPU and it is totally free. The reasons for making it free for public could be to make its software a standard in the academics for teaching machine learning and data science. It may also have a long term perspective of building a customer base for Google Cloud APIs, which are sold per-use basis.

What Colab Offers You?

- Write and execute code in Python
- Create/Upload/Share notebooks

- Import/Save notebooks from/to Google Drive
- Import/Publish notebooks from GitHub
- Import external datasets
- Integrate PyTorch, TensorFlow, Keras, OpenCV
- Free Cloud service with free GPU

Google drive

Google Drive is a file storage and syncing service from Google. Introduced on April 24, 2012, Google Drive allows users to store files in the cloud (on Google servers) synchronize files across devices, and share files.
Along with a web-based interface, Google Drive offers applications with offline features for Windows and macOS computers, as well as Android and iOS smartphones and tablets.
Google Drive includes Google Docs, Google Sheets and Google Slides, which are part of the Google Docs Editors desktop suite that enables collaborative editing of papers, spreadsheets, presentations, drawings, forms, and more. Files created and subsequently edited by Google Docs are stored on Google Drive.

Figure 4.3: google drive Logo

4.2.4 Spyder

Spyder is a powerful scientific environment written in Python, for Python, and developed by and for scientists, engineers and data analysts. It has a unique combination of advanced editing, analysis, debugging and profiling capabilities of a comprehensive development tool with data mining, interactive execution, thorough inspection, and beautiful capabilities to visualize a scientific package.

Figure 4.4: spyder Icon

Thus, it is within the reach of most people in so far as it requires no specific knowledge and work, besides, on the most common operating devices.

4.3 Data Structures

4.3.1 part of dataset used

Network Intrusion Detection Dataset :

The dataset to be checked has been supplied and includes a wide variety of simulated intrusions. For this data set, the TCP/IP dump data was acquired by simulating a typical US Air Force local area network. 41 quantitative and qualitative features are obtained from normal and attack data (3 qualitative and 38 quantitative features).
The class variable has two categories: Normal, Anomalous .

For Train.csv

Our Dataset features for our model :(**duration , src-bytes , dst-bytes , land , wrong-fragment, urgent ,hot, num-failed-logins , logged-in , num-compromised , service-telnet , service-tim-i , service-time , service-urh-i , service-urp-i , service-uucp , service-uucp-path , service-vmnet , service-whois , class-normal)**

AH	AI	AJ	AK	AL	AM	AN	AO	AP
dst_host_	dst_host_	dst_host_	dst_host_	dst_host_	dst_host_	dst_host_	dst_host_	class
0.17	0.03	0.17	0	0	0	0.05	0	normal
0	0.6	0.88	0	0	0	0	0	normal
0.1	0.05	0	0	1	1	0	0	anomaly
1	0	0.03	0.04	0.03	0.01	0	0.01	normal
1	0	0	0	0	0	0	0	normal
0.07	0.07	0	0	0	0	1	1	anomaly
0.04	0.05	0	0	1	1	0	0	anomaly
0.06	0.07	0	0	1	1	0	0	anomaly
0.09	0.05	0	0	1	1	0	0	anomaly
0.05	0.06	0	0	1	1	0	0	anomaly
0.05	0.07	0	0	0	0	1	1	anomaly
0.05	0.07	0	0	1	1	0	0	anomaly
1	0	0.12	0.03	0	0	0	0	normal
1	0	1	0.2	0	0	0	0	anomaly
0	0.07	0	0	1	1	0	0	anomaly
0.01	0.06	0	0	1	1	0	0	anomaly
1	0	0.01	0.02	0	0	0	0	normal
1	0	1	1	0	0	0	0	anomaly
1	0	0.02	0.03	0	0	0.02	0	normal
1	0	0.01	0.04	0	0	0	0	normal
0.09	0.05	0	0	1	1	0	0	anomaly
0.07	0.06	0	0	0.99	1	0	0	anomaly
1	0	0.01	0.02	0	0	0	0	normal
0	0.85	1	0	0	0	0	0	normal
0.01	0.06	0	0	1	1	0	0	anomaly
0.1	0.05	0	0	0.53	0	0.02	0.16	normal
0.05	0.07	0	0	1	1	0	0	anomaly
1	0	0.02	0.14	0	0	0.56	0.57	normal
1	0	0	0	0	0	0	0	normal
1	0	0	0	0	0	0	0	normal
1	0	1	0.51	0	0	0	0	anomaly
0.23	0.04	0.01	0	1	1	0	0	anomaly
1	0	0.11	0.01	0	0	0	0	normal
0	0.31	0.28	0	0	0	0.29	1	anomaly
0.98	0.01	0	0	0	0	0	0	normal
0.12	0.05	0.05	0	0	0	0	0	normal

Figure 4.5: dataset Train.csv

For Test.csv Our Dataset features for our model :(**duration , src-bytes , dst-bytes , land , wrong-fragment, urgent ,hot, num-failed-logins , logged-in , num-compromised , service-telnet , service-tim-i , service-time , service-urh-i , service-urp-i , service-uucp , service-uucp-path , service-vmnet**

, service-whois)

dst_host_	dst_host_	dst_host_	dst_host_	dst_host_	dst_host_	dst_host_srv_rerror_rate
0.06	0	0	0	0	1	1
0.06	0	0	0	0	1	1
0.04	0.61	0.02	0	0	0	0
0	1	0.28	0	0	0	0
0.17	0.03	0.02	0	0	0.83	0.71
0	0.01	0.03	0.01	0	0	0
0.72	0	0	0	0	0.72	0.04
0	0	0	0.01	0.01	0.02	0.02
0	0.01	0.03	0	0	0	0
0.08	0.02	0	0	0	0	0
0.01	0	0	0	0	0.66	0.32
0.03	0	0	0	0	0.33	0
0.07	0	0	0	0	1	1
0.07	0	0	0.69	0.95	0.02	0
0.05	0.03	0.04	0	0.77	0	0.07
0	0.01	0.04	0	0	0	0
0	0	0	0	0	0	0
0	0.03	0.05	0	0	0	0
0	1	0	0	0	0	0
0.07	0	0	0	0	1	1
0.05	0	0	0	0	1	1
0.01	0.01	0	1	1	0	0
0	0	0	0	0	0	0
0	0.03	0.02	0	0	0	0
0.06	0	0	0	0	1	1
0.06	0	0	0	0	1	1

Figure 4.6: dataset Test.csv

4.4 Environment Setup

4.4.1 Linear regression Algorithm

X - Matrix (m x n) Y - last column from data frame (class)

Algorithm fillmissing(df, feature, method):

begin

if method == “mode” then

df[feature] = df[feature].fillna(df[feature].mode()[0])

else if method == “median”:

df[feature] = df[feature].fillna(df[feature].median())

else

df[feature] = df[feature].fillna(df[feature].mean())

End fillmissing

-Mode = most common value

-Median = middle value

-Mean = average

Lets define a cost function which gradient descent will use to determine the cost of each theta. The cost function will implement the following cost equation.

$$J^{(i)} = (\frac{1}{2m}) \sum_{j=1} (h\theta(x^{(i)}) - y^{(i)})^2$$

Algorithm computeCost(X, y, theta) :

begin

m = len(y)

diff = np.matmul(X, theta) - y

J = 1 / (2 * m) * np.matmul(diff, diff)

return J

End computeCost

We now go into our gradient descent loop, where we calculate a new theta on each loop and keep track of its cost. See the equation below:

Repeat untill convergance {

First calculate the hypothesis and then its cost with equation using this equation

$hc^{(i)} = h\theta(x^{(i)}) - y^{(i)}$

Here is the equation to calculate new thetas using the learning rate. This equation is made easier with matrix manipulation and the fact that we added a column of ones to X

$\theta_n = \theta_n - \alpha(\frac{1}{m}) \sum_{j=1} hc^{(i)}.x^{(i)}$

Keep track of the cost of the new theta as we go:

$J^{(i)} = computeCost()$

}

Algorithm gradientDescent(X, y, theta, alpha, num-iters) :

begin

m = len(y)

J-history = [ ]

for all i in range(num-iters) do

hc = np.matmul(X, theta) - y

theta -= alpha / m * np.matmul(X.transpose(), hc)

J-history.append(computeCost(X, y, theta))

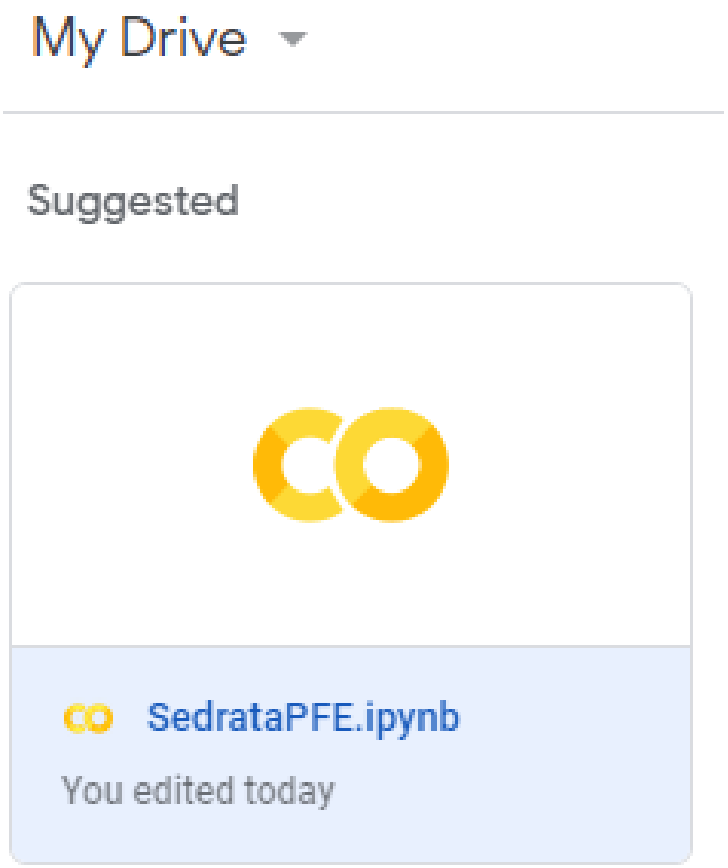
end for

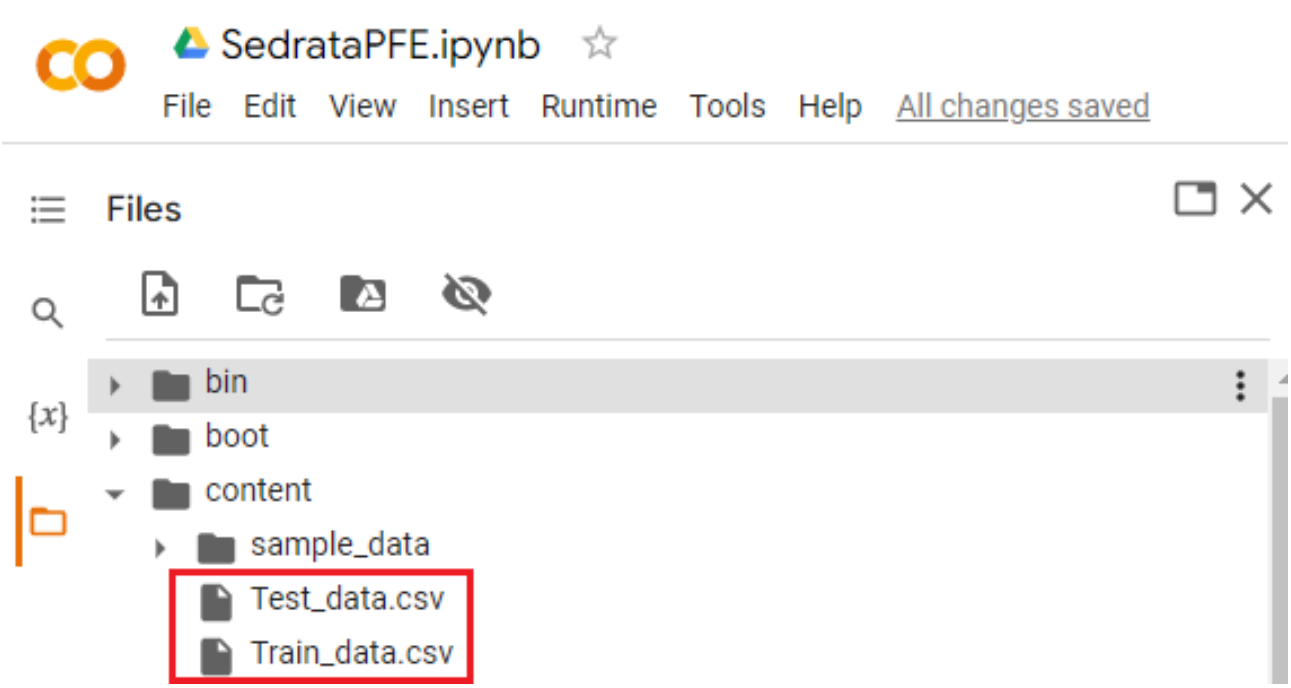
return theta, J-history

End gradientDescent

4.4.2 Work steps with pictures

File and dataset location





Import and Load dataset

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.preprocessing import LabelEncoder
from sklearn import metrics
```

```
[246] df = pd.read_csv('/content/Train_data.csv')
      df.head()
```

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	...	dst_host_srv_count	dst_host_same_srv_rate	dst_host_diff_srv_rate	dst_h
0	0	tcp	ftp_data	SF	491	0	0	0	0	0	...	25	0.17	0.03	
1	0	udp	other	SF	146	0	0	0	0	0	...	1	0.00	0.60	
2	0	tcp	private	S0	0	0	0	0	0	0	...	26	0.10	0.05	
3	0	tcp	http	SF	232	8153	0	0	0	0	...	255	1.00	0.00	
4	0	tcp	http	SF	199	420	0	0	0	0	...	255	1.00	0.00	

5 rows × 42 columns

ıte	dst_host_serror_rate	dst_host_srv_serror_rate	dst_host_rerror_rate	dst_host_srv_rerror_rate	class
.00	0.00	0.00	0.05	0.00	normal
.00	0.00	0.00	0.00	0.00	normal
.00	1.00	1.00	0.00	0.00	anomaly
.04	0.03	0.01	0.00	0.01	normal
.00	0.00	0.00	0.00	0.00	normal

Prepare the data

```
columns = ['protocol_type','service','flag','class']
for feature in columns:
  le = LabelEncoder()
  df[feature] = le.fit_transform(df[feature])

Y = df['class']
df = df.drop(['class'], axis=1)
```

Fill in values for our missing features

```
def fillmissing(df, feature, method):
  if method == 'mode':
    df[feature] = df[feature].fillna(df[feature].mode()[0])
  elif method == 'median':
    df[feature] = df[feature].fillna(df[feature].median())
  else:
    df[feature] = df[feature].fillna(df[feature].mean())
```

```
features_missing= df.columns[df.isna().any()]
for feature in features_missing:
  fillmissing(df, feature= feature, method= 'mean')
Y.fillna(Y.median(), inplace=True)
```

Get X/Y arrays

```
X = df.to_numpy()
y = Y.to_numpy().transpose()
m,n = X.shape
```

X shape

```
array([[ 0.  ,  1.  , 19.  , ...,  0.  ,  0.05,  0.  ],
       [ 0.  ,  2.  , 41.  , ...,  0.  ,  0.  ,  0.  ],
       [ 0.  ,  1.  , 46.  , ...,  1.  ,  0.  ,  0.  ],
       ...,
       [ 0.  ,  1.  , 46.  , ...,  0.  ,  1.  ,  1.  ],
       [ 0.  ,  1.  , 38.  , ...,  1.  ,  0.  ,  0.  ],
       [ 0.  ,  1.  , 17.  , ...,  1.  ,  0.  ,  0.  ]])
```

Y shape

```
0         1
1         1
2         0
3         1
4         1
         ..
25187     0
25188     0
25189     0
25190     0
25191     0
Name: class, Length: 25192, dtype: int64
```

Normalize X

```
mu = X.mean(0)
sigma = X.std(0)
xn = (X - mu) / sigma
```

Add column of ones

```
xo = np.hstack((np.ones((m, 1)), xn))
```

Cost function

```
repeat = 10
lrate = 0.01
theta = np.zeros((n+1))
```

```
def computeCost(X, y, theta):
  m = len(y)
  diff = np.matmul(X, theta) - y
  J = 1 / (2 * m) * np.matmul(diff, diff)

  return J
```

theta shape

```
array([0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0.])
```

Gradient descent

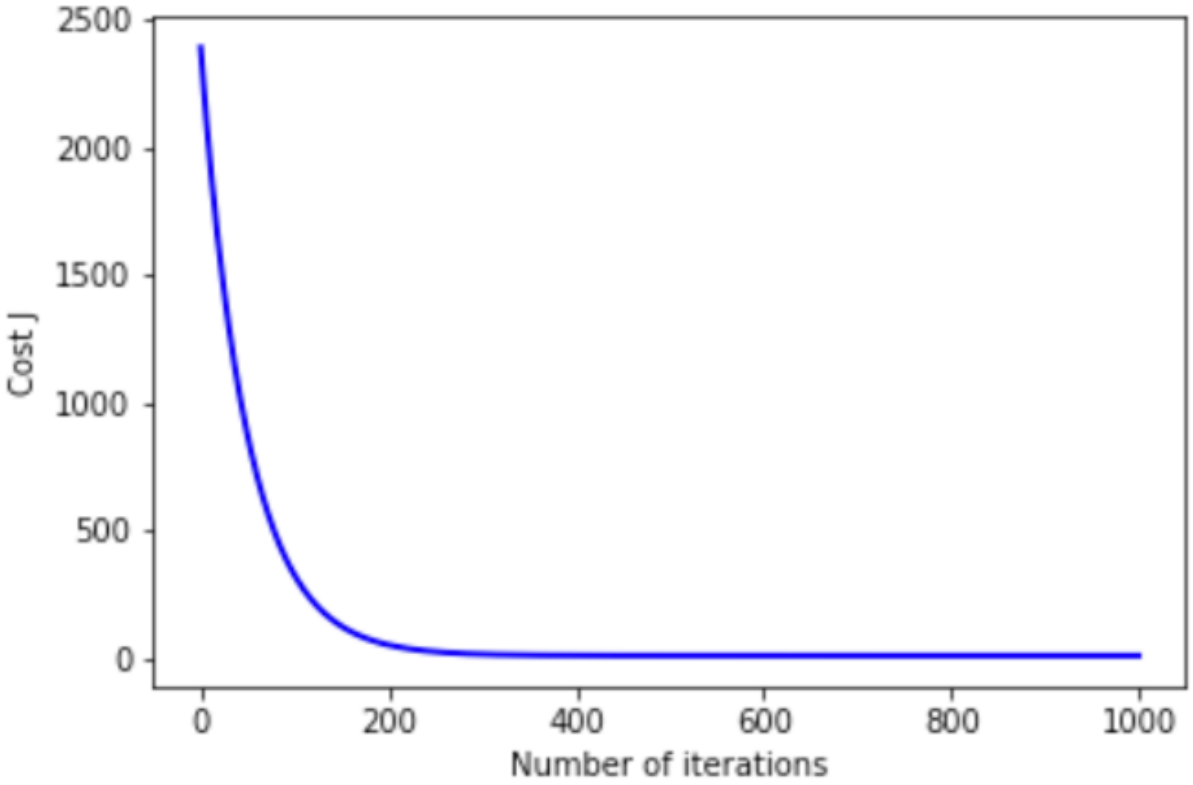
- repeat = number of times to repeat gradient descent
- theta = a theta for each feature of X, add one more column for theta 0
- costhistory = keep the cost of each iteration of gradient descent

```
def gradientDescent(X, y, theta, alpha, num_iters):

  m = len(y)
  J_history = []

  for i in range(num_iters):
    hc = np.matmul(X, theta) - y
    theta -=  alpha / (m * np.matmul(X.transpose(), hc))


    J_history.append(computeCost(X, y, theta))
  return theta, J_history
```

```
theta, J_history = gradientDescent(xo, y, theta, lrate, repeat)

print('Best theta computed from gradient descent: ')
print(f' {theta} ')
```

Plot the cost of gradient descent

```
plt.plot(np.arange(repeat), J_history, '-b', LineWidth=2)
plt.xlabel('Number of iterations')
plt.ylabel('Cost J')
plt.show()
```



Prediction

```
y_pred = np.matmul(xo, theta)
```

Evaluate predictions

```
diff = (y_pred / y * 100)
print('Mean of results: ',diff.mean())
print('Deviation of results: ',diff.std())
print('Results within 10% support/resistance: ', len(np.where(np.logical_and(diff>=90, diff<=110))[0]) / m * 100)
```

```
Mean of results:  100.40691268463144
Deviation of results:  6.765896949654286
Results within 10% support/resistance:  90.09530292716134
```

4.5 Conclusion

In this chapter, we described the implementation of our system, in which we set out the environment and development tools we used. The model of training and classification of the database has been applied with the parameters used and tested, as you can see, this algorithm, which uses many different parameters, yielded an accuracy value of almost 100 percent, as well. Finally, we explained the experiments and the results obtained.

general conclusion

Cyber-attacks are a very common concern at this time, and it's a diversion issue.If a person does not have a suitable security system, the linked information may be easily hacked, making these techniques obsolete, researchers are looking at applying machine learning in cybersecurity to provide more dynamic and robust protection for new types of attacks that we have not seen before.
Machine learning is still a developing field in the field of cyber security, and there is a lack of open source libraries, frameworks, and tools to use for threat and attack issues.

In this work, we have proposed a method of accuracy and approximation of the correct result in the event of attack or not using automatic regression algorithms. and it has worked.

To continue work on this project, we could increase the size of the dataset newtork intrusion detection by collecting more or even using the ones already built to increase efficiency,also, it will be more realistic if we can collect databases from our society not just from Kaggle.
if we can study the effect of choosing the value of alpha(learning rate) and see how results change according to change, it has improved the security process by using machine learning and prediction algorithms and using artificial intelligence (AI) to avoid cyberattacks a good job.

Bibliography

[1] Data mining process: Models, steps, applications, and techniques. URL: https://unstop.com/blog/data-mining-process.

[2] Neural network. URL: https://en.wikipedia.org/wiki/Neural_network.

[3] Qu'est-ce que le data mining ? URL: https://www.tibco.com/fr/reference-center/what-is-data-mining.

[4] Talend, article publier sur tout savoir sur l'exploration de données, ses avantages et sa mise en place. URL: https://www.talend.com/fr/resources/what-is-data-mining%5d.

[5] 2018. URL: "https://www.mindtree.com/insights/blog/reference-guide-implementing-data-mining-strategy

[6] Cyber security in focus, 2020. URL: https://hsfnotes.com/pwtd/2020/05/26/cyber-security-in-focus.

[7] 5 regression algorithms you should know – introductory guide!, 2021.

[8] 2022. URL: https://coretelligent.com/insights/what-is-the-cia-triad-and-why-does-your-cybersecurity-

[9] Data mining : qu'est ce que l'exploration de données ?, 22 mai 2022. URL: https://www.lebigdata.fr/data-miningdefinition-exemples.

[10] Hassen M. Alsafi. Intrusion detection system (ids). 2013. URL: https://www.researchgate.net/figure/Intrusion-Detection-System-IDS-IDS-are-classified-according-to-the-audit-data-to-tow_fig2_260432735.

[11] Ashesh AnandSep. 12 applications of data mining, Sep 17, 2021. URL: https://www.analyticssteps.com/blogs/12-applications-data-mining.

[12] Rashmi Bhardwaj. Ids vs ips vs firewall – know the difference. URL: https://ipwithease.com/firewall-vs-ips-vs-ids/.

[13] CISA CISM By Michael Swanagan, CISSP. How to prevent the top cyber attacks in 2022, 2022. URL: https://purplesec.us/prevent-cyber-attacks/.

[14] Brzychczy Edyta. An overview of data mining and process mining applications in underground mining. URL: https://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-40a08125-efdb-4714-84b1-8f3b0f3c06b0.

[15] Bhandari Behal Sahingoz et al iju Gopal Prakash itnick Simon Breda Barbosa Morais Ghankutkar et al. Ch et al. Swarna Priya et al Eric A. Fischer, aur Chahal. Creating a national framework for cybersecurity:an analysis of issues and options. 2009,2017, 2019 ,2020.

[16] Swati Gupta. A regression modeling technique on data mining. 2015.

[17] Seattle Hazard. City of seattle cemp – shiva, 2019. URL: https://www.seattle.gov/documents/Departments/Emergency/PlansOEM/SHIVA/SHIVAv7.0-Cyber.pdf.

[18] Hunter Heidenreich. What are the types of machine learning?, Dec 4, 2018. URL: https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f.

[19] Brian Jefferson. The 15 most common types of cyber attacks, 2021. URL: https://www.lepide.com/blog/the-15-most-common-types-of-cyber-attacks/.

[20] SuryaNepal JulianJang-Jaccard. A survey of emerging threats in cybersecurity. August 2014.

[21] kaspersky. What is cyber security?, 2021. URL: https://www.kaspersky.com/resource-center/definitions/what-is-cyber-security.

[22] Jack Holsey L. LABOVITZ. Dating mining and implementation. URL: https://brainmass.com/engineering/mining-engineering/dating-mining-and-implementation-43547.

[23] MARK L. LABOVITZ. What is data mining and. URL: https://genderi.org/what-is-data-mining-and.html.

[24] MARK L. LABOVITZ. What is data mining.doc. 08.10.2017. URL: http://cs.furman.edu/~pbatchelor/mis/webarticles/What%20Is%20Data%20Mining.doc.

[25] George Lawton. logistic regression. URL: https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression#:~:text=Logistic%20regression%20is%20a%20statistical,or%20more%20existing%20independent%20variables.

[26] Manas Gaur Manju Khari. Meticulous study of firewall using security detection tools. January 2013. URL: https://www.researchgate.net/publication/265485821_Meticulous_Study_of_Firewall_Using_Security_Detection_Tools.

[27] M.Sowmiya P.S.Seemma, S.Nandhini. Overview of cyber security. 2018. URL: https://www.researchgate.net/publication/329678338_Overview_of_Cyber_Security.

[28] Alfred Chandler Samuel Crashin, Richard Usoni. Data mining methods: Strategies and algorithms on different applications. March 2015. URL: https://www.researchgate.net/publication/277028230_Data_Mining_Methods_Strategies_and_Algorithms_on_Different_Applications.

[29] JitendraAgrawal ShikhaAgrawal. *Survey on Anomaly Detection using Data Mining Techniques*. 2015.

[30] Priyanka Sinha. Multivariate polynomial regression in data mining: Methodology, problems and solutions. 2013. URL: https://www.researchgate.net/profile/Priyanka-Sinha-12/publication/264425037_Multivariate_Polynomial_Regression_in_Data_Mining_Methodology_Problems_and_Solutions/links/5d44d03aa6fdcc370a76b505/Multivariate-Polynomial-Regression-in-Data-Mining-Methodology-Problems-and-Solutions.pdf.

[31] Talend. article publier sur tout savoir sur l'exploration de données, ses avantages et sa mise en place. URL: https://www.talend.com/fr/resources/what-is-data-mining.

[32] Nilsu Goren Theresa Hitchens. *International Cybersecurity Information Sharing Agreements*. 2017.

[33] Nicholas Tsagourias. Cyber attacks, self-defence and the problem of attribution. 2012.

[34] Pavan Vadapalli. 6 types of regression models in machine learning you should know about, 2020. URL: https://www.upgrad.com/blog/types-of-regression-models-in-machine-learning/#Types_of_Regression_Analysis_Techniques.