MINISTRY OF HIGHER EDUCATION AND SCEINTIFIC RESEARCH

ÉCOLE NATIONALE SUPÉRIEURE DE MANAGEMENT
ENSM. Pôle Universitaire de KOLÉA

A Dissertation Submitted in Partial Fulfillment of the Requirements for Master's Degree in Management

Major: Strategic Management and Information System

The Application of Big Data Techniques in Risk Management

Case Study: Algerian Customs

Submitted By:

Belkacem Chemlal

Ramzi Boualem Adjal

Supervised By:

Pr. Redha Tir

Year 2019/2020

Abstract

Big data is the new hot trend in the business world. Lately, big data analytics has been taking massive leaps as a potential and applicable solution to almost every operational challenge company decision-makers are facing nowadays. This research strives to integrate the concepts of Big data, Fraud prediction, and risk management process. The purpose of this research is to find out how big data or big data tools like machine learning can be utilized in the processes of fraud prediction to make the risk management department of Algerian customs more efficient.

Keywords: Big Data, Risk Management, Big Data Analytics, Fraud, Machine Learning.

Resumé

Le Big Data est la nouvelle tendance en vogue dans le monde des affaires. Dernièrement, l'analyse des Big Data a connu une énorme importance en tant que solution potentielle et applicable à presque tous les défis opérationnels auxquels les décideurs sont confrontés de nos jours. Cette recherche s'efforce d'intégrer les concepts de Big data, prévision de la fraude et le processus de gestion des risques. Le but de cette recherche est de découvrir comment les big data ou les outils de big data comme l'apprentissage automatique peuvent être utilisés dans les processus de prédiction de la fraude pour rendre le service de gestion des risques des douanes algériennes plus efficace.

Mots clés : Les mégadonnées, Gestion des risques, Fraude, Apprentissage automatique.

ملخص

تعتبر البيانات الضخمة في الآونة الأخيرة هي الاتجاه الجديد لتحقيق الكفاءة في عالم الأعمال، بحيث اكتسبت تحليلات البيانات الضخمة أهمية هائلة كحل محتمل وقابل للتطبيق لجميع التحديات التشغيلية التي تواجه صناع القرارات تقريبًا. يسعى هذا البحث إلى دمج مفاهيم البيانات الضخمة والتنبؤ بالاحتيال وعملية إدارة المخاطر. الهدف من هذا البحث هو اكتشاف كيف يمكن استخدام البيانات الضخمة أو أدوات البيانات الضخمة مثل التعلم الآلي في عمليات التنبؤ بالاحتيال لجعل مديرية إدارة المخاطر للجمارك الجزائرية أكثر كفاءة.

الكلمات المفتاحية: البيانات الضخمة، إدارة المخاطر، الاحتيال، التعلم الالي.

Acknowledgements

All the praises and thanks be to Allah. May peace and blessings be upon His prophet Muhammad and his family and companions.

We wish to express our gratitude to Pr. Redha Tir, our thesis supervisor, for his guidance throughout the research work.

His experience, insightful comments, and guidance were essential to us in this research endeavor.

Last but most important, we profoundly thanks our family for their innumerable support. Every member of our family has been supportive, not only laying a solid foundation for our career but also in building it higher.

Belkacem Chemlal.

Ramzi Boualem Adjal.

Thank you

List of content

List of tables

List of figures

List of Abbreviations:

BD: Big Data

BDA: Big Data Analytical

IS: Information System

BI: Business Intelligence

OLAP: Online Analytical processing

ML: Machine learning

SQL: Structured Query Language

NoSQL: Not only SQL

HDFS: Hadoop distributed File System

NLP: Natural Language Processing

ACFE: Association of Certified Fraud examiners

INTRODUCTION

1 General Introduction

Big data has been taking massive leaps in companies from different industries as a powerful and applicable solution to the most of the operations challenges that company decision makers are facing nowadays. Fraud is one of the most risks that companies are facing in their daily activities; it has the potential to cause significant financial and non-financial harm to businesses, with wide ranging implications from costing business not only millions of dollars a year but also their reputation.

Furthermore, the proliferation of the internet has exposed financial systems to diverse fraudsters using different mechanisms to exploit financial systems. This provided a huge evolution in attack patterns, which rendered the once effective case-based fraud detection solutions no more effective as the computational complexity increases with each new detected fraud.

The emergence of Big Data analytic tools means to address Fraud detection model property for risk management strategies. Such technology allows the integration of data from various sources used to model and predict financial fraud. For example, location data of a fraudster, social-media activity, and credit card information can be reconciled to trace a fraudulent transaction to him.

Machine learning is one technique that can handle such complex abstractions. It is good at analyzing and organizing a large amount of unsupervised data. Most raw data in Big Data Analytics are largely unlabeled and uncategorized, which are ideally suited for Machine learning algorithms.

In this paper, we will use big data analysis techniques in the fraud detection process and study its impact. The most suitable technique that can handle such a complex operation is machine learning as we mentioned before this is why we choose it to work with, in this work we will find out if it has a good impact and if we will recommend it or not.

2 Theoretical and methodological contributions:

This research was inspired from various RESEARCH PAPER's, where we found that this theme has been discussed in many different ways by several authors. In which some of them talked about big data and its effect in the enterprise in general like Muller (2018). In the other hand, Hossien (2018) and Lackovic (2016) saw that big data is better to be used in banks, and another team of researchers who focused in the department instead of the whole enterprise like IBM (2014), Banarescu (2015), and Chen. Who proved that big data could fit risk management needs and improve its performance in a very impressive way no matter what was the type of the enterprise. Besides that, we have also reviewed the various theoritical and technical studies that exist on the fraud prediction using big data, which was offered by Choi (2017), Gandomi (2015), and many others.

To answer our research question, we adopted a model-based approach and the best one that serves our objectives it was Action research study which was proposed by (Boog, Ben, et al.1996) that allowed us to strengthen our work which is based on fraud detection by using Big Data tool.

3 Problem Statement:

In the present study, our main research question has been formulated as follows: How can Big Data be used to predict and detect fraud in customs**?**

In order to meet our research objective we need first to seek answering the next sub question:

_ What is the followed process to identify risk in customs?

_ How can big data be implemented in risk detecting and predicting?

_ What are the most significant fraud antecedents?

4 Aim of the study:

The aim of this study is to Design a Big Data model by using machine-learning principles that will provide predictive and adaptive fraud detection. The positioned model aims to reconcile some of the well-known risk management challenges. Thus, the simplified model will outline the different stages of improving risk management process through the Big Data solution. Moreover, it will discuss the integration point between risk management and Big Data. Finally, it will demonstrate the model using existing tools and technologies.

5 Scope of the Work

The scope of the work include the following:

1. Design of a Machine-Learning model for fraud detection on a Big Data Platform
2. Implementing the proposal in item 1 above using existing tools and technologies
3. Providing an outline of Big Data analytics challenges solved by Machine-learning

6 Presentation of the host organization:

The General Directorate of Algerian Customs missions are generally set out in customs law and specified by Article 3 of the Algerian Customs Code. Other texts of a legislative or regulatory nature entrust the customs administration with the application of provisions relating to border control, in particular those governing the sectors of trade, finance, national defense, agriculture, transport. Industry, health, transport, tourism, information, and culture.

The development of international trade and the opening of borders have led States to entrust customs with missions to protect public health, public morals, public security and industrial, commercial and intellectual property right.
Its economic missions are represented in the following:
_ Apply, in collaboration with the institutions concerned, the laws and regulations governing the cross-border movement of goods;

_ Promote fair competition by preventing, investigating and punishing unfair and fraudulent practices;

Figure 1 Algerian Customs principal mission

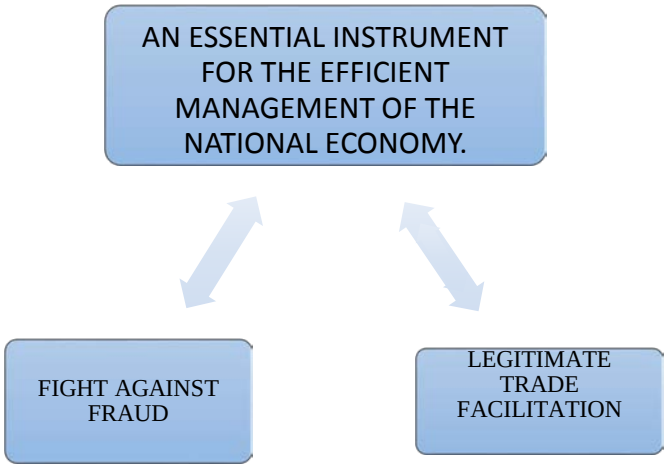
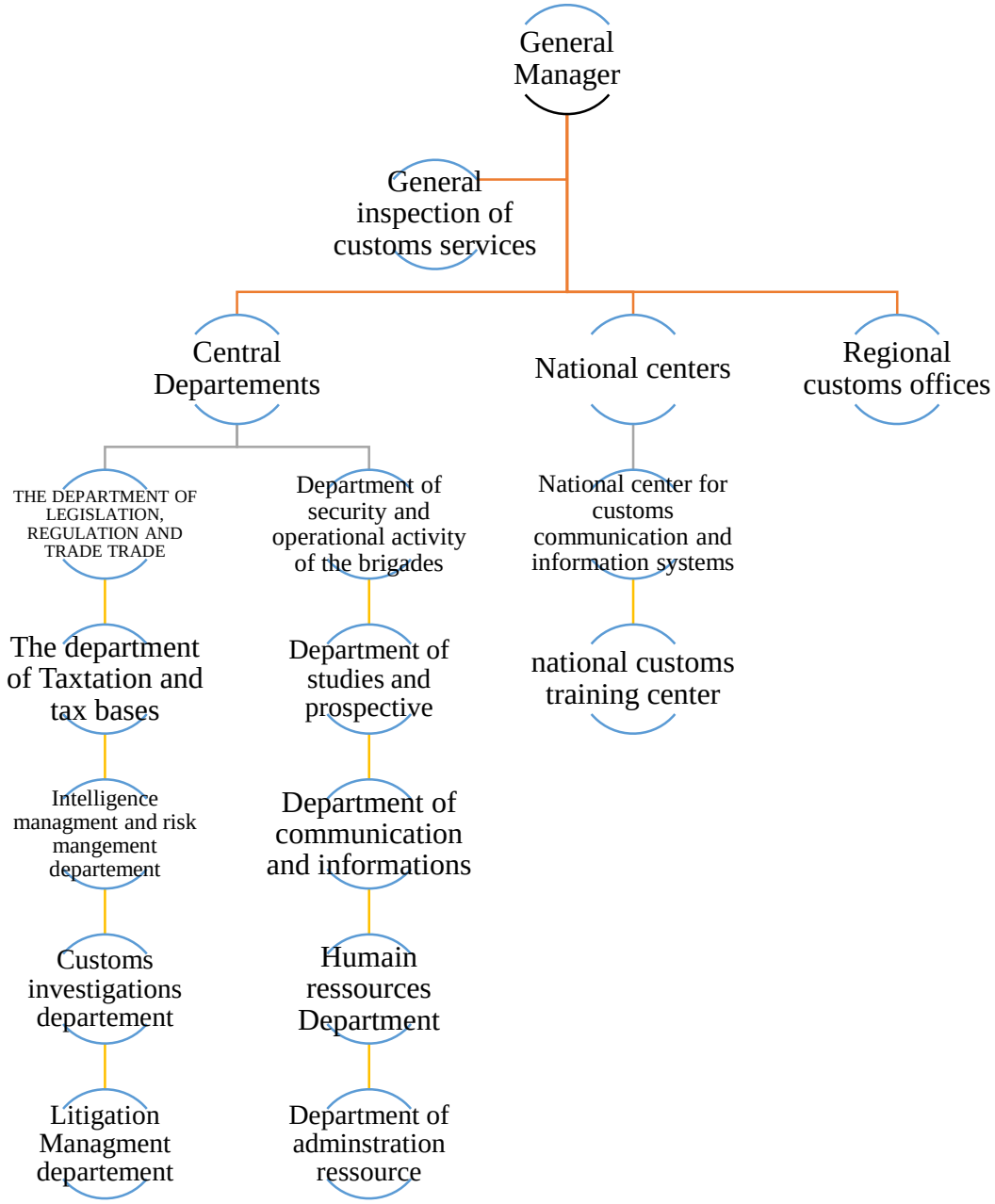


Table 1 Algerian Customs Data sheet

Data Sheet	
Organization name	The General Directorate of Algerian Customs
Logo	
Head Office	Dr Ch. Saadane St, El Djazair 16000
General Manager	Noureddine Khaldi
Creation Date	1964
Affiliation	Ministry of Finance

The Algerian Customs possesses important knowledge through its experience and by adapting foreign trade policy to Algeria's new priorities, upgrading the country's Customs Administration in order to meet international norms and standards, rolling out a new information system relying on digitization and automated Customs procedures, and providing support for economic operators working in foreign trade under the Customs-Business Partnership principle. Moreover, this knowledge and world modernization would not be possible without its highly qualified staff, its management and its organizations. The organization chart below provides an overview of the general structure of the organization.

Organizational chart presentation:



7 Research Organization:

In order to accomplish this work successfully, it has been divided into four parts. The first part determines the Problem Statement, the relevance, the questions and the objectives of the research, thus briefly presenting the studied company. The second part which is the 1st chapter addresses the question of the literature review and the conceptual framework. The seconde chapter deals with the strategy, the approach, the method of collection, and analysis. In the third and last chapter, we presented the results obtained through the ready readable data analysis, then we will interpret them and we will determine the level of organization performance in fraud prediction, beyond we will propose areas for improvement.

CHPTER ONE: LETERATURE REVIEW & CONCEPTUEL FRAME

We will present briefly through this chapter the most important relative research papers in the field of big data and fraud risk management, and we will present the relative conceptual framework in the era of big data in risk management.

1 Literature review:

A literature review is a comprehensive summary of previous research on a topic. The literature review surveys scholarly articles, books, and other sources relevant to a particular area of research (Lawrence, Brenda. 2008).

1. About Big Data:

Big Data already existed at the end of the 1990s and has spread enormously in the 21st century, becoming, in the current context, a key element for modern business. Companies all over the world are exploring these large volumes of highly detailed data to discover previously unknown information that is useful in improving the decision-making process (Hasnat, 2018).

Big Data refers to sets of data so large that they cannot be used with traditional database management systems because their dimensions overwhelm the capability of the software tools and storage systems commonly used to acquire, store, manage and process data within a tolerable period (Hasnat, 2018). Generally, Big Data refers to enormous series of data, both structured and unstructured, with wide, varied and complex structures, which is generated, captured and stored at incredible speed (Sagiroglu and Sinanc, 2013; Srivastava and Gopalkrishnan, 2015). Some authors define Big Data resources as high volume, high speed and high variety, requiring innovative and economical forms of information processing—so-called huge information—for better understanding and decision-making capability. Differently, for other researchers, Big Data indicates not only the set of data, but also the set of technologies that carry out all the functions mentioned and that exploit the value of the data and make its use economical and effective (Lackovic et al., 2016).

2. Big Data Characteristic:

We can identify three main features that characterize Big Data, also known as the 3Vs: (1) Volume, (2) velocity and (3) variety (Ozkose et al., 2015; Sagiroglu and Sinanc, 2013). Volume refers to the quantity of data and, therefore, to the dimensions of the dataset. Regardless of whether it is important, the data is very large. Velocity refers to the speed of

the data flow, that is, the rate at which the information is generated and spread and therefore the rate at which it is processed and analysed (velocity of data and processes). Variety is the characteristic that makes the data 'big' and is related to the typology of the information sources and the generated data, which can be structured, unstructured or semi-structured and can derive from different sources, both internal and external. Some researchers attribute to Big Data two more characteristics: Variability and veracity. Variability concerns the periodicity—or irregularity—and, sometimes, the incoherence of the data (Elgendy and Elragal, 2014). Veracity concerns the accuracy of the data, which can be good, not good or undefined, with the data potentially being incoherent, incomplete or ambiguous (Gandomi and Haider, 2015; IBM, 2014). Some authors identify a further characteristic, value, referring to the potential value of the data (Choi et al., 2017; Ozkose et al., 2015).

3. Big Data sources

In 2015, the United Nations Department of Economic and Social Affairs classified Big Data into three categories according to the different sources from which it derives: Data from social networks, including information from social media, messages, and research conducted on the internet. The second one is data from traditional systems of business, such as data generated by commercial trade transactions, e-commerce, credit cards, and data from the so-called Internet of Things (IoT), referring to machine-generated data, such as that concerning weather and pollution, data from GPS satellites and data from computer-based registers (Hasnat, 2018).

4. Big Data process:

The process of extrapolation of information is articulated in two phases: The first, known as data management, consists of the acquisition, storage, selection and representation of data; the second, called analytics, is composed of all the activities focused on the analysis and interpretation of data (Gandomi and Haider, 2015; Krishna, 2016). Information system tools allows to extract data for external sources, then transformed and loaded into advanced databases or data warehouses. As results, the data can be filtered and classified, then making it available for data mining and other forms of analysis. Finally, it is processed and submitted

to the Big Data Analytics (BDA) tools necessary to make the Big Data useful in the decisionmaking process (Munesh and Mittal, 2014).

5. Big Data Analytics:

Big Data Analytics defined as the processes that, using algorithms, analyse data sets to extract diagrams, reports and useful and unknown information. It is used to extract models and information that is valid, useful and previously unknown or hidden in large data sets, as well as to identify important relationships between variables, thus ensuring a competitive advantage (Elgendy and Elragal, 2014). Other authors consider BDA to be the tools that can generate intuitions useful in the decision-making process, assess company business performance, establish competitive advantages and, therefore, increase enterprise value (Saggi and Jain, 2018). Such a large and varied data set requires a capacity for storage, management and analysis that common software does not have. Indeed, traditional databases or data warehouses are insufficient and are unable to address the problems of selection, adaptability and usability of data—essential characteristics for the use of Big Data to achieve its expected benefits for improvement of the decision-making process and, consequently, an increase of business value. The rapid evolution of technology and the exponential increase in the data flow available has made necessary the development of more rapid and efficient tools for both the conservation and analysis of this data (Elgendy and Elragal, 2014). This has led to the development of advanced BDA based on tools such as NoSQL, BigQuery, Map Reduce, Hadoop, Flume, Mahout, Spark, WibiData and Skytree (Saggi and Jain, 2018), which can collect and analyse large and varied data very quickly to detect hidden models, unknown correlations, market trends, customer preferences and any other information considered useful.

6. Big Data uses in Risk Management:

There are still few studies concerning the use of Big Data in the risk management sector, but, in the last decade, growing interest from researchers and sector experts has been observed. Numerous studies highlight the positive relationship between the use of technological innovations in the risk management sector, including technologies based on Big Data, and business productivity.

Many researchers note that technologies relating to Big Data are applicable in many areas of the risk management sector, including, commercial (risk analysis, customer and sales management), capital markets (negotiation and sales, structured finance) and asset management (wealth management, management of capital investments) (Lackovic et al., 2016; Mohamad et al., 2015). In the following section, after illustrating integrated risk management, we will focus on the employment of BDA in risk management to appreciate the usefulness of these tools of storage, interpretation and management of data.

7. Risk Management in the era of Big Data

The correct functioning of a business activity and the contextual creation of an enterprise's economic value cannot neglect monitoring of the main risk factors, as represented by financial and managerial indicators whose economic effect can compromise performance. The clear and evident interconnections and interdependences between business risks have led to an increasingly global management of enterprise risks following a systemic approach that is coherent with the growth path of a company and a contextual transversal analysis of heterogeneous processes, functions and activities (Bhimani, 2009; Liebenberg and Hoyt, 2011; Ellul and Yerramilli, 2013). It follows that, in the last few years, overhauling of the traditional approach characterized by a mainly sectorial and fragmented view of risks ("silo" management) has resulted in the spread of a new philosophy in the management of business risk that involves the completely organizational structure and affects strategic and operational processes. This approach is known as enterprise risk management (ERM), and provides for integrated risk management through an analysis of business contingencies and an evaluation of uncertainty (Beasley et al., 2008; De Loach, 2000; Idris and Norlida, 2016; Liebenberg and Hoyt, 2003, 2011; Navak and Akkiraju, 2012).). For this, big companies pay particular attention to understand the different typologies of the sources of uncertainty to which the company is exposed (Hossein et al., 2018). There is therefore a greater awareness of risk-taking and greater selectivity of the purposes thereof, thanks to proactive involvement and communication of top management with supervisors.

It is obvious that all the companies especially the one who are active online as GOOGLE and Amazon -which have massive data to deal with every day-, increasingly need to use big data to predict risks, manage them and report them. The quantity and quality of data are essential elements for the formulation and implementation of strategies compatible with risk appetite and suitable for structuring effective and reliable processes and procedures for safeguarding the integrity of the company.

2 Conceptual Work

The conceptual framework explains the path of research and grounds it firmly in theoretical constructs. The overall aim of the conceptual frameworks is to make research findings more meaningful, acceptable to the theoretical constructs in the research field and ensures generalizability.

2.1 Risk Management:

The risk management has been a main subject for a lot of researchers and authors and every one of them has his own understanding to it, which means that the definition of this concept is deferent from one another. Among these definitions, we chose the following: According to (Merna and Smith 1996) risk management is *"an approach by which uncertainty can be understood, assessed, and managed within projects". As such, it forms an integral part of project management, and effective Project Risk Management is a critical success factor for project success". Another definition: "Risk Management refers to the culture, processes, and structures that are directed towards the effective management of potential opportunities and adverse effects. The risk management process involves "establishing the context, identifying, analyzing, assessing, treating, monitoring, and communicating risk"*. From the above, we can say that Risk Management is an organized process that aims to reduce risks if it is possible or minimize damages to reasonable losses that can be contained by studying risks starting from suspicions (analyzing events that may cause it) until it really happens (results, and how to deal with it to limit losses).

2.1.1 Risk Management Fundamentals:

a) **Uncertainty:** is when there is more than one possible out come to a made decision without knowing the probability of each outcome, because of the lack of enough knowledge (web site)

Uncertainty exists in every project because there is no identical project every project is deferent than the other so always, we find a range of uncertainty.

b) **Danger:** the danger indicates a potential nuisance that can harm people, property (deterioration or destruction) or the environment. (CEI, 1998)

c) **Risk:** it is the probability of happening of unexpected events because of the uncertainty associated with a particular course of action, which can affect the objectives or can cause a variation in the desired and planned outcomes. (Phil. Dale, 2005)

d) **Risk Sources:** there are many sources that can cause risk for any corporation or project, and from these sources, we mention:

- Political: change in government policy, public opinion, disorder (war, manifestations)
- Environmental: Contaminated land or pollution liability, nuisance.
- Market: Demand (forecasts), competition, obsolescence, customer satisfaction
- Economic: Treasury policy, taxation, cost inflation, interest rates, exchange rates
- Financial: Bankruptcy, margins, insurance, risk share
- Natural: weather, earthquake, fire or explosion, (COVID-19)
- Human: Error, incompetence, ignorance, tiredness, communication ability,
- Technical: machinery breakdown, adaptation to new technologies.

2.1.2 Risk management tools and techniques:

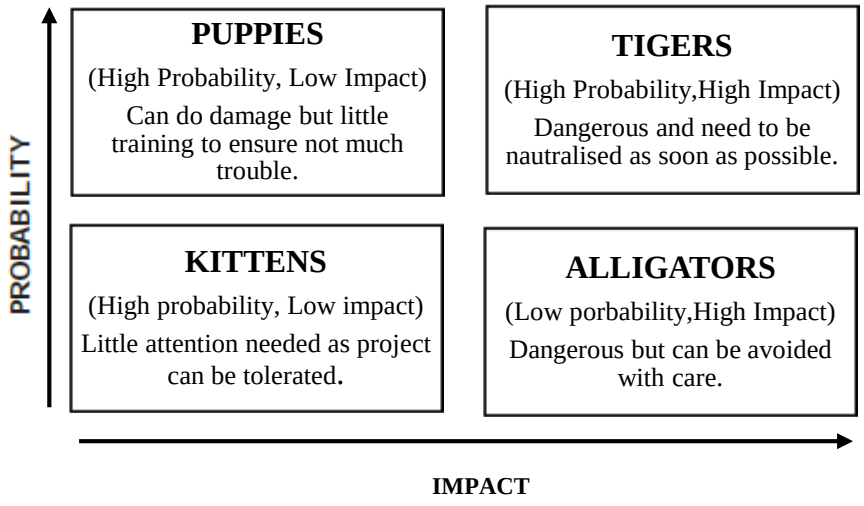
Risk management is currently one of the main areas of interest to researchers and practitioners working in a wide range of projects. In the framework of this research, many developed technics and tools was been provided in the way of improving and increasing its accuracy and effectiveness.

a. Qualitative Techniques in Risk Management:

- **Brainstorming:** it is a situation where a group of people (maximum size 12 people) gathered for a period of time (optimum period 15-45mn) and generate ideas to find solutions to specific problems, this technique has proved its effectiveness in advertising before this why it used later by all kind of business, engineers, scientist, and in risk management has been useful in many ways such as:
  - More brains means too many different views that helps to see the weakness points in the project.
  - More people can take a risky decision than one because of the factors of responsibility.
  - Learning from different experiences to handle the possible risks.

- **Interviews:** in this technique generally, the enterprise personnel interview project personnel about the potential risks at the project. We usually use this technique when group is impractical, or the information required are more detailed than group can provide.
- **Risks Matrix Chart:** this technique starts with identifying risks then assessed them on their impact and their probability of happening. It is allow us to identify high impact risks and low impact risks and classify them in a matrix.

Figure 2Risk matrix chart

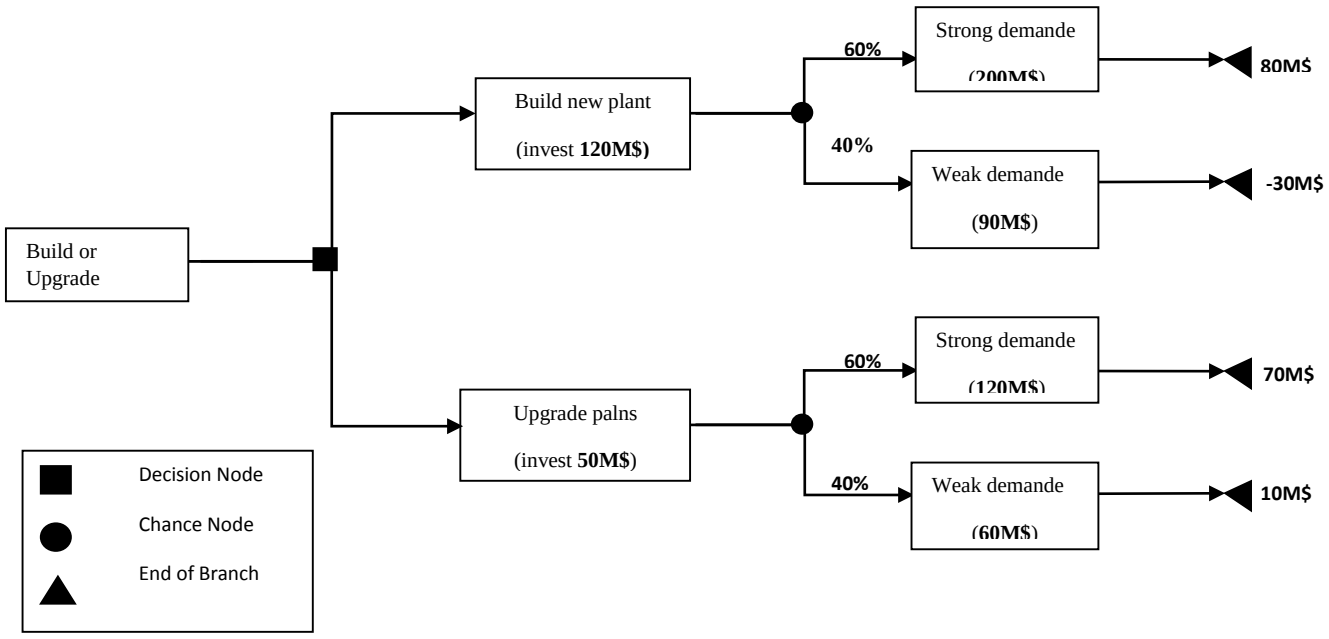


Source: (A.merna. 2005)

b. Quantitative Techniques in Risk Management:

- **Decision tree:** A manager is often faced with many decisions, which are faced by many options. In addition, each option has its outcomes, which complicated to analysis. The decision tree is a diagram that depicts all the possible decisions and its potential events and outcomes under each possible circumstance and that facilitate the analysis and the decision-making.

Figure 3 Decison tree diagrame



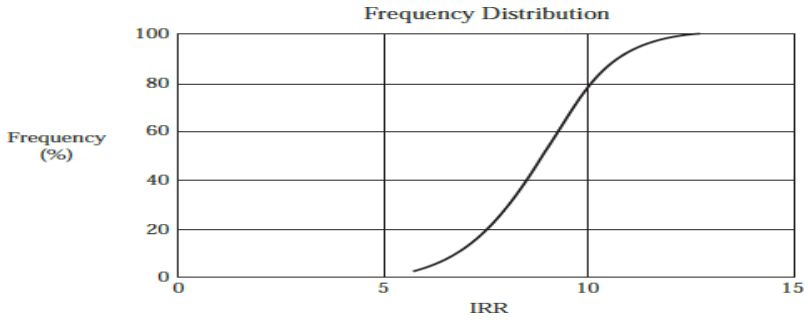
Source: (saade merie, 2016)

- **Monte Carlo Simulation:** A simulation is a science of designing a model system that reacts the same way as the real one. A Monte Carlo simulation is a simulation that depends on using random samples (numbers) and determines how the system reacts to a different input and simulates their different consequences. (minitab, 2020)

- **Sensitivity Analysis:** In the planning stage of any project, the changes that may occur for the data being used is an inevitable event, in other words, there will be a range of uncertainty, for this the sensitivity analysis is used to produce more realistic values supported by a range of possible alternatives, and that's by dividing the uncertainty into a mathematical model or system, it is used to:
  - Identify the most sensitive variables affecting the project.
  - Identify the risks, which have a potentially high impact on the cost or timescale of the project.
  - Identify the point at which a given variation in the expected value of a cost parameter changes a decision.
  - It shows the robustness and the ranking of alternative plans
  - Determine the range of the change of each variable attached with the expected possible range of minimum and maximum effect on the project that leads to determining when risks get important.

Next figure represents the uncertainty in a project in terms of IRR (internal rate of return). In this example, the project has a 40% chance of the IRR being less than 7.5% and a 60% chance of it being greater than 7.5%. Similarly, the project has an 80% chance of the IRR being less than 10% and a 20% chance of it being greater than 10%, with a 50% chance of it being less than or greater than 8%.

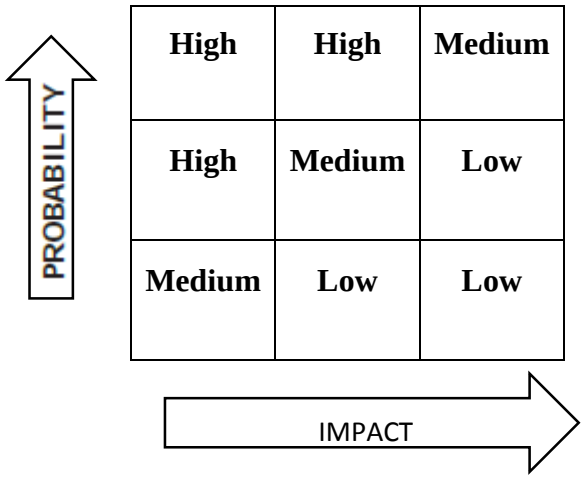
Figure 4 Cumulative probability distribution



Source : (A.merna, 2005)

- **Probability–Impact Grid Analysis:** after determining the impact parameters for a risk (cost, program, performance) in the Risk Management Plan(RMP), and determine its range of impact on the project in the Strategic Business Unit(SBU) and at the project management, this technique is used to rank it in a Probability - Impact Grid (PIG) using a broadband rating system.

Figure 5 Probability–impact grid



Sourche: (W.Hubbard .2009)

c. The choice of the best technique :

As we saw in the above, many techniques analyze risks, and every technique has its own characteristics that may fit some projects and doesn't fit others, so how do we know which technique will fit our project? The determination of the most suitable technique for a specific case we should consider (Toni. Altani, 2006):

- The availability of resources for analysis – human, computational, and time.
- The experience of the analysts with different techniques.
- The size and complexity of the project
- The project phase in which the analysis takes place
- The available information
- The purpose of the analysis.

The data used in the analyses should be considered:

- Accuracy: are the data accurate?

- Adequacy: are they adequate for the purpose of the project?
- Relevancy: are they relevant to the subject?
- Coherence: has the information been classified in an orderly and meaningful way?
- Impartiality: has the analyst remained unbiased?
- Direction: does the analytical procedure lead to conclusions/ decisions?
- Logicality: is the reasoning sound?
- Validity: are comparisons, interpretations, and implications valid?

2.1.3 Risk Management Process:

We have mentioned in this paper that risk management is an organized process that involves determining steps that the responsible should follow to be able to identify, analyze, manage, and monitor risks. From that, we can derive risk management process stages such as (alk. 2008) mentioned:

a) **Identify the risk:** consists of determining risks and sources that might cause risks, internal and external, once which has a high effect on the project, documenting its characteristics, and rank it according to their impact on the project time, cost, and objectives, using historical and current information (Bruce, Lyon. 2013). This process should be carried out in all the stages of the project. The purposes of this stage are:

- Provide a basis for the subsequent management.
- Provide the necessary information to conduct a risk analysis.
- To identify the inherent risks in the project or service.
- To identify the project or service components.

Knowing the needed inputs and the expected outputs will make this stage clearer and easier to understand, so the inputs of identifying risks are:

- project or investment description
- Planning output (cost, time, tools, specification requirements…)
- Historical information.

- Information on similar projects if it was possible

Moreover, its outputs are:

- Potential sources of risks
- Potential risks events

After these identifications, a very important step occurs, which is verification, and the verification of the information used in the identification is a very important step to avoid unreliable information. In addition, risk responses should be taken into consideration.

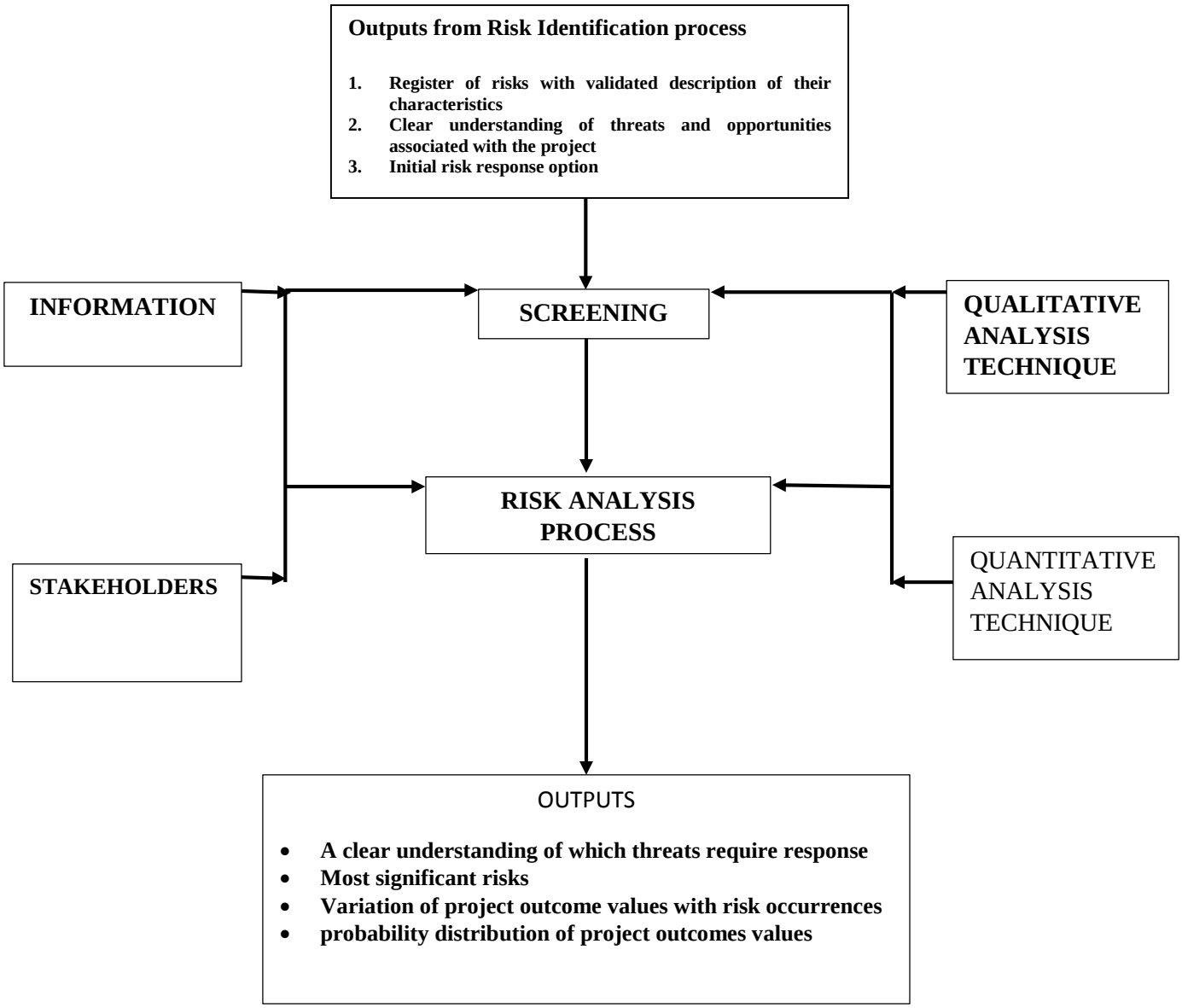
From the above, we understand that this stage depends on information. The better the informational foundation of the risk management process we have, the more accurate its results are. However, this information may or maybe not be available; this is why the process of gathering information must determine what information is needed, where, when, and how it may be collected.

b) Analyze the risk: in this stage, the risk management stuff uses the qualitative and quantitative techniques mentioned above to analyze the output of the previous stage (risk identification) in order to determine opportunities that should be pursued and threats that require more attention. The final purpose of this stage is to determine the balance exists between opportunities and threats. The major outputs of the risk analysis stage are**:**

- A clear understanding of which threats require response and which opportunities Should be pursued.
- A list of risks and their descriptions of their likely outcomes.
- Risks must be deal with, and risks should be avoided.
- How alternative effect the project objectives.

The stockholders take the decision of how to deal with risks and which alternative to be chosen because they make sure that the project does not contradict with objectives of the company itself. From that, we can resume this stage like that:

Figure 6 Risk Analysis Process



Source: (Norman marks. 2015)

c) Evaluate the risk: this step is about ranking and classifying risks based on the combination of their probability and their impact on the project (time, cost, and objectives), probability – impact matrix is the best technique for that. This stage aim is to facilitate decision making about whether the risks are acceptable or serious.

d) Treat the risk: or response to the risk, this is the step of taking an action to deal with the risk. There are three categories of how to deal with the risk:

- **Avoid the risk:** avoiding risks means eliminating the risk and that by eliminating the source of the risk in the project or avoid the project or investment that is expose to the addressed risk.
- **Risk reduction:** this category involves either lowering the probability of risks or minimizing its impact on the project or both.
- **Risk transfer:** Is a technique were risk is a technique where one party assume liability of another party or transfer risks to a third party, the most common example for that is insurance, there are two common methods to transfer risks:
    - **Insurance:** insurance companies charge fees to accept this kind of risks, and the higher the risks the higher the fees.
    - **Indemnification clause in contracts:** In this case, of the contractors will have liability on the risks, by involving a clause that ensure that potential losses will be compensated.

  There are many factors to take in consideration when taking a risk transfer decision:

  Who can best handle the risks if they materialize?

  What is the cost/benefit of transferring risk as opposed to managing the risk internally?

- **Risk Retention:** This technique is used when:
  Risk cannot be avoided
  The company have full knowledge about the risk impact on the project
  Transferring risk will cost more than handling it internally
  Reduction risk may only be cost effective up to a point, thereafter becoming more costly than beneficial.
- **Risk Sharing:** is a way of self-insurance where contractors involved in their contracts that the two part both will take liability of potential risks.

  The choice of the technique in this stage is based on the first stages outputs (risk identification and risk analysis), if there was a mistakes in the first two stages there will be a wrong choice of technique, which leads to unexpected losses that can affect the whole project in time schedule, costs and objectives.

e) Monitor and control the risk: monitor risks is the process of monitoring the implementation of agreed-upon risk response plans, tracking identified risks, identifying and analyzing new risks, and evaluating risk process effectiveness throughout the project. Not all risks can be eliminated- some risks are always present. Market risks and environmental risks are just two examples of risks that always need to be controlled. The purpose of project risks control is to:

- To confirm if risk responses setup is complete
- To monitor the project for new risks
- To monitor risks triggers

2.2 **Fraud**:

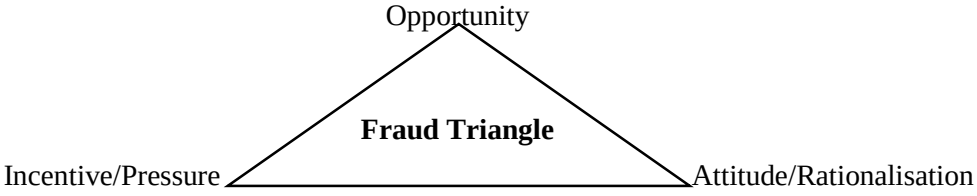
There are two principal methods of getting something from others illegally. Either they can be physically forced, or they can be deceived into giving up their assets. The first type is called robbery and the second is a fraud. (Albrecht et al, 2009) defines fraud as a deception made for personal gain. Deception is key. The most common definition of fraud according to Webster's Dictionary (2001 p.380) is:

"Fraud is a generic term that embraces all the multifarious means which human ingenuity can devise, which are resorted to by one individual, to get an advantage over another by false representations. No definite and invariable rule can be laid down as a general proposition in defining fraud, as it includes surprise, trickery, cunning and unfair ways by which another is cheated. The only boundaries defining it are those which limit human knavery."

2.2.1 Fraud triangle:

There is broad consensus that fraud has its roots in opportunities, incentives, and attitudes Rationalization, which is well-known as the “Fraud Triangle” (Intal T. & Do L.T. 2003, p25) For the purpose of risk assessments for fraud, stakeholders should keep in mind that fraud typically includes these three characteristics:

Figure 7 fraud triangle



Source : (Intal T. Do L.T.2003, p25)

Incentive/Pressure: may be anything from unrealistic deadlines and performance goals to personal vices such as gambling or drugs.

Opportunity is an open door for solving a non-shareable problem in secret by violating a trust.

Opportunity: provided generally through weaknesses in the internal controls.

Attitude/Rationalization: is a crucial component of most frauds because most people need to reconcile their behaviors with the commonly accepted notions of decency and trust.

Fraud-fighters usually work mostly on one of the three elements of the fraud triangle.

a) Fraud prevention :

"Fraud prevention describes measures to stop fraud from occurring in the first place." - Bolton et al. (2000).

Although fraud prevention involves a complex and sensitive process of balancing an organization's diverse interests and limited resources, a pervasive view (Albrecht et al. 2006; etc.) is that preventing fraud is the most cost-effective way to reduce losses from fraud. As the incidence and losses relating to fraud increases rapidly each year, more and more organizations realize that paying for prevention is cheaper than the alternative being the cure they ultimately will need to conquer fraud.

Fraud prevention involves two fundamental activities as below:

1. Creating and maintaining a culture of honesty and integrity, and
2. Assessing the risk of fraud and developing concrete responses to minimize risk and eliminate opportunity.

To eliminate opportunities, the implementation of good internal controls as well as the carrying out of proactive audits would be an important overriding concern.

b) **Detection:**

"Fraud detection involves identifying fraud as quickly as possible once it has been perpetrated." – Bolton et al. (2000)

Fraud detection is a continuously effort because fraud detection is an important part of any overall business security strategy (Clemmons 2007), whenever it becomes known that one detection method is in place, perpetrators with intent to commit fraud will adapt their strategies accordingly. Since fraud is increasing dramatically with the expansion of modern technology,

resulting in substantial losses to the businesses, “fraud detection has become an important issue to be explored” (Lanza R.B. 2004, p1).

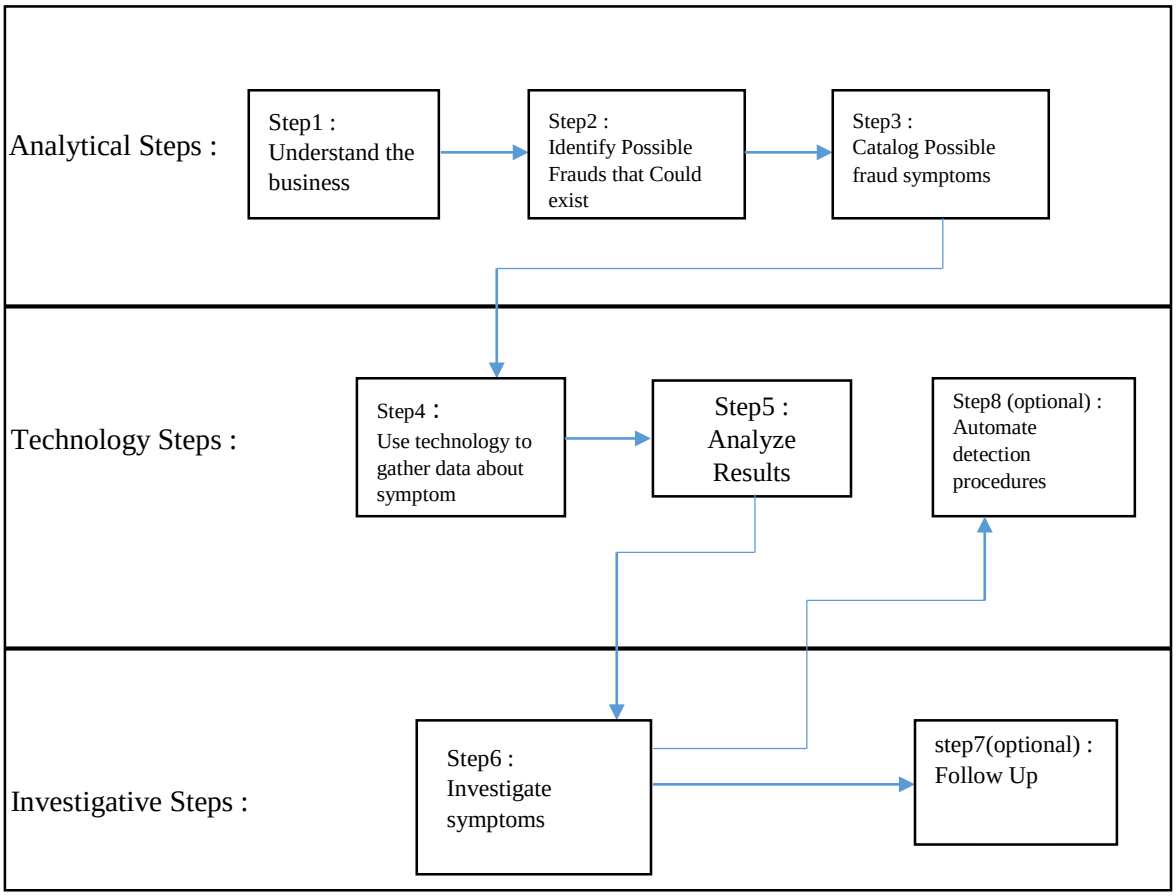
Primarily, there are two ways to detect fraud: 1) by chance and 2) by proactively searching for and encouraging early recognition of symptoms (Albrecht et al., 2006). According to a series of public survey reports from ACFE (2008), KPMG (2006), fraud was mostly detected by accident. Early detection and proactive detection efforts remain critical as most frauds start small and, if not detected, continue to get larger and larger.

2.2.2 Fraud detection process model:

- **_ Proactive Fraud Prevention and Detection:**

The proactive method of fraud detection is an effective way to detect and describe both known and unknown fraud. In 2003a, Albrecht et al. provided a map of categorization of fraud detection method, which is categorized into technology-based and non-technology-based methods, and the strategic method of fraud detection is proposed to match information based on the current widespread use of relational databases to store transactions. The method includes eight stages as illustrated in following Figure:

Figure 8 Strategic Fraud Detection Approach

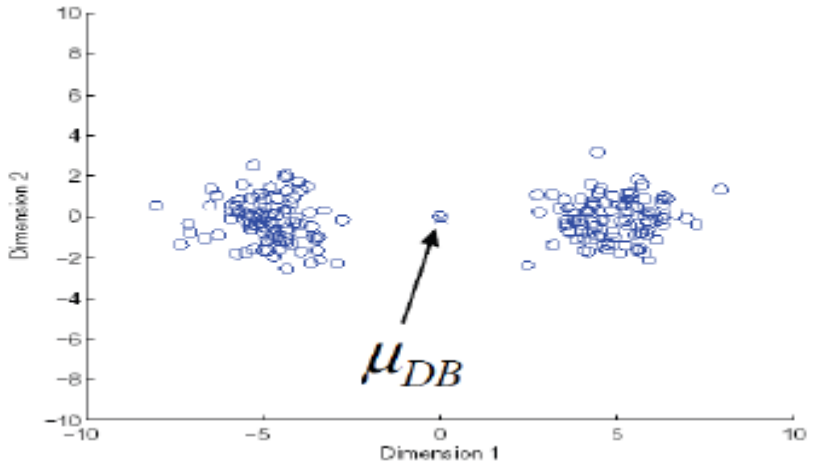


Source: (Albrecht et al. 2001, 2003a, p29, and Accountants’ Handbook2007)

- **Statistical fraud detection:**

Statistical methods have been used to classify and detect frauds. The transactional data is assumed to follow a statistical distribution, and thus, transactional data points that fall out of the normal distribution are considered suspicious. Such methods include Linear Discriminate Analysis and Logistic Regression (Richard and Bolton, 2002). Statistical methods can either be supervised or unsupervised techniques. Supervised methods use known fraudulent cases to model the detector system. A natural problem with the Statistical method is determining the most suitable distribution to fit a data set, and as dimension increases, it becomes more difficult to estimate the distribution, (Steinbach and kumar, 2005)

Figure 9 Figure showing Statistical model in 2-D used to classify data points and detect suspicious data.



(Anna Koufakou, 2009)

2.2.3 Integrating Big Data to Fraud Risk Management:

The utilization of big data has the potential to improve organization's risks strategies. It allows companies to adopt a predictive approach that increase the accuracy of the threat detections. As results, the organizations will institute mitigation mechanisms before the threats affect its normal operations. Big Data helps in the identification of project risks that have the potential to affect the organizations negatively. The evolution of technology has increased the risks of cyber-attacks and companies information system failure, which makes it necessary to devise a framework to detect threats before they affects crucial aspects of the business. Using big data is one of the surest ways to predict the security future of the project.

There are several areas in which big data can be used to prevent any risks that can occur in companies daily activities, the applications include:

- **Prevention of fraud:** The big data analytics approach has been adopted by large organizations, governments (customs ...), so that, the large volume of data is obtained from different sources which guarantee close monitoring of activities in different platform. Hence, the probability of detecting plans to engage in fraud before it happens is very high.
- **Credit Management:** Credit is a high risk that has the potential to paralyze operation of a business. Accordingly, it is paramount to manage this risk by analyzing big data to determine the previous economic history of the organization. This approach will enable to assess the payment patterns, airtime purchases, and other factors that may indicate other possible threats.

Big Data is a significant tool in risk management since it helps in evaluating data from different sources. The approach enables companies to monitor, detect, and mitigate all the risks that can affects its strategy negatively.

2.3 Big Data:

Big data is a concept used for data that has a big and complex volumes, number of transactions and number of data sources that requires a special methods and technologies to analyze it and draw insight out of data because the conventional methods and technologies cannot. in another word, BIG DATA is an approach "*or a set of technologies, architectures, tools and procedures*"(Gill Press, 2013) which consists in collecting and then processing in real time huge volumes coming from various sources, structured and unstructured, difficult to manage with traditional storage and processing solutions.

2.3.1 Characteristics of Big Data (5v):

Big data has been characterized in the beginning by the 3Vs (Volume, Velocity, and Variety), and then it comes more detailed, so the 3Vs become the 5Vs by adding to it (Veracity and Value), and with time and development of it technologies it gets characteristics that are more detailed and the 5Vs:

- **Volume:** they are the most important characteristics that characterize big data; we cannot speak about big data if we have less than 100 terabytes of data (van Rijmenam, 2013). The stocked volume of data is increasing every day from “1.2 Zo per year in 2010 into 40 Zo” per year in 2020 and it is continuing to increase every year. (1Zo = about 1E12 Go).
- **Variety:** The big volume of big data created by different types of data, it can be structured, semi-structured or non-structured for example: texts, voices, videos, glocalization data. This diversity of data in big data creates the need of different treatments methods and tools to make sense of it.

  Structured data: it is data characterized by: Organized logically, with and identical format and understandable by the computer.

  Unstructured data: it is the opposite of structured data; it is illogically crowded, with different format and requires a human intervention for treatment.
- **Velocity:** It is one of the advantages of big data, sometimes data need to be entered and processed as it is collected in real time, as in AliExpress, for example, velocity is the frequency in which data is generated, treated and shared, which means:
  - Data created faster
  - It must be collected faster
  - Treated faster (in the real time)
  - Fast conversion to a decision
  - Share decision at the same time
- **Veracity:** focus on the qualitative side of the used data, it answers the questions: can we trust the available data? Does it contain enough information?
- **Value:** it represents the added value of the data or the extracted information, if the data or the information doesn't have any added value it is not needed and the resources used in treating this data it is considered as a waste.

2.3.2 Big Data Architecture:

There are no standard or norms for big data architecture, problems are different and requires different solution and architectures, we will present the most common one:

a. Collection of data:

The source of the statistical process is the data, Knowledge, since it is the good knowledge of a subject that makes it possible to define the characters to observe, and then action.

There are two kinds of data. Data collected by the investigator himself for a specific purpose, and collected by someone else for some other purpose but it can be useful by the investigator in another purpose, the first one is the primary data and the last one is secondary data and each one has different techniques and sources to be gathered:

Table 2 techniques of collecting data

Technique	Definition	Advantages
Interviews	Involving asking questions and getting answers directly from respondents whatever, face to face, by phone or computer (internet).	• More personal than other methods. • The interviewer has the opportunity to probe or ask follow up questions. • Work directly with the interviewee make it easier to note opinions and impressions
Questionnaires	It is a research instrument for the aim of gathering information from respondents by a Series of questions and other prompts.	• Cheap • Less effort • Contain standardized answers that make compiling data simple and easy
Observation	Observation involves select what to observe, carful planning, record the observation in a way that allows to analyze and interpret the information	• It is relatively free of observer bias • It is precise. • Generalizability. Once you have devised your instrument, large samples can be covered. • Reliability can be strong.

Focus group	An in-depth field method that brings together a small homogeneous group to discuss topics on a study agenda, it aims to make use of the participant feelings, perceptions and opinions.	• Do not require participants to be literate. • Explore way and how is influences their beliefs and values. • useful when exploring cultural values and health beliefs
Case-study	In-depth investigations of a single person, group, event or community. Typically, data are gathered from a variety of sources and by using several different methods (e.g. observations & interviews).	• Study all the related aspects of the unit deeply and intensively • helps to find out the useful data • enables to generalize the knowledge

Source: (done by us)

Secondly sources of collecting second data: it usually costs less money and efforts and it is available, from these sources we mention:

- **Published Printed Sources:** their credibility depends on many factors as the writer, publishing company, time and date when it was published.
- **Published Electronic Sources:** nowadays internet has become more advanced, fast and reachable for most to the researchers, it became a source of gathering data, but credibility was always in question in this case but now problem solved.

b. Data storage:

After collecting data it needs to be stocked, storage in the case of big data will be facing three problems:

- **Volume:** it has very massive volume
- **Velocity:** Among the most important characteristics of big data is the speed, that is, the speed in storage and the speed in extracting information from the store, how much time it needs to store and instore the data.
- **Veracity:** there is a big veracity in big data, sometimes you must store all kinds of data(images, videos, texts, vocals, …) and make it easy and fast to instore

This why requires a special system to deal with, and here we present for you storage systems:

- **Distributed File Systems:** such Hadoop file system (HDFS) offer the capability to store in a reliable way a huge quantity of unstructured data, HDFS is a very important part in the Hadoop framework, it is well suited to quickly ingesting data and processing.
- **NoSQL: "***is an approach to database design that can accommodate a wide variety of data models, including key-value, document, columnar and graph formats. NoSQL, which stands for "not only SQL," is an alternative to traditional relational databases in which data is placed in tables and data schema is carefully designed before the database is built. NoSQL databases are especially useful for working with large sets of distributed data.",* (Margaret Rouse, 2017)

c. Data Analyze:

After collecting and storing data it come the part where collected data should be transformed to a useful information to take a decision through specific techniques, methods and software and as examples we chose:

Data analysis techniques:

- **Natural Language Processing (NLP):** it is the idea of developing a computer system, or program that can understand analyze and generate natural human language, there are many application based on this principle like concept search (Cortana in Microsoft, Siri in apple..), machine translation (Apple translator).
- **Genetic algorithms:** this technique is used for problems with many variables, it is inspired by Charles Darwin theory (the theory of evolution), and to understand the technique we should understand the natural evolution first.

  The natural evolution process starts by the selection of the fittest individuals from a population, use them to get married, and gives us better children and the children replace the other less fittest individuals, so the next generation get better than the one before and so on until we have the best generation. We can use that in a search problem, or decision taking. We consider a set of solutions and find the best one by passing through five phases:

- **Initial population:** at this step we chose the individuals that fits the problem, which means every possible solution to the problem, to make it understandable here an example, a

commercial enterprise they have 30 million $ as capital they want to select which products that they can sell and maximizing out comes at the same time,

Table 3 products with costs and benifits

Products	Costs	Benefits
Televisions	15	15
Phones	3	7
Laptops	2	10
Tablets	5	5
Smartwatches	9	8
Printers	20	17

To define the fittest individuals, we must chose know the problem criteria, the criteria here is the capital so the cost of the products cannot cost more than 30 m$. The next step here is to create chromosomes were chromosomes is a binary string, were for this problem 1 if the product has been chosen 0 if not, from that we create the population and choose the fittest individuals. For example, the first individual:

1	0	1	1	0	1

Individual (1)

This means that the chosen products are: Televisions, Laptops, Tablets, Printers. After creating all the individuals, we chose just the one who fits costs less than 30m$, for individual (1) for example:

Table 4 Prodcuts examle

Products	Costs	Benefits
Televisions	15	15
Laptops	2	10
Tablets	5	5
Printers	20	17
Total	42 > 30	47

This is individual will be eliminated, every individual in the population must cost less than 30m$.

- **Fitness function:** In this step, we set a fitness score that will help us to know which individual fits more than the other does, from that, we can define the fittest, the fitness score here is the benefits. We take two different individuals:

  Individual (2) individual (3)

1	0	0	1	1	0

0	0	1	1	1	0

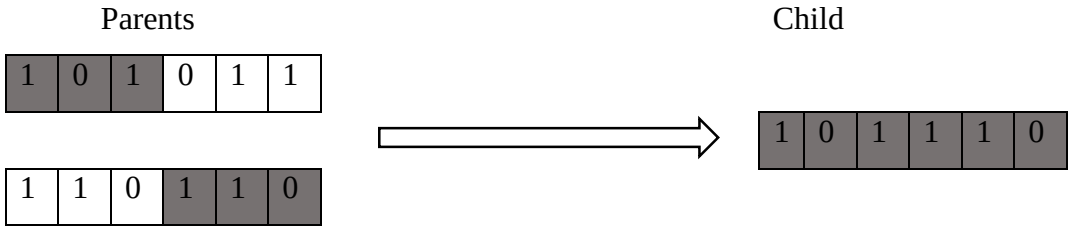
Which gives us:

Products	Costs	Benefits
Televisions	15	15
Tablets	5	5
Smartwatches	9	8
Total	29	28

Products	Costs	Benefits
Laptops	2	10
Tablets	5	5
Smartwatches	9	8
Total	16	23

From that we see that 28>23 which means individual (2) fits more individual (3)

- **Selection:** now we select the fittest individuals to be the parents of the next generations.
- **Crossover:** In this phase, we take the chromosomes of the parents and make a crossover between them to create new chromosome.



- **Mutation:** sometimes times the children does not fit the problem so we eliminate them, and sometimes we add a random mutation to make the fit the problem, mutation is making one random change in the chromosome (change 1 to 0 or 0 to 1). However, most

of time mutation are created to create some difference between children because each new generation they get closet to look like the same.

The operation still happens until generations still look like each other, this operation seems long ant take times but with technologies such (Python and MATLAB) it is taking less time.

- **Machine learning:** it is a science that appeared in the 50s that allows machines to discover patterns and make predictions from data using data mining, statistics, and predictive analyzes. within the development of machine learning tools and algorithms and the incredible increase of the data. machine-learning has been very effective in predictive analysis and decision making by predicting the consequences of a decision and consumer behavior, even it has been so useful in optimizing logistics and production lines.
- Regression analysis.
- Sentiment analysis.
- Social network analysis.

Data analysis tools:

R program: it is a language and a software environment used as a tool for statistical computing and graphics; it is widely used in statistics and data analytics it can be used in descriptive analysis and predictive analytics.

R has many uses such as:

_Linear and nonlinear modelling

_Classical statistical tests

_Time-series analysis

_Classification

R is a tool, used to make sense of big data and to gain use from it, and it can be used for:

_Visualization (Graphs, Charts, etc).

_Cleaning the data to extract useful information.

_EDA (Exploratory Data Analysis)

_Clustering

Application of R in the real world:

Amazon has used R in the aim of increasing their sales

Google has used R to improve search result

Bank of America has used R to predict financial losses.

- **Python:** is an open-source programming language, it promotes structured, functional and object-oriented imperative programming, it is the most used programming language in the world, this language has propelled itself to the forefront of infrastructure management, data analysis, or software development. Python provides a huge number of libraries for big data; it is used also in developing intern codes for big data it much faster than any other programming language which makes it the first choice when it comes to big data big data. It is used generally for:

  Application programming

  Creation of web services

  Code generation

  Metaprogramming (is a program that manipulates other programs (or itself) as its data).

- **Apache Spark:** is an open-source data-processing engine for large data sets. it provides a computational speed, scalability, and programmability required for big data (Holden, Konwinski, 2015). Spark has proved his efficacy in the real world and it is used for many known companies and for example:

Netflix: to offer a movie or Tv series that the viewer might like based on his favorite shows

Pinterest: use spark to get immediate insight into the how the users are engaging to make recommendations for them.

Types of data analysis:

- **Descriptive Analysis:** allows a user to get a detailed view on the business, and answer questions like: what happened? When? Where?
- **Diagnostic Analysis:** it allows you to understand the descriptive analysis and the business by getting answers to questions like, why does it happen. How? Moreover, what are the relations?

- **Predictive Analysis**: it allows a user to get a prospective view of the business and give him the ability to predict what will happen in it by answering questions like what could happen. What if these trends continue? In addition, what might happen next? Good predictive analysis is based on good descriptive and diagnostic analyses.

2.3.3 Big data challenges:

Even though big data has been appeared from the 90s but it still a new technology that have been developing extremely fast but it still has obstacles challenges that must deal with it in the future.

- Information quality: As we know, big data collect many data from many resources that minimize the credibility of the data and if the data used are not true the result is not right and will affect the decision-making.
- Security is a big problem in big data large information create a lack of security that makes it exposed to attacks.
- The human factor: the human mistakes are not detectable in big data that influence the results.
- Integrating disparate data sources: data come from different sources in different types, which made the integration a challenge.

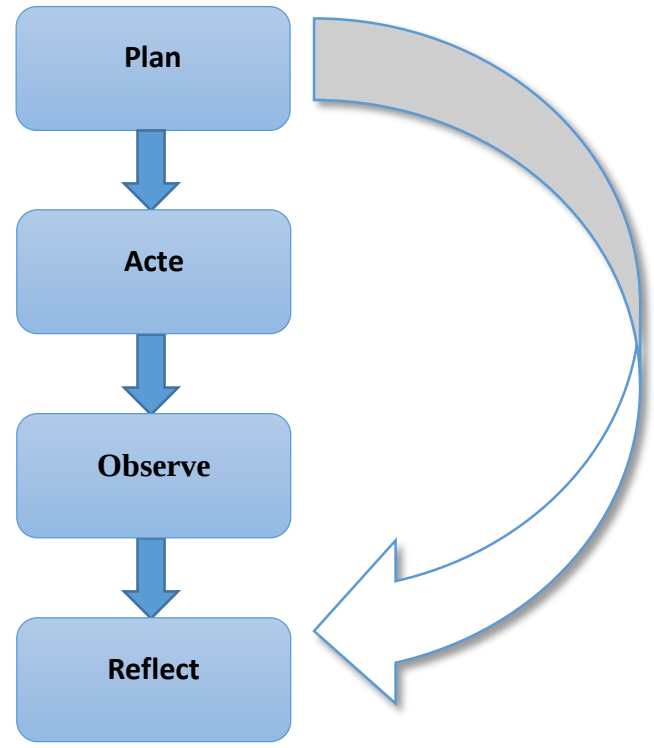
CHAPTER TWO: RESEARCH DESIGN & METHODOLOGY

This chapter presents the design of our research methodology, it also discuss the implementation process of our fraud prediction model based from the data collection step to the data analysis phase.

1 Research Method:

In order to meet our research objectives, we needed a scientific methodology that allow us to explore the studied context and build an intelligent solution, among the existed methods, we opted for Action method to conduct our research. (Thomas, Jim. 1986) define it as *"Action research aims to contribute both to the practical concerns of people in an immediate problematic situation and to further the goals of social science simultaneously. Thus, there is a dual commitment in action research to study a system and concurrently to collaborate with members of the system in changing it in what is together regarded as a desirable direction. Accomplishing this twin goal requires the active collaboration of researcher and client, and thus it stresses the importance of co-learning as a primary aspect of the research process."*

Figure 10 Simple Action Research model



Source: (peter reasen .2001)

The main reason why we choose the action research method because it's a type of applied research that is set on providing practical solutions to specific business problems by pointing the business in the right direction. Typically, action research is a process of reflective inquiry that is limited to specific contexts and situational in nature. Hence, the action research is used in reel situations, rather than in contrived, experimental studies, since its primary focus is on solving real problems. It can, however, be used by social scientists for preliminary or pilot research, especially when the situation is too ambiguous to frame a precise research question. Mostly, though, in accordance with its principles, it is chosen when circumstances require flexibility, the involvement of the people in the research, or change must take place quickly or holistically.

To address the most concerns of a Fraud detection system, the following infrastructural design choices were made.

2 Data Collection:

To answer our research question, we chose the secondary data type. According to (john, melkiv. 2004) *"Secondary data is the data that has already been collected through primary sources and made readily available for researchers to use for their own research. It is a type of data that has already been collected in the past."*

In order to give meaning to our work, it is very important to explain what is and from what these secondary data consists.

2.1 Primary Data:

Data gathered and collected by customs agents from the business partner declarations, customs declaration is obliged for every type of transition (export, import), in these declarations we find different documents and information as: Tax ID, Commercial register, and Document supporting the origin. From these declarations the customs agent extract the necessary informations and upload them to the customs database as excel file format. These informations may contain different data as: The bank used by the client, the origin of the merchandise, the weight, the amount of the merchandise, the legal status of the operator …etc.

2.2 Secondary Data:

Secondary Data: represent the excel file that contains the structured and organized data, we will use this secondary data into the program for the purpose of training the model to identify fraud, that is the more data we use the more precision we have from the model.

3 Implementation design :

As mentioned above that action method based action scientific research consist of testing a built model in reality, here we present our implementation design that we opted to build our model for fraud prediction it consist from statistical and algorithms phases.

Phase One "Regression Analyses":

An analysis that enable us to find a mathematical equation link between dependent variable and independent variables, the equation goes in the following form:

$$y = b1x1 + b2x2 + \dots + bnxn + c$$

Where Y= is the dependent variable in our case is fraud

$x_{1.....n}$= independent variables. $b_{1....n}$ = regression coefficients

The bigger the regression coefficient is the more the dependent factor influence the independent factor, in this way we can identify the factors that have the bigger impact in fraud and eliminate the insignificant factors to work just with the important once.

Phase Two "Machine-Learning model training and rules of extracting":

In this part, we initiate the machine-learning rule, after eliminating the insignificant factors, the machine-learning program will use only the important factors, and these factors data will be used to identify their impact on fraud, the identification process goes like the following:

Using statistical rules:

The program will study all possible relations between the dependent variables and the independent variables for extracting a precise equation of probabilities that aims to predict the impact of each variable on the fraud: $\mathbf{Y} = W_1P_1X_1 * W_2P_2X_2 * \dots\dots\dots\dots\dots W_nP_nX_n$

Where: **Y**= fraud

W= A statistical weight is an amount given to increase or decrease the importance of an item.

A weighting factor is a weight given to a data point to assign it if it's lighter, or heavier, importance in a group. It will be calculated automatically in the program.

P= the probability of fraud by factor.

X= the independent factor.

The customs inspector can change these rules (weight and probability of fraud per factor) if it is not compatible with their master data.

The result of this equation represents the probability of fraud in the collected data, and according to these results, the program will create maximum probability for the clean files and less probability for the suspected possible fraud

Phase three "Model Test":

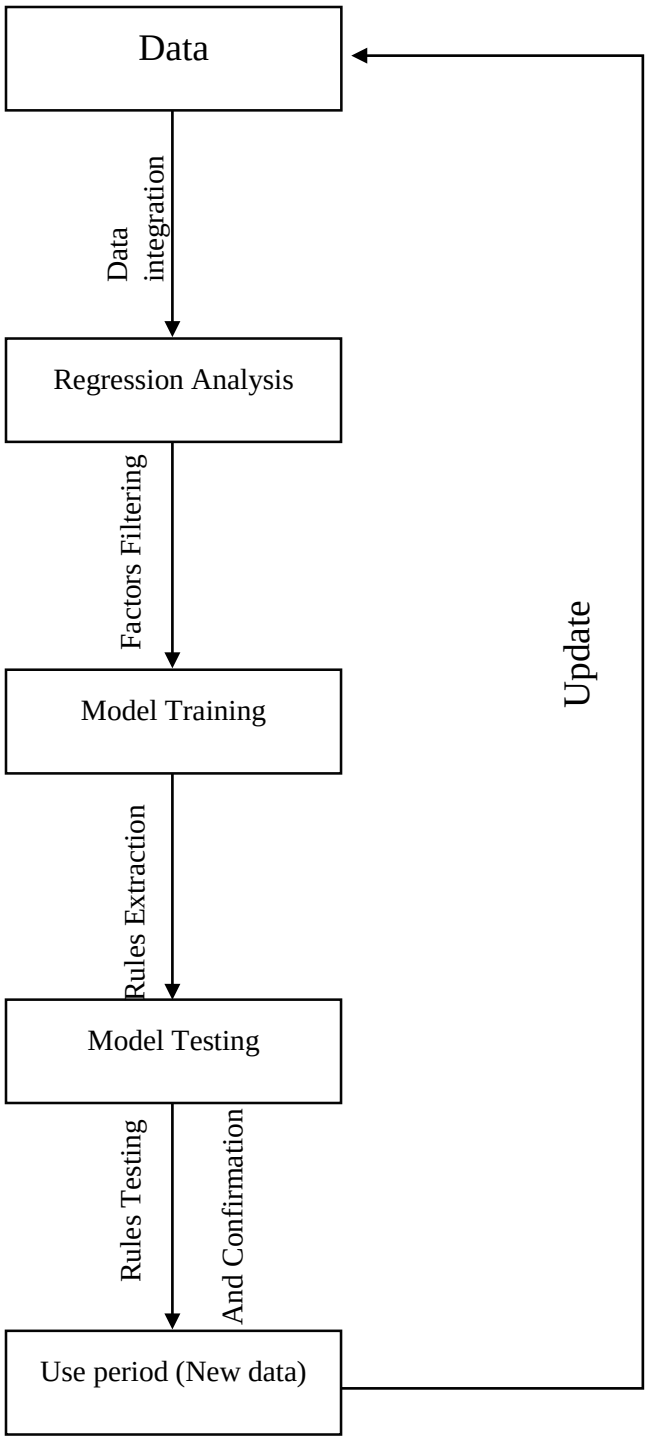
This phase consists of using old data to test the model, then we set the independent data and the program classify the files by using the knowledge he has acquired to determine if it is fraud or not and then we inform the machine if it was correct. After multi tests, the program will be ready to be used by the customs agent with high precision in detecting possible fraud files.

Phase four "Set Up and Updates":

This phase aims to set up the final version in service by the customs agent, who will enter the independent variables and confirm the results after each verification, he must reply to the machine if the file was fraudulent or not. This step will enable the machine to save the correct data and use it to update the model, which means after every declaration the model the legacy system will be upgraded.

To summarize the process, we propose the following schema:

Figure 11 Implementation Design Process



CHAPTER THREE:
PROTOTYPE DESIGN &
RESULTS ANALYSIS

In this chapter, we will present a very simplified fraud prediction prototype using Big Data technologies and regression analysis principles. It also includes the processes of the implementation of the proposed application and the analysis of the results

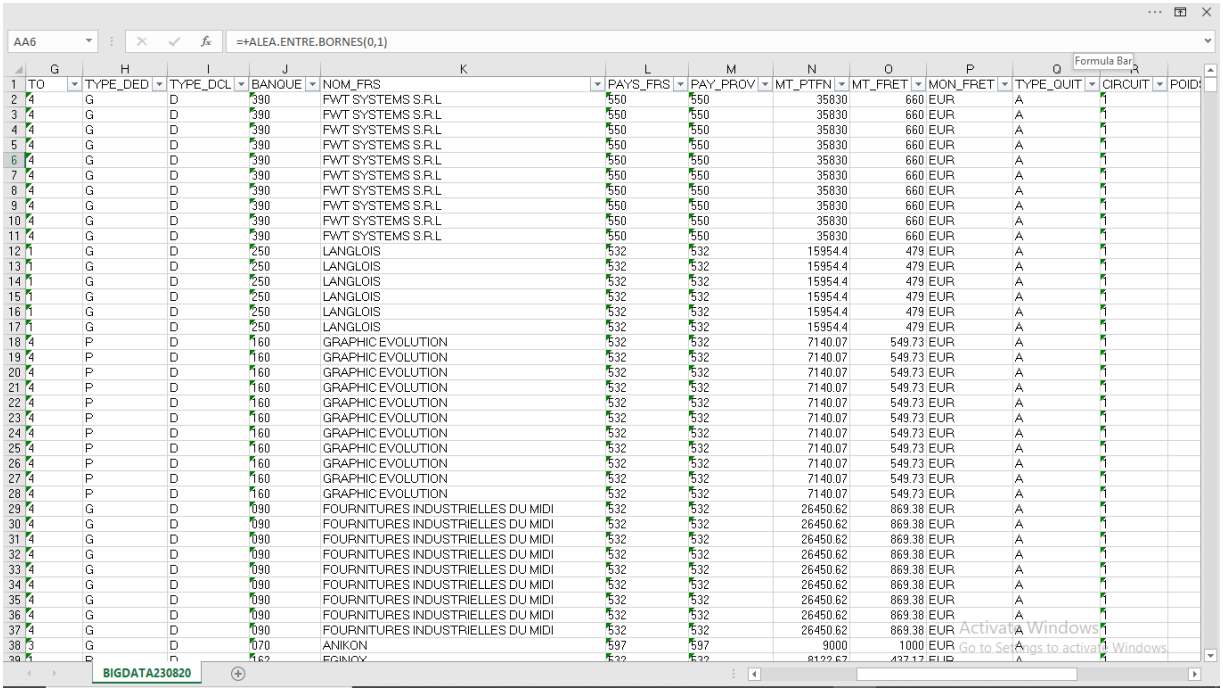
1 Dataset and processing algorithm:

In this part of the chapter, we will present the different steps to follow (algorithm) to setup and process the data of the empirical example studied.

1.1 Dataset:

The data that we feed into our machine-learning algorithm to train our model, it represents the secondary data that we received and contains 65536 files (Volume), in this files we found the information related to exportation and importation transactions of each client during the financial year 2014.

Figure 12 Secondary data (Dataset)



	G	H	I	J	K	L	M	N	O	P	Q	R	
1	TO	TYPE_DED	TYPE_DCL	BANQUE	NOM_FRS	PAYS_FRS	PAY_PROV	MT_PTFN	MT_FRET	MON_FRET	TYPE_QUIT	CIRCUIT	POID
2	4	G	D	390	FWT SYSTEMS S.R.L	550	550	35830	660	EUR	A	1	
3	4	G	D	390	FWT SYSTEMS S.R.L	550	550	35830	660	EUR	A	1	
4	4	G	D	390	FWT SYSTEMS S.R.L	550	550	35830	660	EUR	A	1	
5	4	G	D	390	FWT SYSTEMS S.R.L	550	550	35830	660	EUR	A	1	
6	4	G	D	390	FWT SYSTEMS S.R.L	550	550	35830	660	EUR	A	1	
7	4	G	D	390	FWT SYSTEMS S.R.L	550	550	35830	660	EUR	A	1	
8	4	G	D	390	FWT SYSTEMS S.R.L	550	550	35830	660	EUR	A	1	
9	4	G	D	390	FWT SYSTEMS S.R.L	550	550	35830	660	EUR	A	1	
10	4	G	D	390	FWT SYSTEMS S.R.L	550	550	35830	660	EUR	A	1	
11	4	G	D	390	FWT SYSTEMS S.R.L	550	550	35830	660	EUR	A	1	
12	1	G	D	250	LANGLOIS	532	532	15954.4	479	EUR	A	1	
13	1	G	D	250	LANGLOIS	532	532	15954.4	479	EUR	A	1	
14	1	G	D	250	LANGLOIS	532	532	15954.4	479	EUR	A	1	
15	1	G	D	250	LANGLOIS	532	532	15954.4	479	EUR	A	1	
16	1	G	D	250	LANGLOIS	532	532	15954.4	479	EUR	A	1	
17	1	G	D	250	LANGLOIS	532	532	15954.4	479	EUR	A	1	
18	4	P	D	160	GRAPHIC EVOLUTION	532	532	7140.07	549.73	EUR	A	1	
19	4	P	D	160	GRAPHIC EVOLUTION	532	532	7140.07	549.73	EUR	A	1	
20	4	P	D	160	GRAPHIC EVOLUTION	532	532	7140.07	549.73	EUR	A	1	
21	4	P	D	160	GRAPHIC EVOLUTION	532	532	7140.07	549.73	EUR	A	1	
22	4	P	D	160	GRAPHIC EVOLUTION	532	532	7140.07	549.73	EUR	A	1	
23	4	P	D	160	GRAPHIC EVOLUTION	532	532	7140.07	549.73	EUR	A	1	
24	4	P	D	160	GRAPHIC EVOLUTION	532	532	7140.07	549.73	EUR	A	1	
25	4	P	D	160	GRAPHIC EVOLUTION	532	532	7140.07	549.73	EUR	A	1	
26	4	P	D	160	GRAPHIC EVOLUTION	532	532	7140.07	549.73	EUR	A	1	
27	4	P	D	160	GRAPHIC EVOLUTION	532	532	7140.07	549.73	EUR	A	1	
28	4	P	D	160	GRAPHIC EVOLUTION	532	532	7140.07	549.73	EUR	A	1	
29	4	G	D	090	FOURNITURES INDUSTRIELLES DU MIDI	532	532	26450.62	869.38	EUR	A	1	
30	4	G	D	090	FOURNITURES INDUSTRIELLES DU MIDI	532	532	26450.62	869.38	EUR	A	1	
31	4	G	D	090	FOURNITURES INDUSTRIELLES DU MIDI	532	532	26450.62	869.38	EUR	A	1	
32	4	G	D	090	FOURNITURES INDUSTRIELLES DU MIDI	532	532	26450.62	869.38	EUR	A	1	
33	4	G	D	090	FOURNITURES INDUSTRIELLES DU MIDI	532	532	26450.62	869.38	EUR	A	1	
34	4	G	D	090	FOURNITURES INDUSTRIELLES DU MIDI	532	532	26450.62	869.38	EUR	A	1	
35	4	G	D	090	FOURNITURES INDUSTRIELLES DU MIDI	532	532	26450.62	869.38	EUR	A	1	
36	4	G	D	090	FOURNITURES INDUSTRIELLES DU MIDI	532	532	26450.62	869.38	EUR	A	1	
37	4	G	D	090	FOURNITURES INDUSTRIELLES DU MIDI	532	532	26450.62	869.38	EUR	A	1	
38	3	G	D	070	ANIKON	597	597	9000	1000	EUR	A	1	

Source: (Algerian customs)

1.1.1 Prepare data and reduce (Filtering):

In this step we process the data filtering by keeping only the significant data for our work, and erasing the useless data, e.g.: we found clients that doesn't use banks will cause problems during the training period. The most important step is to identify the circuit cell, this is done by renaming it to “fraud” instead of “circuit”, and we fill the cell by the fraud indicators as following:

(1) Represent a suspecting fraudulent event

(2) Represent not clean files

After making the necessary changes, we got 51341 significant data ready to be integrated in the training process.

1.1.2 Regression analysis:

The aim of using regression analysis is to reduce effort and time by eliminating the insignificant factors and just allow the significant once to integrate the training model.

After analyzing the 51000 files during the training period, the program has identified four significant independent variables (Bank, Supplier Country, Producer Country, receipt type).

We got the following regression coefficient results:

The more the coefficient is far from (0) the more the variable is significant.

PS: the next phase does not rely on insignificant variables.

1.2 Processing Algorithms (Training):

In this section, we will present the phases in which different algorithms have been used to process the various dataset operations from the Backend perspective.

Phase one: consist of setting up the proposed program by using PHP machine-learning library called “Rubix_1”, which is easy to manipulate and gives us much control that allows us to realize our objective.

Phase two: we used Python as the main programming language to develop the machine learning program and we used HTML, JavaScript and CSS to create the whole program and to design the application and make it easier to understand and to use.

Phase three: we determine which model to use among the three models we have created. The chosen model is used in the data analysis process, each model uses a different technique (these techniques are defined and explained in the program overview), and the user uses the technique that matches their needs.

Phase four: consist of uploading data from Excel to the machine-learning algorithm, to do this we created another program which aims to convert data from Excel to Json.

Phase five: The Json files are directly integrated into the chosen model till we could process the training, 50k of data are involved in the training and the other 1k are saved for test phase. After the training is complete, the program generate rules that will be an essential indicator for fraud detection.

Phase six: consist on testing the program, we will use the remaining 1k data to test our program efficiency and this is done by doing the same training process with the new data till we got a huge success average. In this step, the program will be ready for service with the possibility of auto-update.

Phase seven: Implementing data visualization algorithm which one of Big Data characteristic, it aims to simplify the results analysis for us in the forms of graphs and histograms.

2 Presentation of the application:

In this part, it is very important to understand the operating principles of our application, so to remedy the application the work revolves around the following points:

- description of the application steps
- Dataset process
- Data filtering
- Explanation of the role of each field in the interface

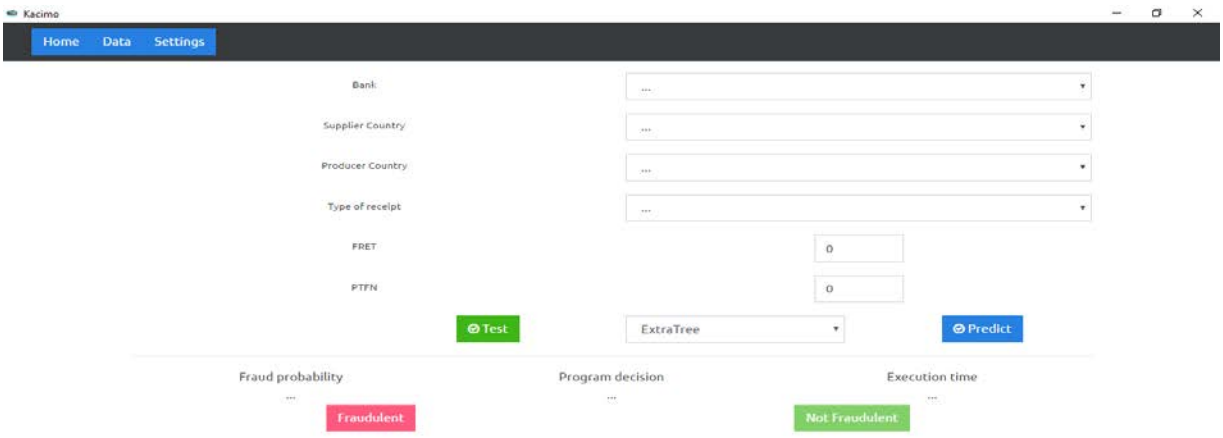
2.1 The aim of the application:

- collect as much as possible data on the various pattern of customs business partners as required
- keep only the useful data by filtering the collected data
- studying the relations between the useful data
- Machine learning logarithm training on the dataset
- Fraud prediction results

2.2 Application Interface :

In this step, we will explain our application interface and the role of each sections and their fields

Figure 13 Application Interface



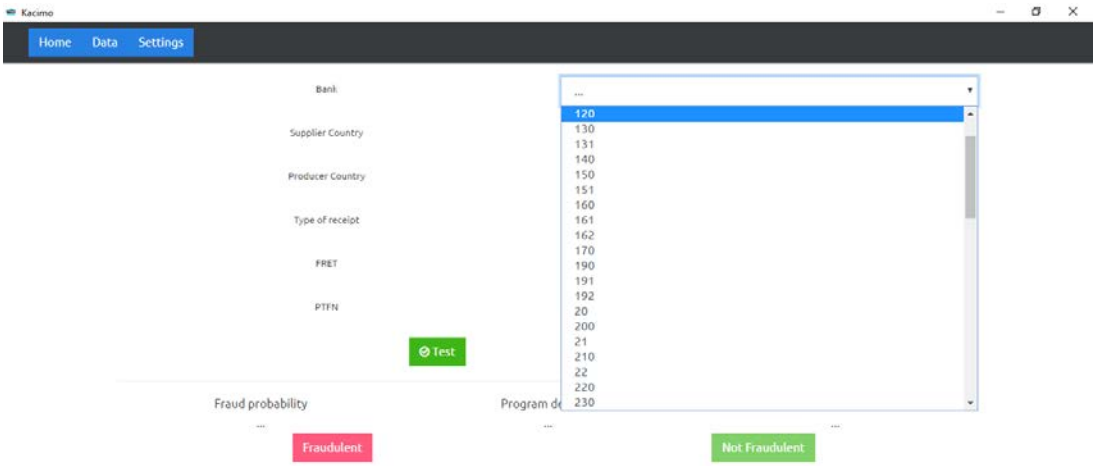
Our interface contain three parts:

- **Part one:** The 1st part “Home” represent the main screen in which we process the fraud prediction via the dataset wish is already maintained in the background of the application. In addition, the main screen “Home” contain different fields where we chose the required information for the fraud prediction Test. These fields represent the useful data that the machine-learning program required them in order to the training process. we will present a demo of how the application work and in the same way an explanation of each field as follow:

➢ **Bank field**: contain all the banks that the customs business partner use and each bank is represented by its unique code.

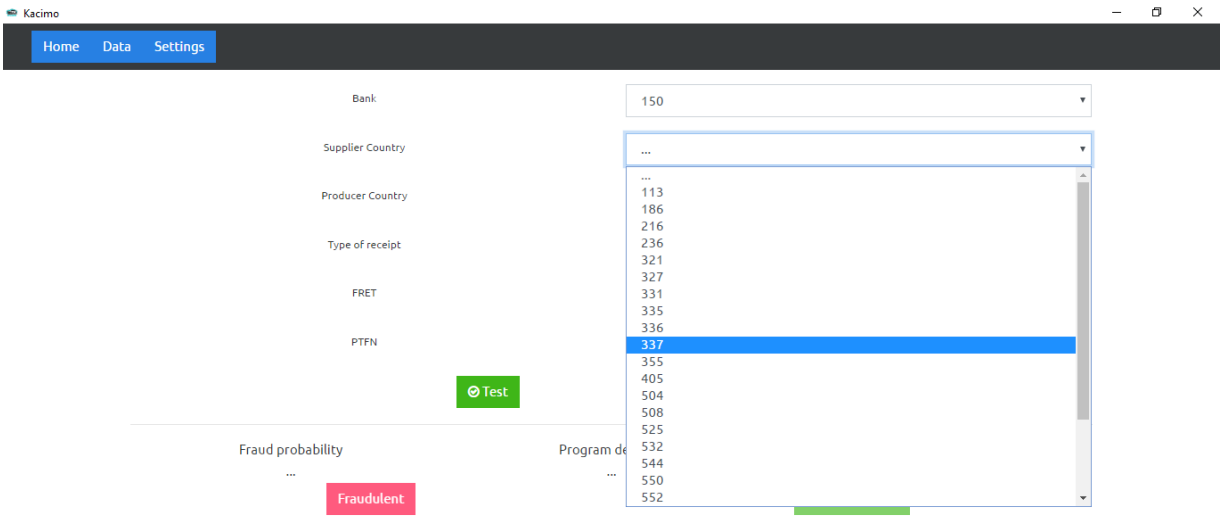
PS: the banks can be local or international.

Figure 14 Bank Selection feild



➢ **Supplier country**: this field contain a list of different supplier's countries where the customs business partners may purchase goods or services and in the same way mentioned above each supplier country is represented by its unique code.

Figure 15 Supplier selection feild

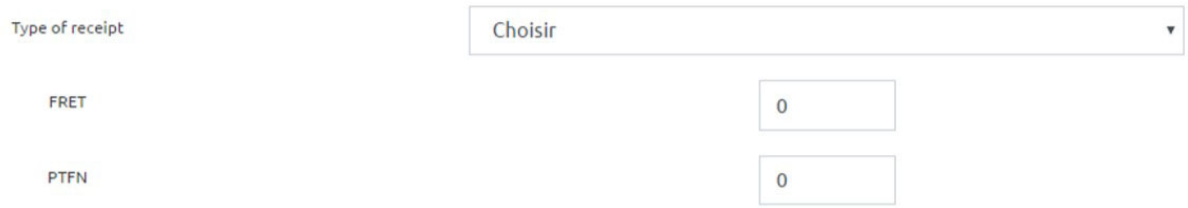


- **Type of receipt:** It is related to incoterms principles that means when the goods delivered are arrived in the countries bounders the customs Agent must make a report about the received goods and gave the buyer's representative a receipt.

- **FRET:** "cost of transporting a shipment of goods by sea". In these fields, we enter an amount related to the fees of supplying goods.

- **PTFN:** "price of goods in foreign currencies"

Figure 16 Receipt & FRET PTFN FEILD



- **Machine-learning algorithms:** There are different types of algorithms that we can use in this application.

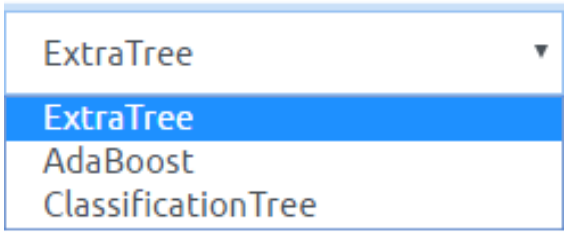
Extra tree algorithm: Extra Trees is an ensemble machine-learning algorithm that combines the predictions from many decision trees. It is related to the widely used random forest algorithm. It can often achieve as good or better performance than the random forest algorithm, although it uses a simpler algorithm to construct the decision trees used as members of the ensemble. It is the most used algorithm among the other's one

Classification tree algorithm: is a supervised algorithm (that is you explain what the input and what the corresponding output is in the training data).

Adaboost algorithm: also known as "meta-learning", an algorithm that use an iterative approach to learn from the mistake of weak classifiers, and turn them into strong one.

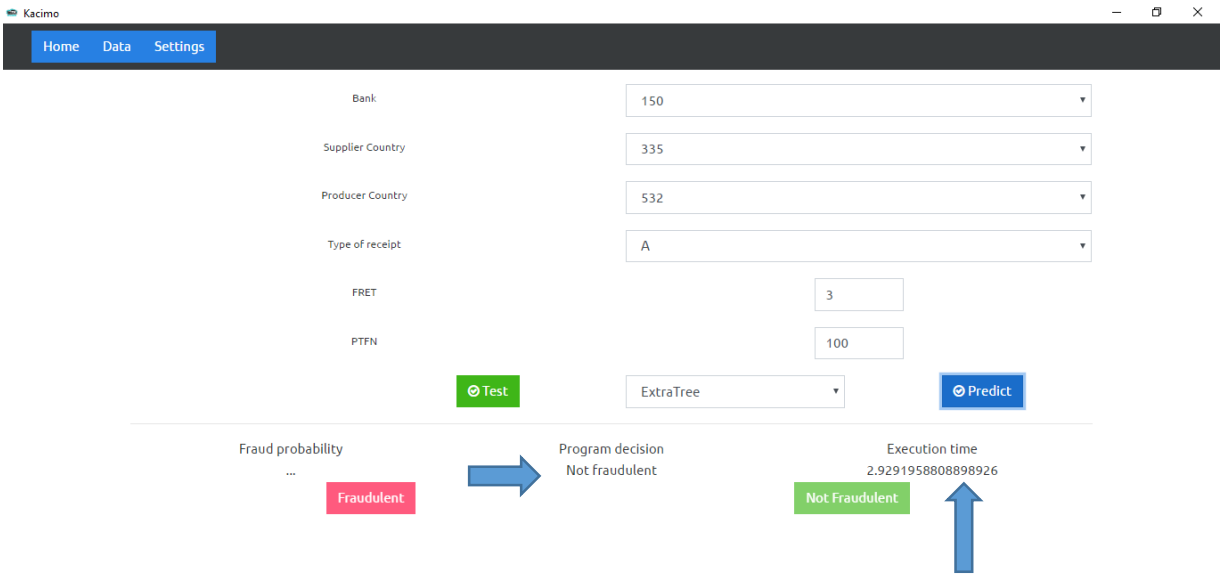
PS: In our demo, we will use the Extra tree algorithm.

Figure 17 Training model option



- **Predict:** in this step and after filling all the required field we will press the predict button, the selected machine-learning algorithm will try to predict the results after making the necessary relations and calculation between dataset variables.

Figure 18 Prediction results



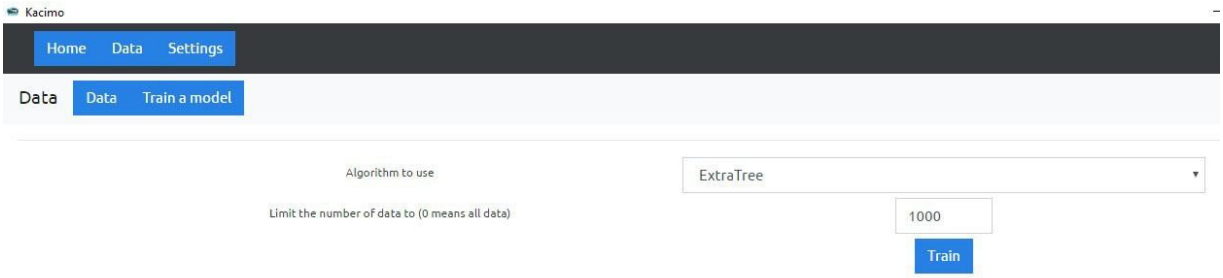
- **Part two :**

The 2nd section "Data" represent a layout page which contain the new results that we get from the 1st part "Home" as we mentioned above that the application already contain 50k dataset in the background which is the basis of the machine-learning training so every new testing results will appear in this section.

Figure 19 Data analyze

In addition this section contain the training field, it consist on adding new dataset and make a request to the machine learning-algorithm for a training for a specific size of data from 10k to 20k. After the model training model is done it will be considered as reference for the future fraud prediction test.

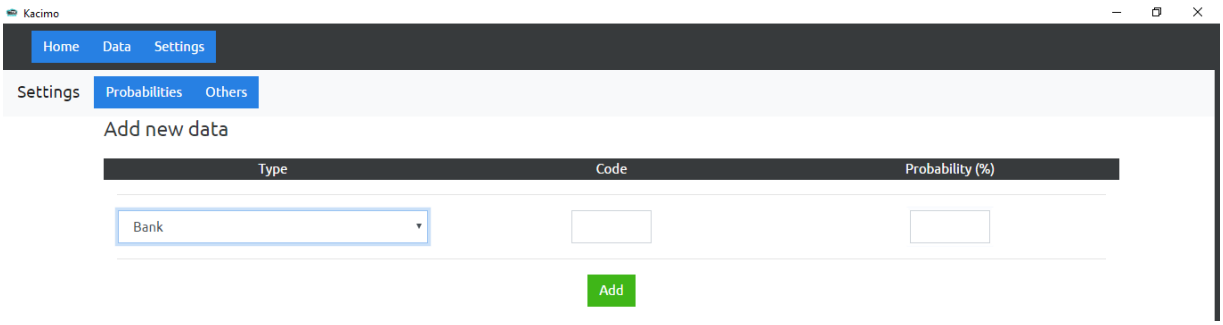
Figure 20 new model training feild



- **Part three**

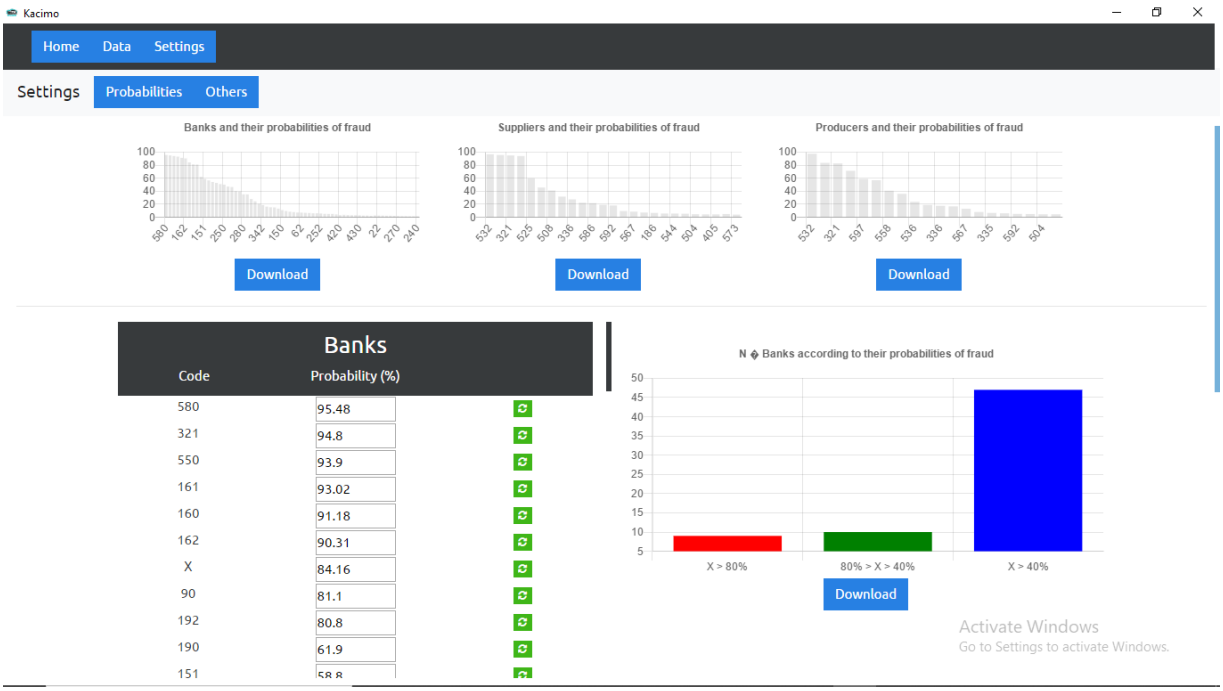
the 3rd section "setting" has different role, the first one consist on the ability of adding a new dataset that doesn't exist in the background of the application, after adding it we should processing a training model as mentioned in the 2nd section "data".

Figure 21 New data feilds



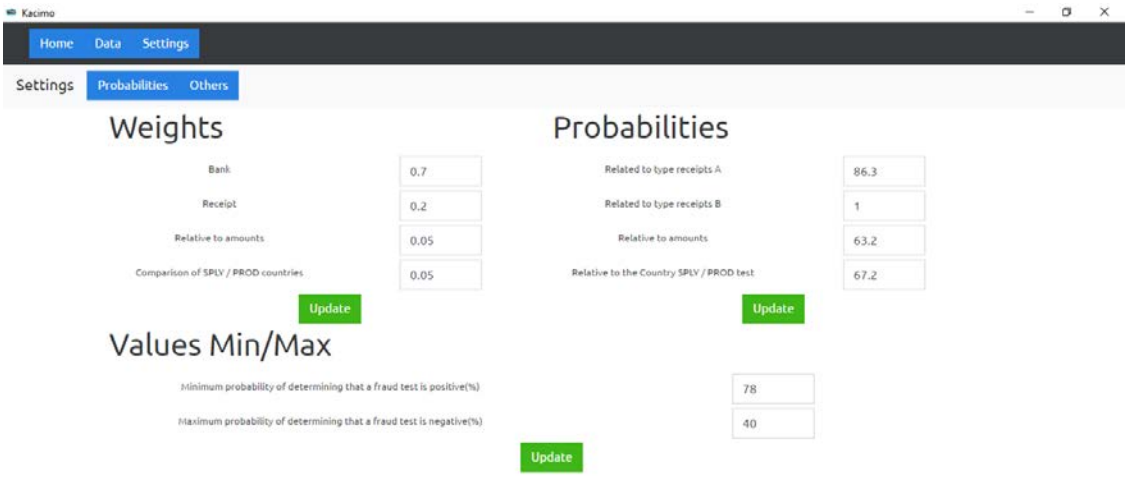
The second features is data visualization, which is one of the famous Big Data characteristic, it consist of the different graphic representation of data. It involves producing image to communicate relationships among the represented data to viewers of the images. This communication is achieved through the use of a systematic mapping between graphic marks and data values in the creation of the visualization.

Figure 22 Data vizualiation



The 3rd part "results" represent the initial results screen in which we can find the weights between each variables and the different probabilities related to each variables for fraud detection. In addition, we find the Max/Min field, which show us the different percentage of the last test that we made.

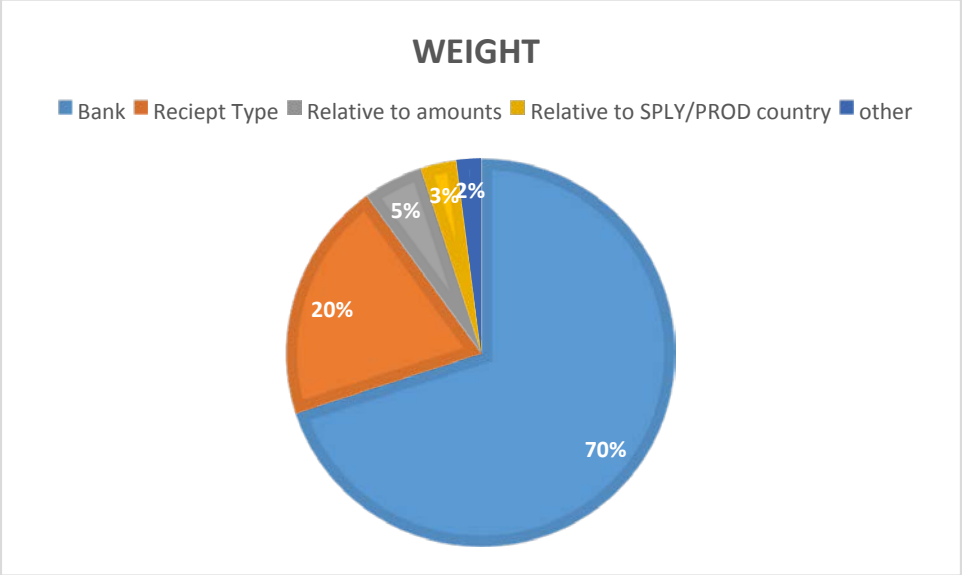
Figure 23 Final results



3 Results Discussion:

3.1 Weight:

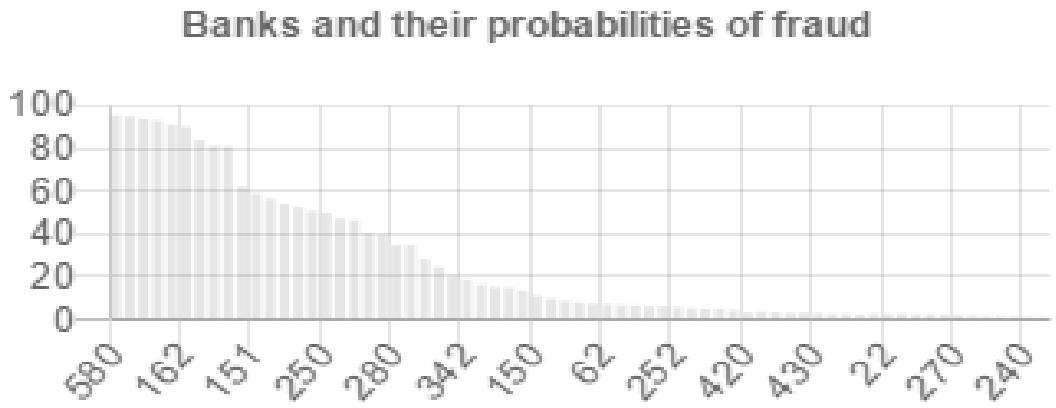
Figure 24 Weight of Independent variables



The weight of the variable represents the severity of its impact on fraud. From the graph we can see that banks represent 70% of fraud weight, which means if the used bank was suspected of fraud, 70% the file will be a risk tolerance. With less effect comes the receipt type (20%), relative to amounts (5%) and relative to Supplier and producer countries with (3%) which are considered as significant in the fraud equation, while the other variables share that last 2% and they are not considered as significant which leads to be eliminated from the equation. This graph represents an effective measure for the risk management department to classify the different fraud variables depending on the importance of each weight in detecting the fraud from different scales.

3.2 Bank Results Analysis:

Figure 25 bank and their probabilities



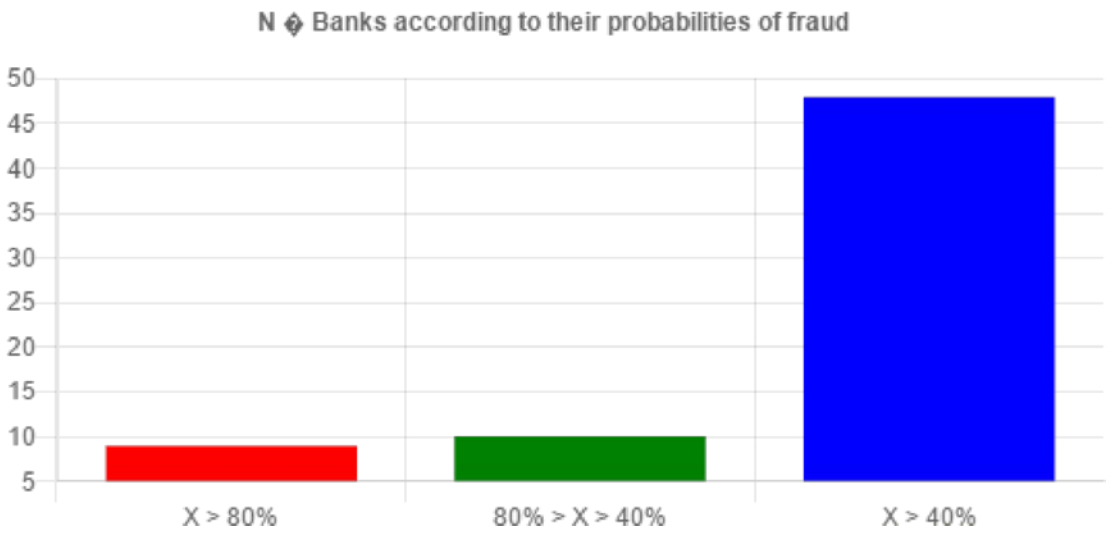
This histogram represents the fraud probabilities in each bank, which was obtained from the program training process that means the knowledge that the machine has acquired after processing 50k of data. From this result, we can remark that there is a considerable variety in bank fraud probabilities; for example, we can find banks like bank n°580 and 162 that has a probability of fraud of 90%, and 70% of the weight of fraud, which leads to considering the file that contains these banks as a high risk. While banks 240 and 270 have not crossed 10% of fraud probabilities and that leads to considering the file unsuspicious.

Since we had a high range of variety and lot of unities, we will categorize them to 3 groups depending on their fraud probabilities and the determined maximum/minimum fraud detection values.

Maximum value of determining fraud is negative = 40% (determined by the machine)

Minimum value of determining fraud is positive = 80% (determined by the machine)

Figure 26 Bank categorization



This histogram represents the results of categorization.

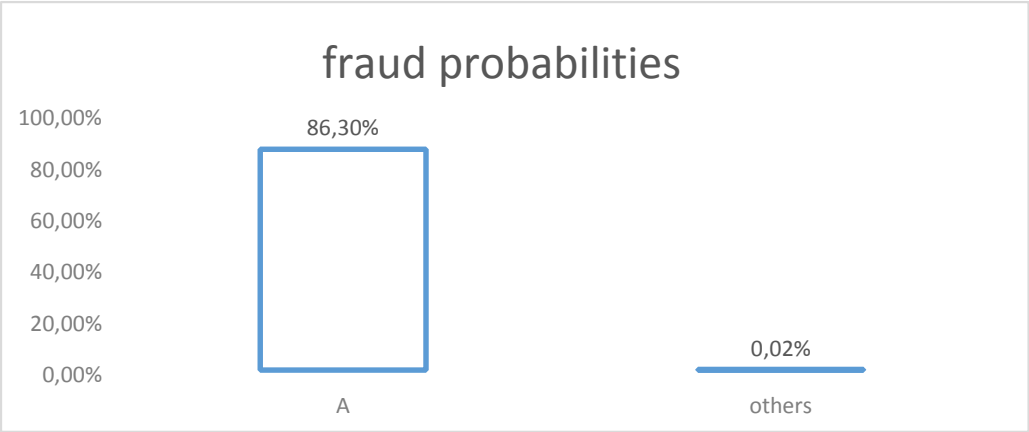
X<40%: The banks which have a fraud probability less than 40% considered unsuspicious so the risk management expert will not be identified as risk which means that no actions are taking against these files.

X>80%: The existence of these banks in the file will lead to consider the file as a risk and the risk management expert will take the process of risk facing and that by stop the packages to be checked and controlled carefully.

40%<X<80%: in this case the file is not considered as a risk nor clean as well so the risk management expert will need more information to trait this file, and these information will be found in the other variables such as receipt type.

3.2.1 Receipt type:

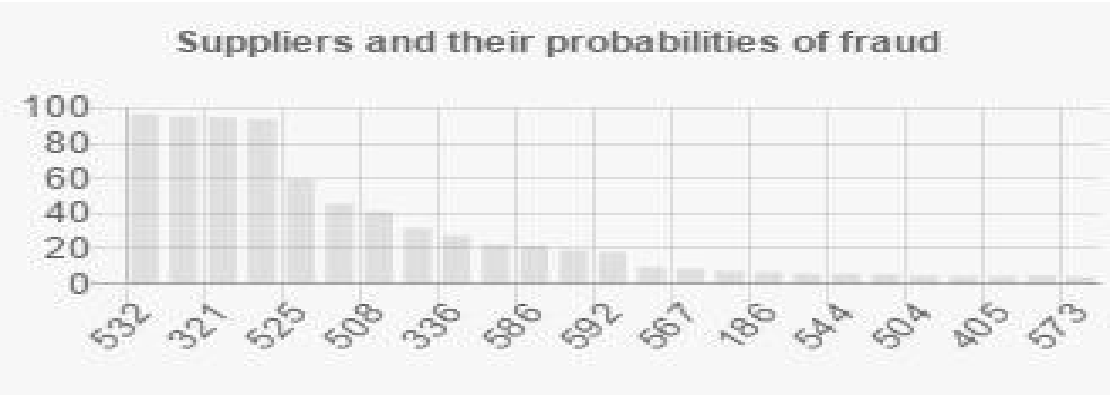
Figure 27 Receaipt type impact analysis



This graph represnts the probabilities of fraud in the different types of receipt used by the clients, as we see the type “A” has a high probbility of fraud while the other types share 0.02%, as we already know that the reciept type has a weight of 20% which means the existance of the type A with 86.3% of fraud probability will effect the clearnace of the file but it still need more information to classify it as a risk.

3.3 Supplier-Country Results Analysis:

Figure 28 fraud probabilities of supplier’s countries

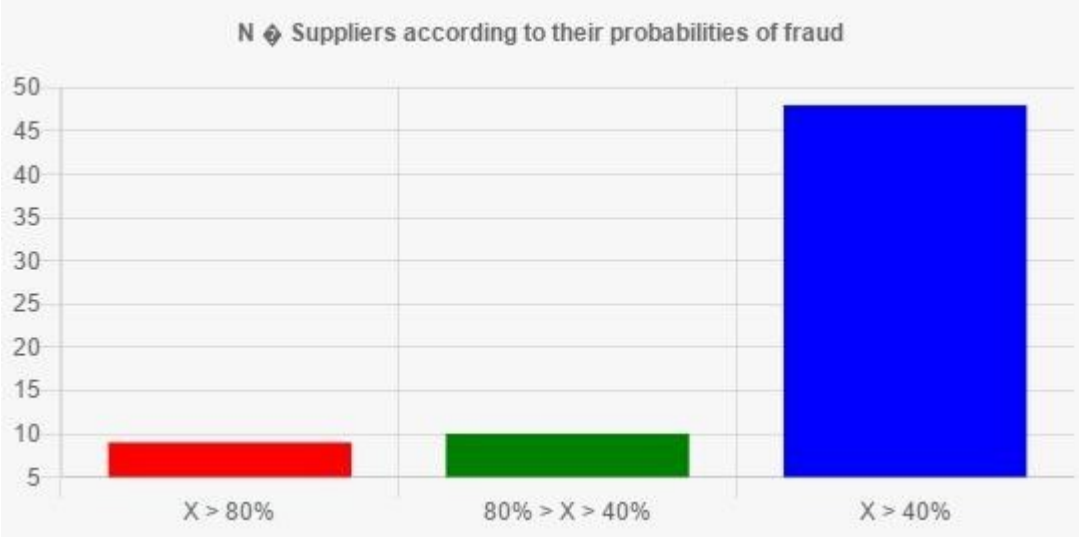


This histogram represents the fraud probabilities for each supplier-country, which mean the country where the goods was supplied. Most of times this case results from the incoterms negotiations between the client and the supplier, which can be different from the suppliercountry. From this result, we can figure that there is a considerable variety in supplier-

country fraud probabilities. We can find suppliers likes supplier-country n°535 and 321 that has a probability of fraud of 92%, and 20% of the weight of fraud, which leads to considering the goods or services received from this country as a risk tolerance, that the organization could not support their damage. While suppliers-country n°573 and 405 have not crossed 10% of fraud probabilities and that leads to considering the goods received from these countries as unsuspicious and also that these countries are already set in the customs database as high-level security boundaries customs.

Since we had a high range of variety, we will categorize them into 3 group according to their fraud probabilities and to the determined maximum/minimum fraud detection values

Figure 29 categorization of suppliers



This histogram represents the result of the categorization of the supplier's countries depending their probabilities we have three categories:

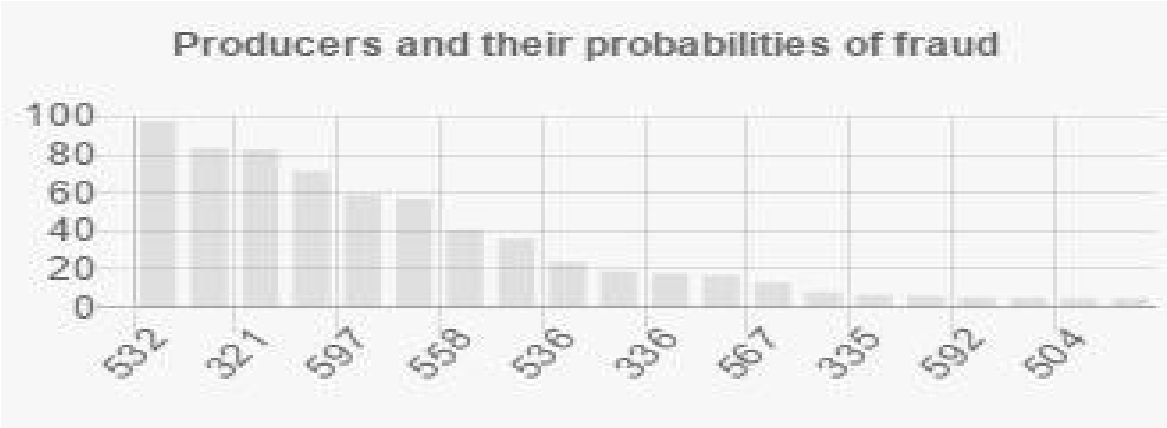
X<40%: Represent the "Target" which means the suppliers-countries which have a fraud probability less than 40% considered unsuspicious so the risk management expert will not be identify it as risk which means that no actions are taking against these files.

X>80%: "The Tolerance" Suppliers-countries belonging to this category will lead to consider the entire file as a risk and the risk management expert will take the process of risk facing and that by stop the packages to be checked and controlled carefully also to get extra information from the supplier country boundaries customs.

40%< X<80%: “Risk Appetite” represent the risks that the risk management department can accept it for a certain period. In addition, this category is not considered as a risk nor clean as well so, the risk management expert will need more information to trait this file and this information.

3.4 Producer Results Analysis:

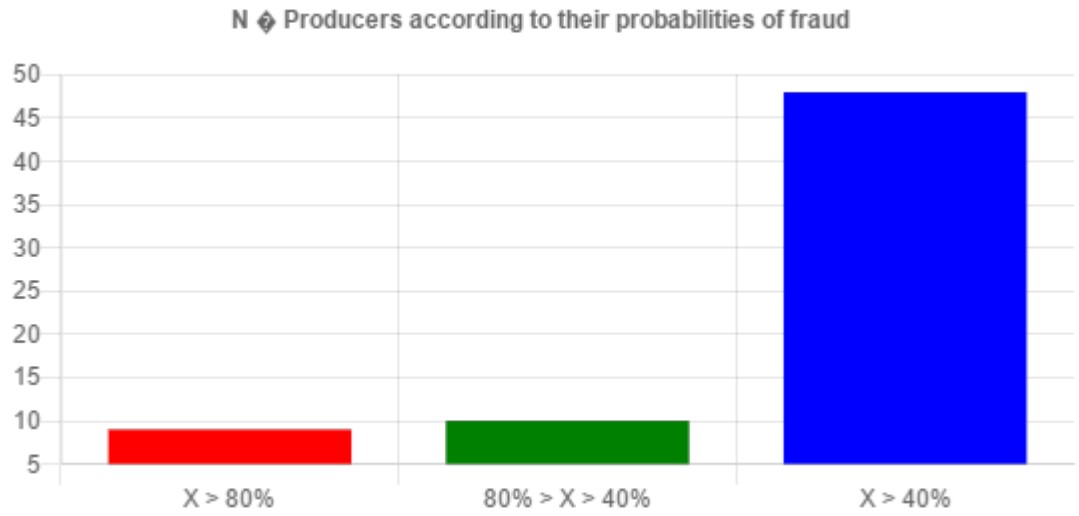
Figure 30 the fraud probabilities of producer’s countries



In this histogram, we find the distribution of producer countries depending their probabilities.

A producer’s country is the country where the merchandise was produces. According to the results, there are countries that most of their incomes found as a fraud like n°532, 321, and 597. Which means that every package comes from these countries will need special measurements, in the other hand we can see that there are safe countries that have insignificant probabilities of fraud. As we know this variable has 3% of fraud weight, which means the treatment of the package, will not be necessary but if there still a confusing the risk management expert will need to treat this variable in this case we will classify the producer’s country in three categories the same as we did with supplier’s countries.

Figure 31 the result of categorization of producer's countries

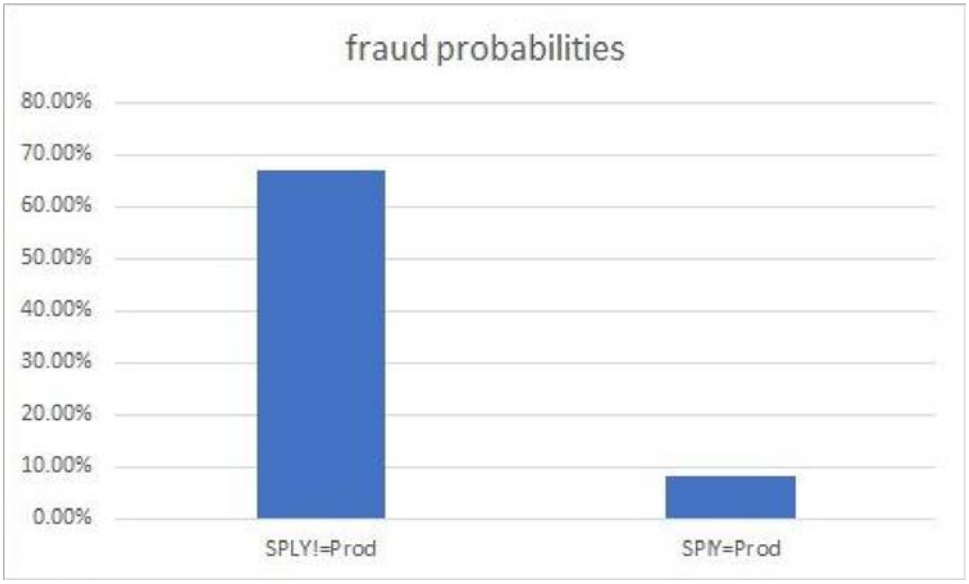


This histogram represents the results of the categorization of countries of producers by fraud probability, as we can see the results here are similar to the results in the supplier's countries case so we will trait this results the same way we did with the results of categorization of the countries of suppliers:

X<40%: the producer's countries which have a fraud probability less than 40% considered unsuspicious so the risk management expert will classify them as "Target" as which means that there is no risk and no actions are taking against these files.

X>80%: this category called "The Tolerance" it is considered as a risk and the risk management expert will take the process of risk facing and that by stop the packages to be checked and controlled carefully also to get extra information from the supplier country boundaries customs.
40%< X<80%: "Risk Appetite" represent the risks that the risk management department can accept it with minimum preventative measurements. In addition, this category is not considered as a risk nor clean as well.

Figure 32 probabilities of fraud when the supplier's country and the producer's country are not the same



This histogram contains the probability of fraud when the when the merchandise comes from the same country where it was produced. In these results we see words when the producer and the supplier are from the same country and the fraud probability is very low in the contrary which means, when they are from different countries the fraud probability is very high, and because this variable has a very low weight in the fraud equation. This result will be taking in consideration by the risk management expert in very rare cases.

Ps: as we mentioned before the result will be taking in consideration the weight. In another world if the variable with higher weight was classified, the whole file will be classified as a risk no matter was the result of the other variables. In this case the result of detecting fraud is high but not as good if we use all the variables in one equation even the one with a very low weight because that will improve the precision of the results. Using all the variables will be almost impossible without using big data analysis techniques, and this is what our program can offer.

CONCLUSIONS & RECOMMENDATIONS

This thesis aimed to find a way how to make use of big data tools and technics in the fraud detection process, and see the consequences of that on all the factors that have a relation with the fraud detection process as much as customs, their clients, risk management department... etc.

The first step was to define and understand the notions related to the main theme of the thesis as big data, fraud, and risk management. Also we explained the fraud detection process and which department is responsible for that in customs, in order to know exactly where we will work, and with whom, plus have more comprehension how they work to identify their weaknesses and try to develop it in our model.

After understanding, the concepts that we will work on, we were able to identify the questions that will help us to set up the guidelines work plan that can lead us to achieve our objectives. The first question was about the fraud identification, to know how the operation works till we be able to figure which big data technic that fit the process the most, secondly found a way how to implement it in the process, in the mean time we wanted to look into the reasons and the variables that impact fraud the most.

The second step was to choose the most suitable big data technic and how can we make it fit the needs of the Algerian customs. After studying many techniques available in big data and taking into consideration the characteristics and quality of information available at the customs level, we found that the machine learning technology is the best technic that will lead us to achieve our goals: detect and predict fraud in the best possible way. In addition, we will be able to define the tools and the sources the most used in fraud, in order to do that in optimized way we decided to visualize the data to be able to analyze in an easier way.

The result of this second step is a software built on a machine-learning algorithm that can detect and predict fraud using statistical mechanisms to calculated and analyze all the existed relation between the several variables with fraud. To setup the rules that we will train on them using historical data to build an artificial knowledge that will add more precision to the detection fraud process and that was proven once we tested the prototype we made.

Building that prototype proved also that big data can be so very useful in the fraud detecting process. It will be able to offer more automation in the customs level besides adding more precision in less time, it guarantees a continuity of improvements after each utilization, and the same time it limits the human intervention and easy to use.

Depending the results from the rules that had been extracted from the customs database we found that there are variables that affect fraud more than the others, which means fraudulent business partner use this variables more than the others. We found that there are banks that have

a very high percentage of being used for fraud that can exceed 90%. same as much as the country origin of the goods, the origin of the supplier, the type of receipt and so on, depending those results we can build an importation/exportation strategy that will eliminates the high suspicious variables if it is possible or at least minimize them .

Otherwise, during our work on this research, we encountered many problems, similar to any other work, especially in this exceptional year. We can divide these limits to two types, the first type is research limitations which was the access for the organization were it was denied to the host organization because of COVID-19, and that stopped us from learning more about how customs do their work, how do they detect risk, what strategies used by the risk department to face it.

The second type is the methodological Limitations, which were represented in the lack of available and/or reliable data; we could not have access to customs database that contains the fraudulent and the non-fraudulent files, so we used the data that they have given to us to create a simulation.

The results of our paper are not the end of the research work in the contrary it opens the door to researchers to start considering using big data technologies that have proved its efficacity in other domains, and start asking questions like, how can big data affect decision-taking? How to improve business using big data techniques? What is the difference between a strategy based on big data and other strategies? In addition, what makes big data good for business? The questions we can ask and the researches we can do about this theme are infinite all we need is to have is to change the angle view for business.

BIBLIOGRAPHY

- Arena, M., Arnaboldi, M., Azzone, G. (2010), The organizational dynamics of enterprise risk management. Accounting, Organizations and Society, 35, 659-675.

- Banarescu, A. (2015), Detecting and preventing fraud with data analytics. Procedia Economics and Finance, 32, 1827-1836

- Beasley, M.S., Pagach, D., Warr, R. (2008), Information conveyed in hiring announcements of senior executives overseeing enterprise-wide risk management processes. Journal of Accounting, Auditing and Finance, 23(3), 311-332.

- Bhimani, A. (2009), Risk management, corporate governance and management accounting: Emerging interdependencies. Management Accounting Research, 20, 2-5.

- Bolton, R & Hand, D 2002. 'Statistical Fraud Detection: A Review (With Discussion)', Statistical

- Boscal, K.H.Y., Lai, I.K.W., Chan, S.K.C. (2010), Supply Chain Risk Management Model ERM Approach. 8th International Conference on Supply Chain Management and Information.

- Chen, J., Tao, Y., Wang, H., Chen, T. (2015), Big data base fraud risk management at Alibaba. The Journal of Finance and Data Science, 1, 1-10

- Choi, T.M., Lambert, J.H. (2017), Advances in risk analysis with big data. Risk Analysis, 37(8), 1435-1442.

- CIRREL (Interuniversity Research Center on Enterprises Networks, Logistics and transportation), George Dionne, Risk Management, P: 06

- Clemmons, D. 2007, 'The never-ending fight against fraud', Internal Auditor, Dec, 2007

- De Loach, J.W. (2000), Enterprise-wide risk management. London: Financial Times-Prentice Hall.

- elements of senior executives overseeing enterprise-wide risk management processes. Journal of Accounting, Auditing and Finance, 23(3), 311-332.

- Elgendy, N., Elragal, A. (2014), Big Data Analytics: A Literature Review Paper. Switzerland: International Publishing. p214-227.

- Ellul, A., Yerramilli, V. (2013), Stronger risk controls, lower risk: Evidence from U.S. bank holding companies. The Journal of Finance, 68(5), 1757-1803.

- Florio, C., Leoni, G. (2017), Enterprise risk management and firm performance: The Italian case. The British Accounting Review, 49(1), 56-74.

- Gandomi, A., Haider, M. (2015), Beyond the hype: Big data concepts, methods and analytics. International Journal of Information Management, 35, 137-144.

- Gordon, L.A., Loeb, M.P., Tseng, C.Y. (2009), Enterprise risk management and firm performance: A contigency perspective. Journal of Accounting Public Policy, 28, 301-327.

- Grace, M.F., Leverty, J., Phillips, R., Shimpi, P. (2015), The value of investing in enterprise risk management. Journal of Risk and Insurance, 82(2), 289-316

- Hasnat, B. (2018), Big data: An institutional perspective on opportunities and challenges. Journal of Economic Issue, 580-588.

- Hossein, H., Xu, H., Emmanuel, S.S. (2018), Digitalisation and big data mining in banking. Big Data and Cognitive Computing, 2(3), 18.

- Hu, D., Zhao, J.L., Hua, Z., Wong, M.C. (2012), Network-based modelling and analysis of systemic risk in banking system. MIS Quarterly, 36(4), 1269-1291.

- IBM. (2014), Operational Risk Management in the World of Big Data. IBM Software. Business Analytics. United States: IBM. p1-12.

- Idris, A., Norlida, A.M. (2016), Influence of enterprise risk management success factors on firm financial and non-financial performance: A proposed model. International Journal of Economics and Financial Issues 830-836.

- Intal T. & Do L.T. 2003, 'Financial Statement Fraud - Recognition of revenue and the auditor's responsibility for detecting financial statement fraud', Master Thesis, Goteborg University, Sweden.

- Krishna, D. (2016), Big data in risk management. Journal of Risk Management in Financial Institutions, 9(1), 46-52.

- Lackovic, D.I., Kovska, V., Lakovic, V.Z. (2016), Framework for Big Data Usage. Risk Management Process in Banking Institutions. Central European Conference on Information and Intelligent System. p49-54.

- Liebenberg, A.P., Hoyt, R.E. (2003), The determinants of enterprise risk management: Evidence from the appointment of chief risk officers. Risk Management and Insurance Review, 6(1), 37-52.

- Liebenberg, A.P., Hoyt, R.E. (2011), The value of enterprise risk management. The Journal of Risk and Insurance, 78(4), 795-822

- Mohamad, S.H., Rashila, R., Marwan, Y.I.M., Azzam, I.R. (2015), Reputation risk and its impact on the Islamic banks: Case of the Murabaha. International Journal of Economics and Financial Issues, 5(4), 854-859.

- Muller, O., Fay, M., Vom Broke, J. (2018), The effect of big data and analytics on firm performance; an econometric analysis considering industry characteristics. Journal of Management Information System, 35(2), 488-509.

- Munesh, K., Mittal, P. (2014), Big data: A review. International Journal of Computer Science and Mobile Computing, 3(7), 106-110.

- Navak, N., Akkiraju, R. (2012), Knowledge Driven Enterprise Risk Management. California: Annual SRII Global Conference.

- Nocco, W.B., René, M.S. (2006), Enterprise risk management: Theory and practice. Journal of Applied Corporate Finance, 18(4), 8-20.

- Ozkose, H., Ari, E.S., Gencer, C. (2015), Yesterday, today and tomorrow of big data. Procedia Social and Behavioural Sciences, 195, 1042-1050.

- Ravisankar, P., Ravi, P., Raghava, R.G., Bose, I. (2011), Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems, 50(2), 491-500

- Saggi, M.K., Jain, F. (2018), A survey towards an integration of big data analytics to big insights for value-creation. Information Processing and Management, 54, 758-790.

- Sagiroglu, S., Sinanc, D. (2013), Big Data: A Review. San Diego, CA, USA: International Conference on Collaboration Technologies and Systems. p42-47.

- Srivastava, U., Gopalkrishnan, S. (2015), Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks. Procedia Computer Service, No. 50. 2nd International Symposium on Big Data and Cloud Computing, 50, 643-652.

- Woods, M. (2007), Linking risk management to strategic controls: A case study of Tesco plc. International Journal of Risk Assessment and Management, 7, 1074-1088.

- ACFE 2008, 2008 Report to the Nation on occupational fraud & Abuse, US, www.acfe.com

- Albrecht, W.S., Albrecht C.C. & Albrecht C.O. 2006, Fraud Examination, 2nd ed, Mason, Ohio, South-Western.

- Barry Klein, (2001), The word's first company, https://www.irmi.com/articles/expert-commentary/the-worlds-first-insurance-company date of consultant 16/06/2020.

- CFI, (2015), uncertainty https://corporatefinanceinstitute.com/resources/knowledge/other/uncertainty, consultation date: 06-16-2020.

- Dale F. Cooper, Stephen Grey, Geoffrey Raymond, Phil Walker, Project Risk Management Guidelines Managing Risk in Large Projects and Complex Procurements (2004), P: 3; 126
- Encyclopedia of Science, Technology, and Ethics (2020), SOFT SYSTEMS METHODOLOGYhttps://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/soft-systems-methodology Date of consultation: 6/28/2020 at 10:15 P.M
- INTERNATIONAL STANDARD CEI IEC 61508-4 (2006), P:18.
- KPMG 2006, 2006 Survey of Fraud in Australia and New Zealand, Sydney.
- Lanza R.B. 2004b, 'How to Use a New Computer Audit Fraud Prevention and Detection Tool Information Systems Audit and Control Association, Volume 1, 2004, www.isaca.org
- Osama Azmi Sallam, Shukairi Nuri Musa, Risk and Insurance Management, Hamed Publishing and Distribution House, 1st Edition, Amman-Jordan, 2007, p: 55.
- Practice Standard for Project Risk Management by PMI (2009), P:13

Science, 17(3): 235-255.

- Smith, N.J. (1999). Managing Risk in Construction Projects. Blackwell Science, P: 66-67
- Tony Merna and Faisal Al-Thani, Corporate Risk Management (2008), 2nd Edition, P: 68-69
- Webster (2001) Webster's New World College Dictionary. 4th ed., Cleveland, IDG Books Worldwide
- www.minitab.com date of consultation 06/18/2020
- *Xu, Wei*, *Chen, Yuehuan*, *Coleman, Conrad*, *Coleman, Thomas F* (2017), Moment matching machine learning methods for risk management of large variable annuity portfolios