NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Implementation of an Effective and Efficient Anti-Money Laundering & Counter Terrorism Financing System

The Adoption of a Behaviorally View

David José Santos Guerra

Dissertation / Project Work / Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2019

Title: Implementation of an Effective and efficient Anti-Money Laundering & Counter Terrorism Financing System

Subtitle: The Adoption of a Behaviorally View

Student David Guerra
full name David José Santos Guerra

MAA

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INTERNSHIP REPORT – IMPLEMENTATION OF AN EFFECTIVE AND EFFICIENT ANTI-MONEY LAUNDERING & COUNTER TERRORISM FINANCING SYSTEM

by

David José Santos Guerra

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

Advisor: Prof. Dr. Rui Gonçalves

Co Advisor*:* Dr. Mário Neves

02 2019

DEDICATION

This thesis is wholeheartedly dedicated to my beloved parents, who always pushed me forward and encouraged me to follow my ideals, providing me with the security I required in order to dive into the challenges that I deemed worthy. To my family, that always had a word of motivation and advise, helping me in this pursuit. To my mentors, that always made room for my questions regardless of their busy schedule. To my colleagues, both academic and professional, that were always ready to help. To my closest friends, for always being there. And lastly, to all that provided me the chance to be here.

ACKNOWLEDGEMENTS

I would like to acknowledge the financial institution, office and department whereas I was integrated, for making this thesis possible, allowing me to integrate such amazing team and relevant project, and the time and comprehension given for the concretization of this thesis. In particular, I would like to thank Eng. António Félix and Dr. Raimundo Martins for introducing me to the professional workflow in an extraordinary way and always backing me up in the writing of this thesis, regardless of the work situation; Dr. Mário Neves, for always being available to answer my questions and providing useful insights; and Prof. Rui Gonçalves for being with me in every step of the way of the development of this thesis.

ABSTRACT

Money laundering and the financing of terrorism has been and is increasingly becoming an even greater concern for governments and for the operating institutions. The present document explores the methodologies and approaches to be partaken in order to achieve an implementation of an effective and efficient anti-money laundering & counter terrorism financing system (AML/CFT system), in regards to a behaviorally view on the European banking context. To do so three main requirements were delimited: the legal, that states that there must exist a compliance with the legal authorities' requirements and proposals; the adaptive, that denotes the necessity of a high adaptability on the system's capability to adjust to the client's behavior; and the continuous fine-tuning, implying the importance of the capability to do a continuous fine-tune to the system so that it can achieve its most optimal efficiency, along its life cycle. A data exploration and analysis was made to the bank's customer database, to uncover discrepancies and inconsistences, and to further provide solutions to correct such situations in preparation for the AML/CFT system implementation. The environment whereas the system will be integrated was analyzed, assessing the impact of such implementation in the pre-existing systems' ecosystem (market abuse, transaction and account opening validation, and AML/CFT thematic), and utilizing the parallelism between the applications to extrapolate key points, at both the logical and technical level. At the technical level a set of procedures were made, regarding the data mapping, the staging areas' creation (ETL processes), the data structure and organization delimitation, the jobs definition, and the output and interface design (query, reporting and ML/FT analysis). At the logical level, the behavior evaluation was modelled (creation, validation and implementation of customer due diligence rules (CDDs), risk factors and ML/FT's scenarios), the suspicious behavior's analysis was defined, the processes' validation delimitated, and a population segmentation and classification was made. As to assure that a compliance with the legal requirements and the expected functioning of the operations was verified, a delimitation of internal controls was applied. Having that the system was to be implemented throughout multiple geographies, a multitenancy capability was developed, replicating the system's technical and logical view, and making the appropriate changes in order to adjust to each of the geography's context, without altering the core of the system. As of this implementation, the newly implemented system demonstrated an alignment with the best practices; coverage with about 90% of the standards; a high efficiency, assertiveness and customizability; and an alert generation alignment. This translated into an AML/CFT system, that is considered fully functional, meeting all the imposed requirements and being efficient; that validates the transactional behavior as a whole, being more adjusted; and that possesses a high level of easiness on the assessment and optimization, being that in this regard a set of future procedures to be implemented were delimitated, reflecting on the population validation and optimization, scenarios validation, threshold optimization, risk factors validation and CDDs validation.

KEYWORDS

AML/CFT; Analysis; Segmentation; Clustering; Customer Due Diligence; Risk Factor; Machine Learning; Compliance;

INDEX

1. Introduction 1
   1.1. Internship Objectives 2
   1.2. Organization Characterization 2
      1.2.1. The Banking Institution 2
      1.2.2. Compliance Office 2
      1.2.3. Operations and Processes Monitoring Department 3
      1.2.4. AML/Counter Terrorism Monitoring Department 3
      1.2.5. Information Systems and Analytics Department 3
2. Literature review 4
   2.1. Business/Concepts Review 4
      2.1.1. Money Laundering 4
      2.1.2. Financing of Terrorism 4
      2.1.3. AML/CTF Legislations 5
      2.1.4. AML/CFT System 7
      2.1.5. RBA 10
      2.1.6. Regulatory Entities 12
   2.2. Technique Review 13
      2.2.1. Classification & Regression 13
      2.2.2. Clustering 14
      2.2.3. Similarity Criterion 15
      2.2.4. Hierarchical Clustering 19
      2.2.5. Partitional Clustering 21
      2.2.6. Validation Measures 21
      2.2.7. K-Means 26
      2.2.8. K-Means ++ 29
      2.2.9. K-Medoid 30
      2.2.10. BIRCH 30
      2.2.11. CURE 33
      2.2.12. RFM 34
      2.2.13. Decision Tree 35
      2.2.14. Ensemble 38
      2.2.15. Artificial Neural Networks 40

2.2.16. Association Rules 43
2.2.17. Algorithms Final Intakes 44
3. Methodology/Idea 46
4. Project 'Environment' 47
4.1. Project Objectives 48
4.2. Project Technical View 49
4.3. Project Logical View 50
4.4. Internal Control 51
4.5. Project Multitenancy 52
4.6. Project Implementation Plan 52
5. Description Of The Tools 55
5.1. SQL Server Management Studio 55
5.2. Microsoft Excel 55
5.3. R 55
5.4. SAS® Software 55
5.4.1. SAS® Compliance Solution 55
5.4.2. SAS® Visual Analytics 55
5.4.3. SAS® Enterprise Guide 56
6. Data And Experimental Settings – Implementation Phase 57
6.1. Parallelism Between Systems 57
6.2. Data Exploring 57
6.3. Data Parallelism 61
6.4. Data Mapping 61
6.5. Populations/Segments 62
6.6. Scenarios 63
6.7. Risk Factors 72
6.8. Thresholds 76
6.9. CDDs 76
6.10. Practical Usage of The System 82
6.11. Reporting (VA) 84
6.12. System's Assumptions/Restrictions 86
7. Ongoing Phases 88
7.1. Populations Validation 88
7.2. Populations Optimization 88
7.3. Scenario Validation 89

7.4. Threshold Optimization 89
7.5. Risk Factors Validation 89
7.6. CDDs Validation 89
8. Conclusions 91
8.1. Feedback 91
8.2. Final Statements 93
9. Future works 94
10. Bibliography 95
10.1. Papers & Books 95
10.2. Websites 97
11. Annexes 99

LIST OF FIGURES

Figure 1 - Compliance Office Constitution (Source: Author) 3
Figure 2 - X, Z and Y representation in a bi-dimensional space (Source: Author) 15
Figure 3 - Manhattan Distance Example (Source: Author) 16
Figure 4 - Impact of the Mahalanobis Distance (Source) 17
Figure 5 - Single Linkage Example (Source: Author) 18
Figure 6 - Complete Linkage Example (Source: Author) 19
Figure 7 - Average Linkage Example (Source: Author) 19
Figure 8 - Centroid Linkage Example (Source: Author) 19
Figure 9 - Dendrogram example (Source: Author) 21
Figure 10 - K-Means operational example (Source: Author) 27
Figure 11 - Elbow graph example (Source: Author) 28
Figure 12 - Same dataset, different initializations, different results, example (Source) 29
Figure 13 - Decision Tree (Source: Author) 36
Figure 14 - Decision Tree Incorrect Structure (Source: Author) 37
Figure 15 - Artificial Neural Network Schema (Source: Author) 41
Figure 16 - Processment of the system's information (Source: Author, based on SAS documentation (2018)) 50
Figure 17 - System's Multitenancy (Source: Author, based on SAS documentation (2018)) 52
Figure 18 - Implementation Timeline (Source: SAS Documentation (2018)) 53
Figure 19 - Example of a threshold distribution (Source: Author) 60
Figure 20 - Scenario Distribution by Monetary Type (Source: Author) 69
Figure 21 - Scenario Distribution by AML/CTF Areas (Source: Author) 70
Figure 22 - Scenario Distribution by Monetary Flux (Source: Author) 70
Figure 23 - Scenario Distribution by Intervenient (Source: Author) 71
Figure 24 - Risk Factor Distribution by Evaluated Data Nature (Source: Author) 74
Figure 25 - Risk Factor Distribution by Evaluated Risk Areas (Source: Author) 75
Figure 26 - Risk Factor Distribution by Intervenient (Source: Author) 75
Figure 27 - CDD Rule Distribution by Attribute Category (Source: Author) 79
Figure 28 - CDD Rule Distribution by Data Nature (Source: Author) 80
Figure 29 - CDD Rule Distribution by Intervenient (Source: Author) 81
Figure 30 - System's Operation Flowchart (Source: SAS documentation) 82

Figure 31 - Dashboard of the generated and analyzed alerts (Source: SAS-Bank's Solution) . 85

LIST OF TABLES

Table 1 - Money laundering simplified three step process 4
Table 2 - Preventive duties of the bounded organizations 7
Table 3 - Objectives related to the Intermediate objective 'A' 8
Table 4 - Objectives related to the Intermediate objective 'B' 8
Table 5 - Objectives related to the Intermediate objective 'C' 9
Table 6 - RBA as four stage process 10
Table 7 - True positives, false positives, false negative, true negative definition 24
Table 8 - Four stage building phase 54
Table 9 - Old Population Distribution 58
Table 10 - Portugal's clients' segmentation 62
Table 11 - Macau's clients' segmentation 62
Table 12 – Bank's employees' segmentation 63
Table 13 - Scenario logical classification 72
Table 14 - Risk Factor logical classification 76
Table 15 - CDD logical classification 81

LIST OF FORMULAS

Equation 1 - Symmetry property 15
Equation 2 - Identification of the indiscernible 15
Equation 3 - Non-Negativity property 15
Equation 4 - Triangle inequality property 15
Equation 5 - Euclidean distance formula 16
Equation 6 - Manhattan distance formula 16
Equation 7 - Minkowski distance formula 17
Equation 8 - Mahalanobis distance formula 18
Equation 9 - Mahalanobis distance formula, for each observation 18
Equation 10 - Euclidean distance weighed formula 18
Equation 11 - Bayesian information criterion index formula 22
Equation 12 - Calinski-Harabazsz index formula 22
Equation 13 - Davies-Bouldin index formula 22
Equation 14 - Silhouette index formula 23
Equation 15 - Dunn index formula 23
Equation 16 - Recall and Precision formulas 24
Equation 17 - Recall and Precision formula (using true positives, false positives, true negatives and false negatives) 24
Equation 18 - F-Measure formula 25
Equation 19 - Normalized mutual information formula 25
Equation 20 - Individual and overall Purity formulas 25
Equation 21 - 'Individual' and 'Overall' Entropy formulas 25
Equation 22 - K-means ++ seed selection probability 30
Equation 23 - BIRCH's Merging Process 31
Equation 24 - Centroid and Radius (Sub-cluster) 32
Equation 25 - Cluster Radius 32
Equation 26 - Size of sample conditions 34
Equation 27 - SOM Learning Expression 43
Equation 28 - Example of Neighborhood Function 43
Equation 29 - Worse time complexity of CURE 44
Equation 30 - Structuring Factor formula 64

Equation 31 - Velocity Factor formula 72

LIST OF ABBREVIATIONS AND ACRONYMS

.db	Database
ACAMS	Association of certified Anti-Money Laundering Specialists
AGP	Alert Generation Process
AML/CFT	Anti-money Laundering and Counter Financing of Terrorism
Art	Article
ATM	Automated Teller Machine
BI	Business Intelligence
BIRCH	Balanced iterative Reducing Clustering Hierarchies
BMU	Best Mapping Unit
BP	Banco de Portugal
CDD	Customer Due Diligence
CF	Cluster Features
CGV	Golden Visa Clients
CMVM	Comissão de Mercados de valores Mobiliários
CORP	Corporation
CTR	Currency Transaction Reporting
CURE	Clustering using REpresentatives
DFA	Disaster Financial Assistance
DNFBPs	Designated Non-Financial Business and Professions
EDD	Enchased Due Diligence
EFTA	European Free Trade Association
ESA	EFTA Surveillance Authority
ETL	Extract, Transform, Load
EU	European Union
FAFT	Financial Action Task Force
FCA	Financial Conduct Authority
KYC	Know Your Customer
MLP	Multiple Layer Perceptron
MLRO	Money Laundering Reporting Officer
MSB	Money Service Business
NPO	Non-Profit Organization
PB	Private Banking
PEP	Politically Exposed Person
RBA	Risk Based Approach
RFM	Recency, Frequency and Monetary

SAR	Suspicious Activity Report
SAS	Statistical Analysis System (Institute)
SEE or ENI	Self Employed Entrepreneur
SMC or SME	Small and/or Medium Companies
SOM	Self-Organizing Maps (Konohen's)
SOM	Konohen's Self-Organizing Maps
SVM	Support-Vector Machine
UAT	User Acceptance Test
UNSCRs	United Nations Security Council Resolutions
VA	Visual Analytics

1. INTRODUCTION

In the current era that we live in, money laundering and the financing of terrorism has been gaining more focus and is becoming an even greater concern for the governments. As such, even tighter legislations in the matter of AML/CFT, have been imposed onto the institutions that deal with sectors/markets, within which these kinds of activities tend to flow through[1].

The increase of AML/CFT mechanisms has as a direct objective the increase of the detection of money laundering and the financing of terrorism, and a consequential decrease in the practice of these kinds of activities. This objective might have been met by some percentage, but as consequence, it triggered the creation of newer and more complex money laundering and the financing of terrorism techniques. Translating into an even more complex and complete task to be done by the institutions affected by these legislations, in order to efficiently combat these kinds of activities[2].

The banking institution in question being a financial institution, has been affected by these regulations ever since they came into action, but with this increase of strictness and higher frequency, by which these newer legislations have been imposed on, and by the increase of complexity on the money laundering and the financing of terrorism stratagems/techniques used, its analytics capabilities must also adapt and improve in order to respond to such necessities.

To date, this banking institution has been working with software that is considered the norm to when it comes to AML/CFT, a 'rule' based software. 'Rule' based software is to be understood as, an AML/CFT system that relies and analyzes the financial activities by account and inter-related accounts (logical entities), accounts that have financial relationship between them. This methodology produces detailed and accurate results[3], in accordance to the entity in analysis. Having this kind view can negatively impact this methodology effectiveness, making it too restrictive and focused in local financial activities, in cases where the wallet of clients is characterized by multiple accounts of common ownership and parameter wise intra-related accounts. This can be translated into having a segregated and disconnected view of the client activities, making the analysis process of the generated alerts heavier. Another inherent problem is also the incapability of adapting to the more complex and wide money laundering and the financing of terrorism strategies that keep on being created, due to the restrictive and fixed rule creation options.

This banking institution, identifying itself in that scenario, takes the decision to change its AML/CFT software into one with a more 'behavior' like methodology. 'Behavior' based software is to be understood as, an AML/CFT system that relies and analyzes the financial activities of the client's related accounts. This methodology inherits the qualities of the 'rule' based one, but mitigates its flaws, having the ability of monitoring the behavior of clients across all his accounts without losing the detailed analysis in each of his account's activities.

The present paper describes the implementation process of a 'behavior' based AML/CFT software, SAS® Anti-Money Laundering 7.1, and the planning of its continuous optimization, to better fit the this banking institution's environment and clients.

[1] (Edmonds, 2018)
[2] (Hoque, 2009)
[3] Considering that the segmentation and parameterization of the rules and respective values were done correctly.

1.1. INTERNSHIP OBJECTIVES

The main objective is the implementation of a fully functional and efficient behavior-based AML/CTF system. To support this objective, three main secondary objectives came into play, namely, the compliance with the legal authorities' requirements and proposals; the adaptation of the system to the client behavior and not to a set of strict and rigid rules, so that a more adjusted spotting of suspicious behaviors can be achieved; And the capability to do a continuous fine-tune to the system so that it can achieve its most optimal efficiency, across its life cyclic.

1.2. ORGANIZATION CHARACTERIZATION

This chapter serves as an introduction to the institution whereas the internship toke place, so that the reader has contextualization for the decision-making process and data reported in this paper.

1.2.1. The Banking Institution

This banking institution was founded following the deregulation of the Portuguese bank system, which allowed for the fixation and creation of commercial banks of private capital in the Portuguese market. Since then, this banking institution has gone through bank acquisitions and absorptions, market expansions, both geographically and product wise, and restructure actions.
As of now, this banking institution is the largest Portuguese private bank, having a dominant position in the Portuguese market, being it's second largest bank and first banking institution, in terms of market share. This translates into an expressive national bank distribution network, with more than 500 branches and an international network of more than 1000 branches, which serves a client wallet that amounts to more than 5.4 million of clients worldwide.

1.2.2. Compliance Office

The Compliance Office has as its main mission the insurance of the adoption, by all of the group's institutions, of both the internal and the external norms, which work in conformity with their respective activities. Helping in the mitigation of the imputation risk to those institutions, of sanctions or support, by those, of patrimonial loses or reputational damage. This banking institution's Compliance Office is divided into 3 main departments, across which responsibilities were distributed, in order to ensure the accomplishment of previously mentioned mission (Figure 1).

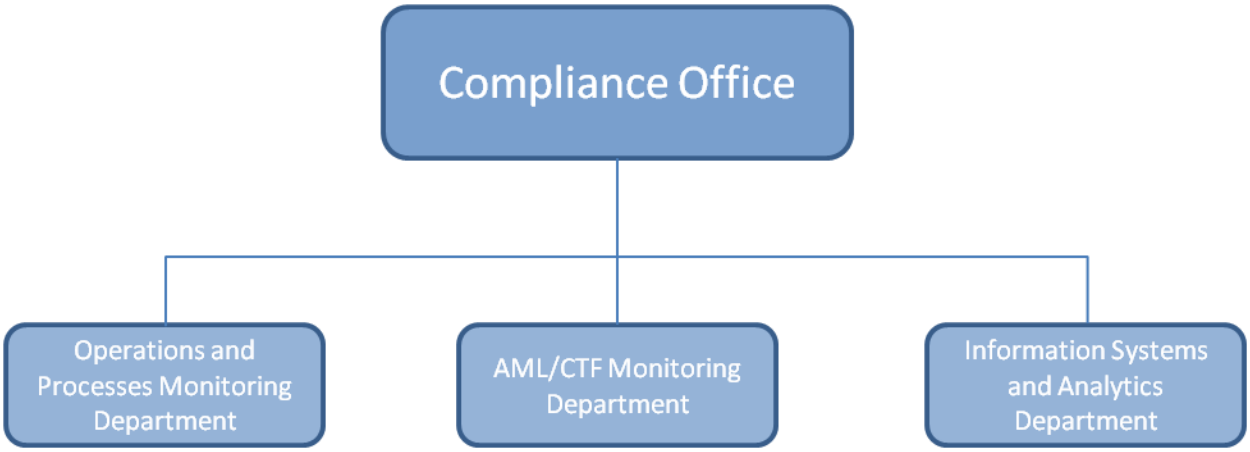

Figure 1 - Compliance Office Constitution (Source: Author)

1.2.3. Operations and Processes Monitoring Department

This department is responsible for the processes', legislations', correspondent banks' and third party risk assessment's compliance; The monitoring and support of the foreign operations, the market abuse compliance, as well as the control over the approval of new products; The interactions and relationship with supervisory entities, supervision authorities and the police; and lastly, the management of the internal compliance rules review, the compliance training development and the strategic planning definition of the compliance office.

1.2.4. AML/Counter Terrorism Monitoring Department

This department ensures the Customer Due diligence/Know Your Customer; the transactions' risk-based approach; the monitoring and control of filtered transactions (Enhanced due diligence); and the reporting of deemed suspicious transactions.

1.2.5. Information Systems and Analytics Department

This department supports the other compliance office departments, by managing the AML/CFT system's populations, scenarios and parameters; Managing a variety of high relevance lists, such as, the sanctioned jurisdictions', the PEPs' and the watch/high risk's lists; and providing data analyses.

2. LITERATURE REVIEW

This chapter will be responsible for introducing and explaining all of the required notions for the understanding of the thesis. These notions will vary greatly, from theoretical concepts, legal obligations, relationships, theorems, algorithms and much more. The chapter will be divided into two major groups, a more conceptual one and another more technical.

2.1. BUSINESS/CONCEPTS REVIEW

The first group of topics to be touched upon is the conceptual segment. It is here that the major concepts to be used throughout the thesis will be introduced and defined. These concepts and their respective definitions will have a major role in the work developed, even if just as contextualization or even just as constraints.

2.1.1. Money Laundering

Fundamentally, the process by which the legal activities generate earnings is very similar to those that are considered illegal, differing mainly on the way, of how they are perceived by the authorities, by which they are governed. Imposing this, another fundamental difference, the facility of usage of earnings. Illegal origin earnings still can circulate, but not as freely as its legal counterpart.
Money laundering comes into play as a way to bring down these constraints, transforming/masking as legal, the illegal funds[4](Table 1). The processes by which the transformation/masking is accomplished varies widely, but one of the most commons is by entering the banking circuit, without being identified/detected as illegal funds, since once funds have entered the banking circuit, they are considerate as legal funds. This happens because the institutions[5] have the obligation to validate theirs incoming funds, in other words, any funds movement[6] by which the institution takes part in.

Placement	Furtive integration of the illegal funds into the legal financial circuit.
Layering	Movement of the funds, within the legal financial circuit, in order to create confusion, increasing the difficulty on the tracing to the original source of the illegal funds.
Integration	Integration into the legal financial system, by the means of the last few transactions, until it appears to be legal origin funds.

Table 1 - Money laundering simplified three step process

2.1.2. Financing of Terrorism

The definition of terrorism can be a very controversial topic, since its definition changes accordingly with the perspective of the person viewing it. For instance, a suicide bomber can be viewed as a terrorist act, from the perspective of the country which the suffered the bombing, and at the same

[4] Definition in Art.1 (3)Directive (EU) 2015/849
[5] To be specified in the AML/CTF Legislations Chapter
[6] With certain exceptions/nuances to be specified in the AML/CTF Legislations Chapter

time, as an act of rightful retaliation against a repressive regime, from the perspective of the group or organization responsible for the bombing.

In accordance with the context of this paper, one good description of terrorism is in the Framework Decision on Combating Terrorism (2002), which is denoted by the EU for official and legal purposes, as:

"(a) attacks upon a person's life which may cause death;

(b) attacks upon the physical integrity of a person;

(c) kidnapping or hostage taking;

(d)causing extensive destruction to a Government or public facility, a transport system, an infrastructure facility, including an information system, a fixed platform located on the continental shelf, a public place or private property likely to endanger human life or result in major economic loss;

(e) seizure of aircraft, ships or other means of public or goods transport;

(f) manufacture, possession, acquisition, transport, supply or use of weapons, explosives or of nuclear, biological or chemical weapons, as well as research into, and development of, biological and chemical weapons..."[7]

The practice of these acts usually requires long periods of planning and network complex and wide enough to support them, and as such the maintenance of such things require a considerable amount of income in order to assure their functioning. Having in consideration the previous topic, the financing of terrorism fundamentally distinguishes itself apart from the topic of money laundering, in the way that it can use, either legal or illegal funds, to finance terrorist activities[8], regardless if it was done direct or indirectly, or even if it was done with the intent to do so.

2.1.3. AML/CTF Legislations

Terrorism and money laundering have always been and are currently becoming a more apparent and frequent reality in countries all around the globe. As such, it is a natural response, from each country, and by their respective governments, the creation and tightening of rules and legislations, surrounding these topics, to mitigate and ultimately end the practice of such activities.

In the context of this paper, the legislation that is going to be considered is the EU and Portuguese government-imposed legislation[9][10] and duties (Table 2), which are very similar, if not the same, as the up to date and globally accepted AML/CTF directives.

It is relevant to note that not all organizations are subjected to these legislations, being the ones that are, namely these ones:

- Financial Institutions;
- Non-Financial Institutions;
- Equivalent Institutions, to the legislation obliged Institutions;
- Payment service providers, obliged by the (EU) 2015/847 regulation;
- Curators and registry officials.

[7] The Council of EU 2002, 475-JHA

[8] Definition in Art.1 (5) Directive (EU) 2015/849

[9] Banco de Portugal Warning 5/2013

[10] Lei n.º 83/2017

Identify Duty	The institutions have the mandatory duty to identify each customer that made business with them and/or made transactions of equal or greater amount than 15.000 Euros, regardless of the number of operations that compose the transaction. They are also obliged to adopt, complement or repeat the client's, representative's and effective beneficiary's identification process, if there are suspicions that they are involved in money laundering and/or terrorism financing, or if there are doubts about the veracity or adequacy of the identifying data, in the normal course of business interaction.
Diligence Duty	The diligence duty dictates that the institutions must do a 'context' analysis to the transaction, to the client or to the organization structure, if there is any suspicion of the activities previously introduced in the Identify duty. As for 'context' analysis, it is to be considered as, analysis of the participations that client has, the nature of their activities, the structure of the organization (BEF's), their familiar bonds, their relationship connections and the nature of their work position (PEP's). The implementation of this duty should be done in parallel with the Identify duty.
Control Duty	The present duty dictates, the necessity of a definition and implementation of an internal control system that integrates policies, means and procedures focused on ensuring the compliance of the legal normative and regulations related to the theme of AML/CFT, in such a way that it avoids, its involvement in operations related to these kinds of crimes. As of this, it can be advisable to have as best practices, the limiting to a writing format of the policies, the means and the procedures that integrate the internal control system, including the client acceptance policy; The insurance of the sufficiency and adequacy of the human, financial, material and technical resources that are related to the AML/CFT; The divulgement, to the relevant parties, of permanent access to up-to-date information, about the fundamental principles of the internal control system, related to AML/CFT, as well as about its instrumental norms and procedures necessary for the execution; Implementation tools and informatics systems adequate to the client and operations record and control, having as objective their monitoring, suspicious operation detection and emission of the corresponding alert indicators; The effective necessity of an internal control system's continuous quality evaluation, as well as its regular adequacy and efficiency testing.
Train Duty	The train duty enforces the definition and appliance of a training policy, adjusted to the functions exerted by the collaborators, relevant in the context of AML/CFT. Ensuring that they have permanent and total knowledge about the juridical framework, in effect and applicable in the AML/CFT domain; The preventive policies, means and procedures defined and implemented by the institution. With this in mind it advisable to propagate complementary and contextual knowledge regarding this thematic, providing orientations, recommendations and information originated from judiciary and supervision authorities or representative associations of the sector; The typologies, tendencies and techniques associated with money laundering and terrorism financing; The vulnerabilities of the products and services made available by the institutions, as well as its specific emergent risks; The reputational risks and the

	consequences, of misdemeanor nature, derivative from the AML/CFT's preventive duties negligence; Specialized professional responsibilities, in the theme of AML/CFT, more specifically, the operational procedures associated with the compliance with the preventive duties.
Others	Refusal duty (Exertion of this duty when at least one the of the required documents are not made available, requiring a context analysis to see if there is any need to report); Conservation duty (This duty dictates that the institution should maintain copies or data extracted from all documents presented by the clients); Exam duty (Duty to examine the conducts, activities or operations, which characterization elements become particularly susceptible of being related with money laundering and terrorism financing crimes); Communication duty (Duty to communicate suspicious operations); Abstention duty (Exertion of this duty when the institution's abstention to execute operations is not possible, being a document with the reasons, as to so, required); Collaboration duty (Adoption of an activity information archive into the institutions structure, to allow relationships of collaborations with the designated entities); Secrecy duty (This duty dictates that the institution, in its regular activities and when implementing and respecting the other duties, acts with the necessary cautions in the interactions with the clients, that are related with suspicious communicated operations).

Table 2 - Preventive duties[11] of the bounded organizations

2.1.4. AML/CFT System

The AML/CFT systems surges as an answer for the increase of the legislation load that has been being imposed, and for the consequential heavier/more detailed analysis required from the organization, in order to ensure so. These systems can differ in methodology, effectiveness and completeness, but the underlying logic and general objective is the same, the strengthening of the financial sector integrity, by enforcing that the financial systems and economy are protected from threats related to money laundering and terrorism financing.

2.1.4.1. Intermediate Objectives

A. *"Policy, coordination and cooperation mitigate the money laundering and terrorism financing risks*." (FATF)
B. "Proceeds of crimes and funds in support of terrorism are prevented from entering the financial and other sectors or are detected and reported by these sectors." (FATF)
C. "*Money laundering threats are detected and disrupted, and criminals are sanctioned & deprived of illicit proceeds. Terrorist financing threats are detected and disrupted, terrorists are deprived of resources, and those who finance terrorism are sanctioned, thereby contributing to the prevention of terrorist acts.*" (FATF)

[11] Banco de Portugal Warning 5/2013

These intermediate objectives work together in order to achieve the AML/CFT system's main one, but there are some immediate objectives (Table 3, 4 and 5) that are necessary, in order to build a solid foundation that can sustain each of these individually.

2.1.4.2. Immediate Objectives

Risk, policy and coordination	"*Money laundering and terrorist financing risks are understood and, where appropriate, actions coordinated domestically to combat money laundering and the financing of terrorism and proliferation*." (FATF)
International cooperation	"*International cooperation delivers appropriate information, financial intelligence and evidence, and facilitates actions against criminals and their assets.*" (FATF)

Table 3 - Objectives related to the Intermediate objective 'A'

Supervision	"*Supervisors appropriately supervise, monitor and regulate financial institutions and DNFBPs for compliance with AML/CFT requirements commensurate with their risks.*" (FATF)
Preventive measures	"*Financial institutions and DNFBPs adequately apply AML/CFT preventive measure commensurate with their risks and report suspicious transactions*." (FATF)
Legal persons and arrangements	"*Legal persons and arrangements are prevented from misuse for money laundering or terrorist financing, and information on their beneficial ownership is available to competent authorities without impediments*." (FATF)

Table 4 - Objectives related to the Intermediate objective 'B'

Financial intelligence	"*Financial intelligence and all other relevant information are appropriately used by competent authorities for money laundering and terrorist financing investigations*." (FATF)
Money laundering investigation & prosecution	"*Take advantage of the physically borderless digital world, to facilitate processes, improve client relationships and to enchant efficiency and effectiveness. Being digital without forgetting the world of emotions within which we live in*." (FATF)
confiscation	"*Being close is the fastest way to be the first to react, respond and correct. It is to know what is needed, even if it is required to do so before there is a necessity. Being close does not translate into to being everywhere, but into to being where it is needed.*" (FATF)
Terrorism financing investigation & prosecution	"*Money laundering offences and activities are investigated, and offenders are prosecuted and subject to effective, proportionate and dissuasive sanctions.*" (FATF)

Terrorism financing preventive measures & financial sanctions	"*Terrorists, terrorist organizations and terrorist financiers are prevented from raising, moving and using funds, and from abusing the NPO sector*." (FATF)
Proliferation financial sanctions	"*Persons and entities involved in the proliferation of weapons of mass destruction are prevented from raising, moving and using funds, consistent with relevant UNSCRs*." (FATF)

Table 5 - Objectives related to the Intermediate objective 'C'

All these objectives are then considered when developing, implementing and upgrading the institution's AML/CFT system. Systems which, even with all the diversity of methodologies and institution's adjustment specifications, follow a general operational guideline/structure:

Screening

Individual's and entities' record match search against an official distributed list. This search uses the individuals and the businesses identifying records, such as names and addresses. After this match, there is a necessity to review the maintained accounts and transactions, affiliated with the current context, within a specific period of time. In accordance with the ACAMS, this stage can be intimidating and overwhelming, in case of a poorly done calibration of the system, since the cheer number of records produced, and number of false positives would be unmanageable.

Communications (Screening related)

The share of information, between financial institutions, that are related to client's communication logs and mutual clients.

'Currency transaction reporting' (CTR)

Monitoring of the client's behavior, delimitating scenarios and behaviors parameters, within which the suspicious ones fall into. In the cases where suspicious behaviors are detected, alerts are generated, aggregating and providing information about the owner and the specifications of such behavior, allowing for an analysis to be done, in order to assert the veracity of the suspicious behavior.

'Currency transaction reporting '(CTR) exemptions

Financial institutions are bound to have a certain level of trust relationship with a small range of its client base, such as banks, state and government agencies, or certain trust in a set of specific behaviors, in such a way that a 'CTR' exemption is justified.

Case management

Centralized location for the building and storing of case notes, attaching documents and tracking of communications. This assists undergoing and/or the re-opening of client investigations, by having a

centralized place, where all the relevant information is stored and arranged accordingly. Structurally wise, this can also be extremely helpful, since it allows for an information interconnection and availability between co-workers, which can be helpful when there multiple treating the same or interconnected cases.

Watch list screening

Additional watch list screening, to complement the activities' filtering with additional lists from complementary sources. Existing also the concept of custom lists, that allows the creation of 'unwanted' client's lists for the institution.

Suspicious activity report (SAR) filing

Once the analyses are done, and the case/alert is still considered as suspicious, a SAR is filled. SAR's are then processed to the competent entities to further investigate or to complement ongoing investigations. SAR's can be an invaluable resource, to these entities, because it is only through them that they have access to this kind of information.

Dashboard reporting

The dashboard reporting is an essential part of the system, because they are the system's retrospective tool, allowing the view, in trending, of the institution's overall risk profile and validation of parameters. There are systems that have a custom report functionality, which allows for more detailed view, on the areas that the institution deems desired.

2.1.5. RBA

FAFT since 2007 has been outlining the importance of implementing the RBA as a part of the AML/CFT system, publishing several specific guides in accordance with the implementation sector, being recommended as a methodology to be followed through, by regulatory bodies such as the EU directives, FCA, DFA and others. In a general case, RBA should be understood as a methodology[12] that does not eliminate the risk, but that tries to mitigate its impact, enabling the understanding of the risks, through a process of identification, classification and scoring of the risk factors (Table 6).

RBA can be viewed as a chain process of 4 stages:

Identification	Risk factors identification
Assessment	Evaluation of the risk factors, and assessment of their level of risk
Knowledge	Exploration of the risk factors, to extrapolate and discover the extent to which they can impact the organization
Action taking	Creation of a mitigation plan

Table 6 - RBA as four stage[13] process

[12] (Touil, in the name of Association of Certified Anti-Money Laundering Specialists, 2016)

[13] Based on the Groundwork RBA implementation

In order to correctly function it is required that the information surrounding the client characterization, such as its potential as customer, background, business line, relationship history with the institution, and other complementary information are available. This information should fall into one of four main client's characterization areas, correctly illustrating it:

1. Customer(background)
2. Geographic locations (nationality, residence, incorporation, …)
3. Products
4. Services

Accordingly with the 'meaning' of the information, and with the category to which it belongs, a weighting schema is run, and the risk level of the client is asserted, usually ranging from one to five ([1; 5]). Going into further detail, the evaluation of the 'meaning' of the information is an ever-changing paradigm, requiring continuous maintenance. This is so, because the information context is systematically changing, meaning, the way that a specific characteristic is viewed as, risk wise, can change according with the influences of its context, being it direct or indirectly (per example, a country that used to have a huge problem with corruption, could have made considerable effective efforts to resolute this problem, can no longer be considered a potential risk country of origin).

2.1.5.1. Equilibrium point between the risk assessment and the institution's risk appetite

Each institution is unique, either on customer base or on product and services diversity, having each one a very particular risk appetite (risk which they consider but decide to accept, in trade off with some benefit). This risk appetite management usually, translates into a segmentation of the risk levels into groups of control and analysis procedures:

- Due Diligence group, defines the segmentation of the analysis process, requesting more detailed procedures to the higher risk levels. This group usually has three analysis options: the Customer Due Diligence (CDD), which is used for low-risk scores analyses, implementing the regular KYC procedures; the Simplified Due Diligence, which is used for the medium-risk scores, complementing the previous procedure with the request of further information that might justify the risk; the Enhanced Due Diligence (EDD), which is used for the high-risk scores, that on top of the previous procedures, involves a thorough search of the client and subsequent related information on the potential clients, being it either on the web, or thought the form of questioners.
- Control/Validation group, defines the level of authority necessarily in order to process and approve the acceptance of the client. Usually, for the lower risk scores, the relationship officer/manager approval suffices, for the medium risk scores, the unit's head approval is then necessary, and for the higher risk scores, the necessary approval escalates to the AML committee's and/or MLRO's.

These procedures require some internal control actions, mainly two types of actions, one that enforces quality assurance on to the processes, and another in order to assure their correct and appropriate implementation and usage.

2.1.6. Regulatory Entities

The financial institutions have a set of regulatory entities with specific monitoring duties, all culminating in the assurance and validation of the AML/CFT legislation compliance and the correct enforcement of the AML/CFT's system operation. Usually it does this through the process of audits. In the Portuguese context, the entities that the financial institutions have to 'answer' to are Banco de Portugal (BP) and Comissão de Mercados de Valores Mobiliários (CMVM).

2.1.6.1. Banco de Portugal (BP)

"Banco de Portugal acts in the field of prevention of money laundering and terrorist financing (ML/TF).

The Bank has the supervision responsibility on the compliance of preventive duties regarding ML/TF (refer to the previous AML/CTF Legislation segment) by the credit institutions, financial companies, payment institutions, electronic money institutions, branches established in Portugal and entities providing postal services as well as financial services.

Supervised institutions are required to comply with several duties such as, (i) customer identification and due diligence, (ii) duty to keep documents and records on customers and operations, (iii) scrutiny and reporting of suspicious operations and (iv) adoption and implementation of internal control systems that are adequate to the ML/TF risk intrinsic to each institution.

In addition to providing for compliance with these duties, Banco de Portugal has regulatory functions and actively participates in the preparation of the legal framework governing ML/TF.

Banco de Portugal is also represented in national and international bodies dealing with these issues, among which the AML/CFT Coordination Committee and the Financial Action Task Force (FATF)." (BP)

2.1.6.2. Comissão de Mercados de Valores Mobiliários (CMVM)

"The Portuguese Securities Market Commission, also known by its initials "CMVM", was established in April 1991 with the task of supervising and regulating securities and other financial instruments markets (traditionally known as "stock markets"), as well as the activity of all those who operate within said markets.

The CMVM is an independent public institution, with administrative and financial autonomy. The CMVM derives its income from supervision fees charged for services and not the General State Budget. "(CMVM)

"The supervision carried out by the CMVM includes the following:

> *Constant supervision of the acts of individuals or entities, which operate in capital markets, for the purpose of detecting unlawful acts, particularly in stock market trading;*
>
> *Monitoring rules compliance;*

Detection of criminal offences;

Punishment of infringers, namely by the imposition of fines;

Grant registrations of individuals and operations to check the compliance with applicable rules;

Information disclosure, particularly on listed companies, through its website on the Internet. " (CMVM)

2.2. TECHNIQUE REVIEW

The technique review chapter will serve as an introduction to the techniques, methodologies and algorithms that exist and that could be used in a beneficial way in the implementation, validation and maintenance of the AML/CFT. The chapter will explore the techniques in different depth levels, in accordance to their relevance in the project and thesis context.

Before introducing any technique, it is important to note that there is an elementary division of techniques in this field. It distinguishes techniques in accordance with the way that they use the data to validate and to improve their results, existing main two groups, the supervised learning and the unsupervised learning techniques.

The supervised learning is characterized by the existence of guidance on the learning process, existing a value/attribute that can serve as a comparison and validation tool, to assert the success of the technique. In contrast the unsupervised learning is characterized by the non-existence of guidance on its process, not being able to ensure that the technique's result was successful, at least not in the assertiveness level of the supervised one, existing more subjectivity in this type of techniques. There are cases where the distinction can be dubious, as in the case of the neural networks, but these cases will be explained further down as of the introduction of such techniques.

Having introduced these notions, a more knowledgeable and sustainable approach can be used in the analysis and exploration of the techniques to be presented.

2.2.1. Classification & Regression

Classification and Regression are both methodologies that fall within the supervised learning category, being that these methodologies, specifically, use a set of pre-classified data, a data set which not only contains variables used to describe the observation, but also contains the class to which the observation belongs to (the predictor variable), in order to develop models that can predict how new records should be classified in the best way possible. Trying to draw conclusions from observed values.

The difference between the classification and regression is fundamentally the type/category of the model result/prediction. The classification methodology focuses on the categorical predictions, the labeling of the observations, meanwhile, the regression methodology attempts to predict continuous variables/values, not trying to label the observation, but trying to quantify it.

2.2.2. Clustering

Clustering is a methodology that falls within the unsupervised learning category. As the main methodology in it, it has been a very widely discussed, since its relevance covers a wide range of business and interest areas/sectors. As such, innumerable definitions have been developed and worked upon. A brief collection of widely accepted definitions should be introduced, in order to illustrate the different views that exist towards Clustering.

The definition given by (Anil, K. Jain, 2010) states that "*Cluster analysis is the formal study of methods and algorithms for grouping, or clustering, objects according to measured or perceived intrinsic characteristics or similarity*"; (Rendón, et al., 2008) says that "*determine the intrinsic grouping in a set of unlabeled data, where the objects in each group are indistinguishable under some criterion of similarity*"; (Jain, Murty and Flynn, 1999) states that "*Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters)*"

With this said, it is possible to extrapolate that generally speaking, clustering tries to aggregate similar objects into groups, in such a way that different groups have different characteristics. Meaning that, it tries to group the objects so that, there is an increased intra-group similarity and at the same time a low inter-group similarity, requiring this the determination of the similarity's 'structure'. It does so without any need for pre-classified data, but working under the belief that similar individuals, in terms of the variables used, will have similar behaviors.

This definition can fit pretty much any interest area that makes use of clustering, being some examples of these the following:

- Biology, assessing the similarity of the genetic data in order extrapolate the populations structures;
- Business, by allowing the segmentation of their clients accordingly with the value that they bring to the company;
- Social awareness, identifying areas whereas greater incidents of a particular type of crime takes place, and enforce adequate counter measures to the case in question;
- Social network, recognizing specific communities within the large groups of users.

Until now, the clustering results have been called groups, for simplicity sake, but from now on they will be referred to as clusters.

In order to correctly and completely understand the ways of use, as well as, the use in itself of clusters, the notion of 'cluster center' is a necessity. This 'center' can mainly have two distinct forms, centroid or medoid. A centroid is generically known as the center of the mass of a geometrical object in the multivariate context, being the point whereas all the variables mean intersect. This last definition is the one that better fits in the context of the clustering, and as such, a centroid can be considered as the conceptual mean of all the variables/observations belonging to a cluster, and to which they belong to, in relation to the proximity/similarity to the cluster's centroid in itself.

A medoid is very similar to the definition centroid but distinguishes itself by the form that it assumes. In the case of the centroid, its representation is a conceptual one, within the universe/context of the dataset in usage, but in the case of the medoid, its representation is required to be of an existing one in the dataset, meaning, it assumes the form of an existing dataset observation. The selection of this observation is made through the process of calculi of the average dissimilarity to all the other observations in the cluster. The observation that has the lower value of average dissimilarity in is considered the medoid.

2.2.3. Similarity Criterion

The ability to calculate/quantify the degree of similarity between observations is a fundamental and essential requirement in the context of this thesis, since it is this degree that will fuel the algorithms functioning, and as such the topic of the similarity criterions must be introduced. These are the methodologies, as previously said, that support the usage of the concept of similarity. They do so through the calculi of distances between the population objects/observations. These methodologies use measures based on the Euclidean space, which consider the observations as points; the cosine space, which considers the observations as vectors; and the Jaccard space, which considers the observations as sets.

The geometric measures based on Euclidian space have dominated the analyses of relations of similarities, and as such these will be the main focused ones in the thesis. As complementation to what was previously said, these distances represent the observations as points in the multidimensional space, so that the dissimilarities between the objects/points correspond to their distances. Being that these metrics usually is required to satisfy the following proprieties:

1. Symmetry, given two objects, x and y, with x ≠ y, the distance between verifies the following property

$$d(x,y) = d(y,x) > 0$$

Equation 1 - Symmetry property

2. Identification of the indiscernible, given two objects x and y:

$$d(x,y) \neq 0 \Rightarrow x \neq y$$

Equation 2 - Identification of the indiscernible

3. Non-Negativity, given two objects x and y

$$d(x,y) \geq 0$$

Equation 3 - Non-Negativity property

4. Triangle Inequality, given three objects, x, y and z, distances between them satisfy the property:

$$d(x,y) \leq d(x,z) + d(z,y)$$

Where $d(x,y)$, represents the distance between the points x and y (Figure 2).

Equation 4 - Triangle inequality property

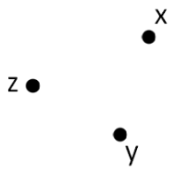

Figure 2 - X, Z and Y representation in a bi-dimensional space (Source: Author)

The way of obtaining the distance itself, between observations/points can vary from case to case and from interest area to interest area. The following, are a few of the most recognized and used distances:

2.2.3.1. Euclidean Distance

Distance two elements (i, j) is the square root of the sum of squares of the differences between the values of i and j for all the variables (v = 1, 2, …, p):

$$d(i,j) = \sqrt{\sum_{v=1}^{p} (X_{iv} - X_{jv})^2}$$

Equation 5 - Euclidean distance formula

2.2.3.2. Manhattan Distance

To better understand this distance, let's imagine a grid with two intersections, where, if the Euclidean distance would be the hypotenuse, the Manhattan would be the two peccaries. Interestedly enough, this distance got its name from the New York's city borough, because its roads were built as a grid, and the distance to travel between two points through its roads, would be pretty much equivalent to the distance calculated through the Manhattan distance method (Figure 3). Being the method, the following:

$$d_{ij} = \sum_{v=1}^{p} |X_{iv} - X_{jv}|$$

Equation 6 - Manhattan distance formula

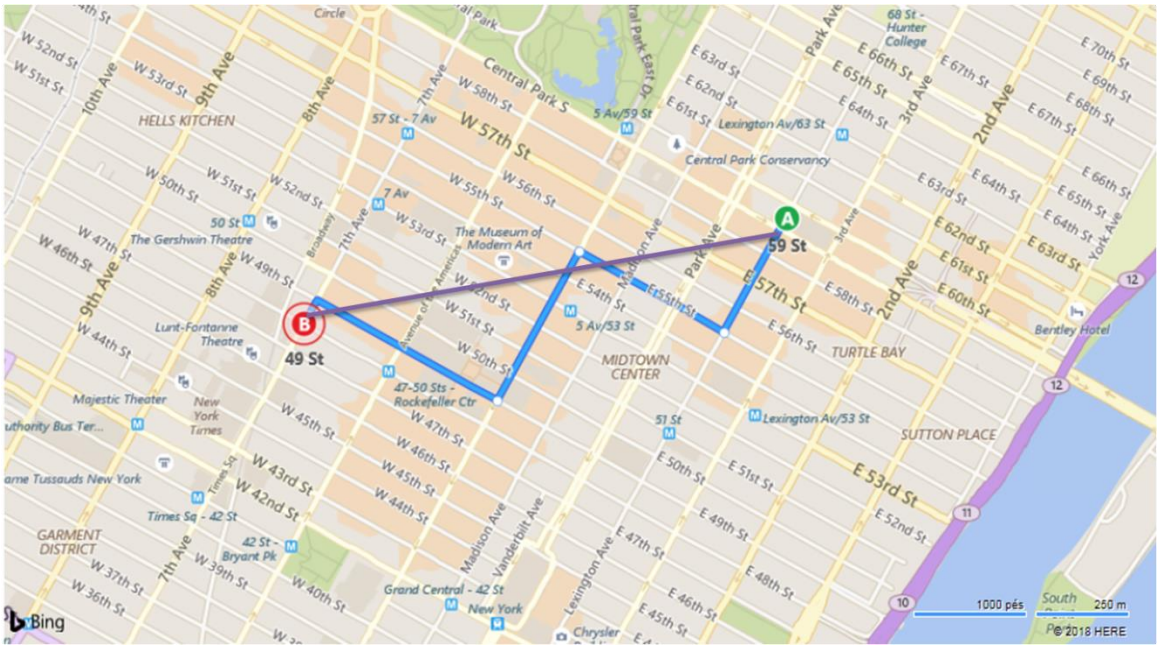

Figure 3 - Manhattan Distance Example (Source: Author)

2.2.3.3. Minkowski Distance

This distance can be defined as an absolute distance, having the ability to calculate different 'similarity criterions', by changing the value of the parameter r. It is viewed as a generalization of the Euclidean and Manhattan distance, since it coincides with the Euclidean distance, when its parameter r = 2, and with the Manhattan's, when r = 1:

$$d_{ij} = \left(\sum_{v=1}^{p} |X_{iv} - X_{jv}|^{r} \right)^{\frac{1}{r}}$$

Equation 7 - Minkowski distance formula

2.2.3.4. Mahalanobis Distance

The Mahalanobis distance represents the distance between two points in the multivariate space. This distance is particularly useful in the cases where two or more variables are correlated. When the variables are uncorrelated, the distance between two points can be measured in a rule like way, since the graphical representation of the axes in the Euclidean space, denotes them at right angles, which makes this distance identical to the Euclidean one. In the case of two or more variables being correlated, the axes representations are no longer at right angles, making the usage of the rule like way of measuring the distance between two points suffer a huge detriment (Figure 4).

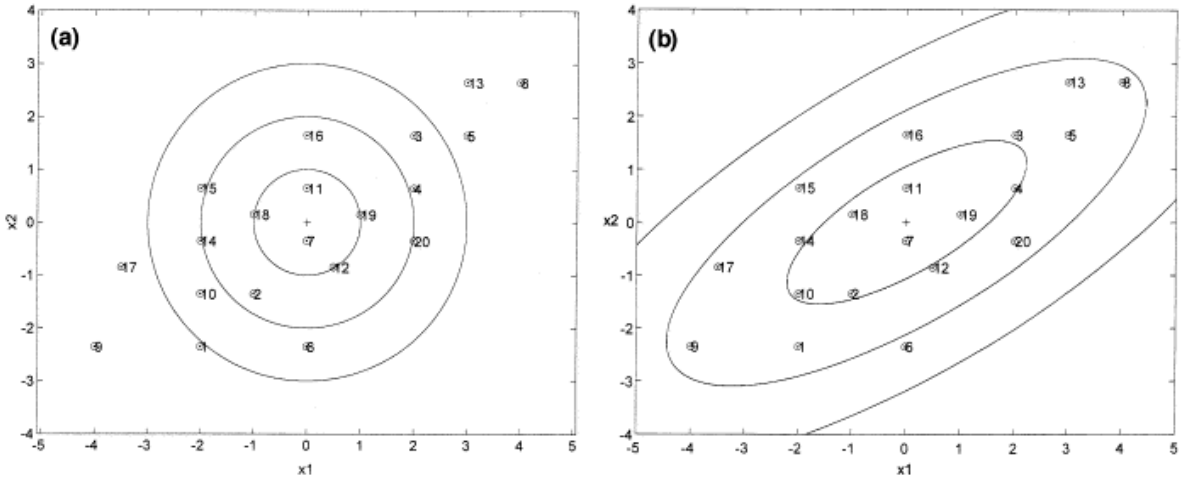

Figure 4 - Impact of the Mahalanobis Distance (Source[14])

The Mahalanobis distance is able to overcome this 'problem' through the structure of its logic, which translate into a measurement of the points/observations distance to the central point (centroid), which can be considered as the overall mean for the multivariate data, ensuring that it takes into consideration the points' distribution (covariance matrix) and by being independent from the variable scaling. It does so, through the following formulas:

$$d(Mahalanobis) = [(X_B - X_A)^T \cdot \boldsymbol{C}^{-1} \cdot (X_B - X_A)]^{0.5}$$

Where:

X_B and X_A are pairs of objects;

C is the sample covariance matrix.

[14] https://ars.els-cdn.com/content/image/1-s2.0-S0169743999000477-gr1.gif (2018)

Equation 8 - Mahalanobis distance formula

It can also be computed, from each observation to the data center, through the following formula:

$$d(X_i) = [(x_i - \vec{x})^T \cdot \boldsymbol{C}^{-1} \cdot (x_i - \vec{x})]^{0.5}$$

Where, ^{i} = 1, 2, …, n

Equation 9 - Mahalanobis distance formula, for each observation

The major problem with this distance lies in the dependence that it has on the inverse of the correlation matrix, which cannot be calculated in cases where the variables are highly correlated.

Regardless of the distance calculation method in usage, there might be cases when certain observations variables' are more relevant than others. In such cases it is advisable to proceed with a weighting process, giving more importance, and/or decrementing the relevance of certain variables. Per example, for the Euclidian case, its weighed format should look like the following:

$$d(i,j) = \sqrt{\sum_{v=1}^{p} w_v (X_{iv} - X_{jv})^2}$$

Where, w_v represents the weighting schema defined for each variable of the points/observations.

Equation 10 - Euclidean distance weighed formula

In regards of how to critically assess the distances between clusters, there are a few methodologies that can be used, being the most used and accepted ones the following:

2.2.3.5. Single Linkage

Single Linkage is a methodology based on the nearest-neighbor, meaning that, the distance between clusters is obtained by calculating the distance between the two most similar points/observations, from each cluster (Figure 5). This tends to form longer and slender clusters.

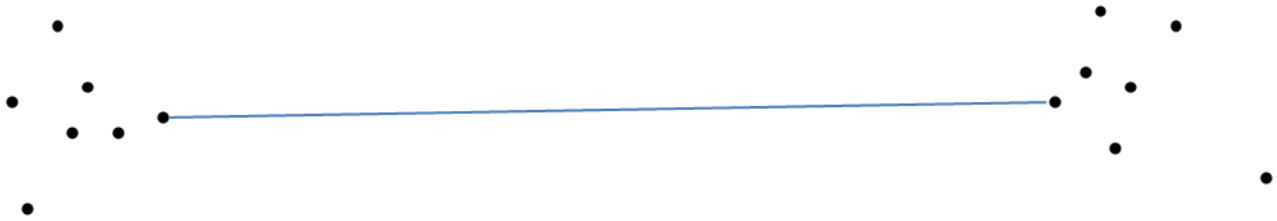

Figure 5 - Single Linkage Example (Source: Author)

2.2.3.6. Complete Linkage

Complete Linkage is a methodology based on the farthest-neighbor, meaning that, the distance between the clusters is obtained by calculating the distance between the two most different

points/observations, from each cluster (Figure 6). This tends to form more compact and sphere-like clusters.

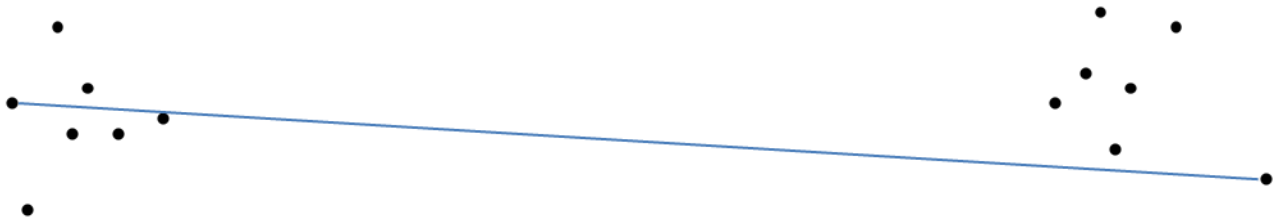

Figure 6 - Complete Linkage Example (Source: Author)

2.2.3.7. Average Linkage

Average Linkage is a methodology based on the average distance between all the points/observations that constitute each cluster (Figure 7). Since this methodology considers all the points/observations, it is less dependent on the extreme values, than the two previous methodologies.

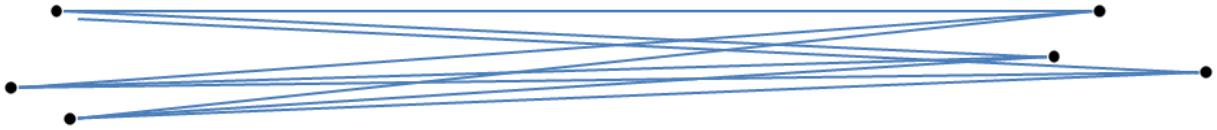

Figure 7 - Average Linkage Example (Source: Author)

2.2.3.8. Centroid Linkage

Centroid Linkage is a methodology based on the distance between each of clusters' centroids, meaning that, the distance between the clusters is obtained by calculating the distance between the point, from each cluster, whereas all the points/observations variables' means intercept (Figure 8).

Figure 8 - Centroid Linkage Example (Source: Author)

As of now, clustering has been referred and bordered as a theme and not as a methodology or technique, because clustering is not a unique methodology but rather a culmination of them, deriving this from the wide range of uses that clustering serves across all the interest areas that use it. From this culmination is possible to extrapolate, two main technique groups, the Hierarchical and the Partitional Clustering.

2.2.4. Hierarchical Clustering

Hierarchical clustering is a group of clustering techniques that are characterized by imposing an order to the clusters, meaning that, the clustering discovery process works by deriving, hierarchically,

clusters from other clusters. As it is known, a hierarchy can only go two ways, from the top-down or from the bottom-up, and as such, these techniques can also be either, divisive (top-down) or agglomerative (bottom-up).

2.2.4.1. Divisive

The divisive methods start with the whole population as one cluster, recursively splitting each cluster, until a termination condition is found. By termination condition, it is to be understood as, a condition by which the method stops its process, if met, and if there is none specified, the process will only end when each cluster is composed by a unique observation/point. As for the combination of the two 'nearest' clusters, its components may vary according with the measurement technique used.

2.2.4.2. Agglomerative

The agglomerative methods start with each point/observation as a cluster, and then repeatedly combine the two 'nearest' clusters into one, until a termination condition is found. By termination condition, it is to be understood as, a condition by which the method stops its process, if met, and if there is none specified, the process will only end when every single observation/point belongs to the same cluster. As for the combination of the two 'nearest' clusters, its components may vary according with the measurement technique used.

Both these methods can use the same graphical tool to visualize the clustering that occurred during their process, the dendrogram (Figure 9). This tool represents graphically, the cluster tree created through the methods process to represent the data, having each point connect to two or more sequential points (each point represents a cluster). The Y-axis represents the distance between points (clusters), the longer the line (Y-axis wise) the more distant/different the points/clusters that it connects. The X-axis represents the points (clusters), and their ancestry, points which form them or which it forms.

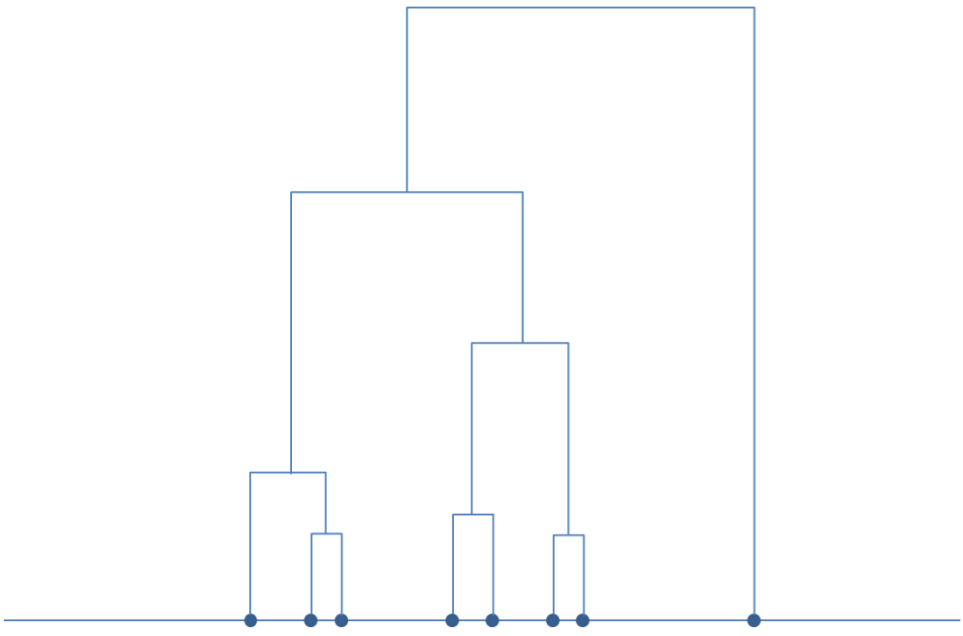

Figure 9 - Dendrogram example (Source: Author)

The hierarchical clustering, having the characteristics that it has, imposes onto itself some limitations: Firstly, it requires an a priori definition of the number of clusters, on the dataset processing through algorithm; Secondly, once an iteration is performed (either splitting or merging), a retrocession is no longer possible. It is to be noted that this rigidity has its benefits, because it allows for an avoidance of the combinatorial different choices' computational costs; Lastly, there is the need of calculation, which imposes in a lack of capability when it comes to the processing of large data sets.

2.2.5. Partitional Clustering

Partitional clustering is a group of techniques that take a set of points/observations and divides them into non-overlapping subsets/clusters, where each point/observation exclusively belongs to only one subset/cluster.

This group's techniques all follow same logic: Seed Selection ->Center Adjustment->Cluster Refinement. Seed Selection is the process of selecting the clusters' centers, so that the remaining points/observations can be assign to a cluster. This can be done through a series of selection procedures, being the most common and simple, the random selection[15]; Center Adjustment, uses the selected stipulated number of clusters' seeds, to the assign the points/observations to the closer cluster's center. Sequentially adjusting its center in accordance with observations that constitute the cluster; Cluster Refinement, is the composed by the implementation of refinement algorithms whose objective is to improve the quality of the partitioning.

All the techniques belonging to this type of clustering, end up having only slight differences on the process of how to follow this logic, retaining the criterion of a good partition as: having objects closely related belonging to the same cluster (Compactness), and differ substantially from the objects from other clusters (Separability).

2.2.6. Validation Measures

The validation measures serve to quantify the extent of match between the cluster labels and the externally supplied labels, and the goodness of cluster structure with disregard for the external information about the data. Each of these quantifications, represent distinct types of measures, external and internal respectively.

2.2.6.1. Internal Measures

The internal measures, being based on the knowledge that is intrinsic to the data will be constituted by metrics that evaluate the internal context of the observations. The five most referred indexes on literature[16], BIC, CH, DB, SIL and DUNN, will be the internal measures that are analyzed in this thesis.

[15] As the names suggests, it consists on randomly selecting, a specified number of points for the set.

[16] (CE MATH, 2018)

BIC, also known as Bayesian information criterion index, has as its main objective the avoidance of overfitting. It accomplishes this by considering both the fitness of the model to the data and the model complexity[17]. It does so through the following formula:

$$BIC = -\ln(L) + v\ln(n)$$

Where:

n: Number of objects

L: Likelihood of the parameters to generate the data in the model

V: Number of free parameters in the Gaussian model

The smaller the BIC, the better the model.

Equation 11 - Bayesian information criterion index formula

CH, also known as Calinski-Harabazsz index, uses the average between and within the cluster sum of squares, to evaluate the cluster validity[18]. It does so through the usage of the following formula:

$$CH = \frac{trace(S_B)}{trace(S_W)} \cdot \frac{n_p - 1}{n_p - k}$$

Where:

S_B: Between-cluster scatter matrix

S_W: Internal scatter matrix

n_p: Number of clustered samples

k: Number of clusters

Equation 12 - Calinski-Harabazsz index formula

DB, also known as Davies-Bouldin index, has as its main objective the identification of the set of clusters that are compact and well separated[19], through the usage of the following formula:

$$DB = \frac{1}{c}\sum_{i=1}^{c} \max_{i \neq j}\left\{\frac{d(X_i) + d(X_j)}{d(c_i, c_j)}\right\}$$

Where:

c: Number of clusters

i and *j*: Cluster labels

$d(X_i)$ and $d(X_j)$: All the samples' distance, in the cluster, to their respective cluster centroid

$d(c_i, c_j)$: Distance between these centroids

The smaller the *DB*, the better the clustering solution

Equation 13 - Davies-Bouldin index formula

SIL, also known as Silhouette index, is based on the pairwise difference between and within clusters distances, validating the clustering performance[20], identifying the data structure and highlighting possible clusters. Meaning that, for a given cluster, $X_j\ (j = 1, \dots, c)$, the silhouette technique assigns to the ith sample of X_j a quality measure, $s(i) = (j = 1, \dots, m)$ known as the silhouette width.

[17] (CE MATH, 2018)

[18] (CE MATH, 2018)

[19] (Batistakis, Halkidi & Vazirgiannis, 2001)

[20] (Gao, Li, Liu, Wu & Xiong, 2010)

Which value represents the confidence indicator on the membership of the ith sample in the cluster X_j. This process translates into the following formula:

$$s(i) = \frac{(b(i) + a(i))}{max\{a(i),\ b(i)\}}$$

Where:

$a(i)$: Average distance between the ith sample and all of samples included in X_j

$b(i)$: Minimum average distance between the ith and all the samples clustered in X_k $(k = 1, \dots, c;\ k \neq j)$

Equation 14 - Silhouette index formula

The Dunn index's main objective is the estimation of the most reliable number of clusters that should exist, through the combination of the dissimilarities between clusters, minimum pairwise distance between objects in distinct clusters, and their diameters, maximum diameter among all clusters[21][22]. This is translated into the following formula:

$$\min_{1 \le i \le c} \left\{ min \left\{ \frac{d(c_i,\ c_j)}{\max_{1 \le k \le c} (d(x_k))} \right\} \right\}$$

Where:

$d(c_i,\ c_j)$: Intercluster distance between the clusters X_i and X_j

$d(x_k)$: Intercluster distance of cluster x_k

c: Number of clusters

The larger the values of DUNN, the better the clustering solution

Equation 15 - Dunn index formula

2.2.6.2. External Measures

The external measures being based on the knowledge that is foreign/external to the data (a priori knowledge), will evaluate context around the observations. The external measures that will be analyzed are four of the most referred indexes, F-measure, MNIMEASURE, Entropy and Purity.

F-measure, is a measure based on two information retrieval concepts, the precision and the recall. The precision represents the fraction of relevant observations among all the retrieved ones, and the recall, represents the fraction of relevant observations among the total amount of relevant observations. This translates into:

$$Recall\,(i,j) = \frac{n_{ij}}{n_i} \text{ AND } Precision\,(i,j) = \frac{n_{ij}}{n_j}$$

Where:

n_{ij}: Number of objects of class i that are in cluster j

n_i: Number of objects in cluster j

n_j: Number of objects in class i

[21] (Gao, Li, Liu, Wu & Xiong, 2010)

[22] (Raphael, Saitta & Smith , 2007)

Equation 16 - Recall and Precision formulas

Another way to view the logic behind the recall and precision is through the usage of the concepts, true positives, false positives, true negatives and false negatives. These concepts can be better understood through Table 7:

	Condition Results		
Predicted Conditions	Total population	Positive Condition Result	Negative Condition Result
	Positive Predicted Condition	True Positive	False Positive
	Negative Predicted Condition	False Negative	True Negative

Table 7 - True positives, false positives, false negative, true negative definition

In short, Table 7 describes, True positives as the correctly predicted positive values, True false as the correctly predicted false values, False negatives as the incorrectly predicted negative values, and False positive as the incorrectly predicted positive values.

Having these concepts introduced, the other way of viewing the recall and precision can introduced, and is as follows:

$$Recall = \frac{TP}{TP + FN} \text{ and } Precision = \frac{TP}{TP + FP}$$

Where:

TP: True Positives

FN: False Negatives

FP: False Positives

Equation 17 - Recall and Precision formula (using true positives, false positives, true negatives and false negatives)

In this context, the meaning of recall and precision can be translated into, the fraction of true positives among all values, and the fraction of true positives among all the positive values, both correctly and incorrectly classified ones, respectively.

Now that both recall and precision were e introduced and explained, the F-Measure can be approached. This measure tries and uses both of the previous concepts in a single measure, through the following formula:

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Where:

The $F - Measure$ values are contained between 0 and 1, inclusively ([0, 1]), being that the larger the values, the higher cluster quality. This is so because its values increase, as both the recall's and the precision's values increase, and as previously verified the higher their individual values, the better.

Equation 18 - F-Measure formula

MNIMEASURE, also known as Normalized Mutual Information (NMI), has as its objective the insurance of comparability of results from algorithms with distinct metrics[23], measuring the MNI between two objects through following way:

$$MNI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}}$$

Where:

$I(X,Y)$: Mutual information between two random variables X and Y

$H(X)$: The entropy of X

X: Consensus clustering

Y: True labels

Equation 19 - Normalized mutual information formula

Purity is a simple and transparent evaluation measure, which assigns the most frequent class in the cluster, to the cluster itself, dividing it by the cluster size. The overall purity of all the clusters is calculated by a weighted average of the purities of each cluster, which weight is obtained accordingly with the size of the respective cluster. Resulting in the following formulas:

$$Individual\ Purity\ (Pj) = \frac{1}{n_j} \max_i \left(n_j^i\right) \text{ and } Overall\ Purity = \sum_{j=1}^{m} \frac{n_i}{n} P_j$$

Where:

n_j^i: Number of objects in j that has the class label of i

n_j: Size of cluster j

m: Number of clusters

n: Total number of objects

The purer the models, the better, since it will better reflect its labels. Being that, the higher the values (closer to 1), the purer the model.

Equation 20 - Individual and overall Purity formulas

Entropy, measures the purity of the clusters' class labels, measuring the diversity in terms of class variety, within each cluster and across all clusters. It does so through the following formulas:

$$'Individual\ Entropy'\left(E_j\right) = -\sum_i p_{ij} \log_2 p_{ij} \text{ and } Entropy\ (E) = \sum_{j=1}^{m} \frac{n_j}{n} E_j$$

Where:

n_j: Size of cluster j

m: Number of clusters

n: Total number of objects

Equation 21 - 'Individual' and 'Overall' Entropy formulas

The higher the value of the entropy, the higher the variety existent (e.g. when all the objects belong to the same class, existing only one class, the entropy value will be zero).

[23] (Ghosh, Mooney & Strehl, 2000)

2.2.7. K-Means

K-means is a clustering algorithm that is very widely known and used, and as such many variations of it have surfaced along time. In this segment the original/vanilla algorithm will be explained, along with the more relevant variations of it. By relevant variations, is to be understood, variations that have been recognized and proved to improve on the original/vanilla algorithm and that have relevance to the context of this paper (explained when each variation is presented).

Before explaining the logic of the algorithm, it is necessary to introduce the notion of partition. This notion is of high importance to this algorithm, because it is required to specify the exact number of partitions that are desired before the algorithm is run. This definition has to follow a set of rules, given a dataset of n objects, and the constitution of k partitions representing each partition a cluster, the number of observations must always be greater or equal to the number of clusters ($k \leq n$). The data will then be classified into the k clusters, in such a way that each group contains at least one object, and each object belongs to one, and only one cluster.

Having introduced this notion, the k-means algorithm logic can now be explained. Its logic is straight forward, given k the algorithm creates an initial partition, which is typically random (randomly selects k observations from the dataset), using then an iterative reallocation techniques in order to improve the partition, by moving objects from one cluster into another, accordingly to their similarities. K-means can be translated into a five-step algorithm (which is illustrated and explained in the Figure 10):

1. Seed Definition;
2. Association of each observation/individual to the closest seed;
3. Calculate the centroid of the formed clusters;
4. Return to step 2;
5. End when the centroid no longer changes.

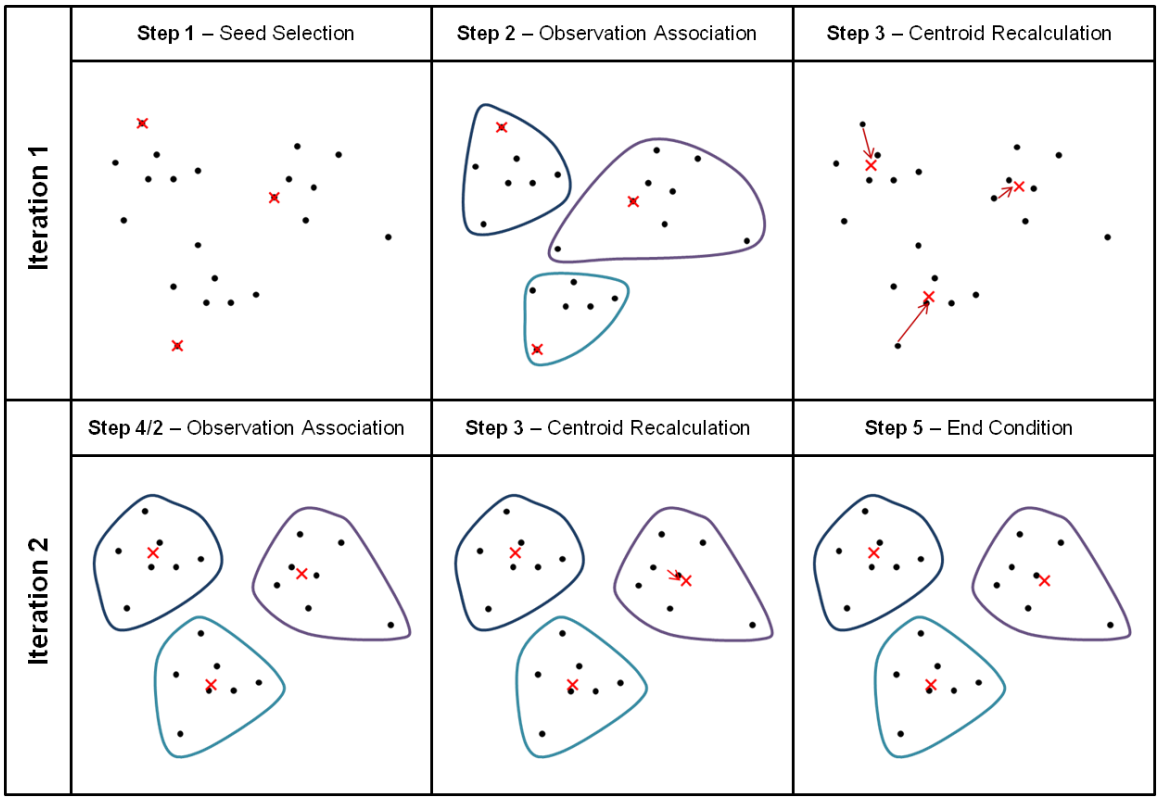

Figure 10 - K-Means operational example (Source: Author)

The figure 10 illustrates the processing of a dataset by the k-means algorithm, where k is equal to 3, meaning that the dataset is to be partitioned into 3 clusters. As it is seen the algorithm starts with the seed selection (random selection), where 3 of the dataset observations are selected and used as centroids (Step 1, Iteration1). The remaining observations are then associated to the centroid to which they are closer (more similar) forming the first three clusters (Step 2, Iteration 1). A recalculation of the previously defined centroids is made (Step 3, Iteration 1), being that in this case the centroids change, meaning that a new observation assignment is due (Step 4/2, Iteration 2). Having now a new set of clusters and centroids a centroid recalculation is made, maintaining 2 of the 3 centroids. A new reassignment of the dataset observations is made, but no changes occur, which deems the recalculation of the centroids, unnecessary (Step 5, Iteration 2). With this, the algorithm reaches its final output, the final 3 clusters.

Having understood the k-means functioning it is possible to critically look into it. It is possible to assert that the need to set the starting number of clusters, a priori to the processing of the algorithm, is a limitation, having to set an essential parameter of the algorithm just with a conceptual basis. The real problem with this limitation is that there is not a defined process in order to define the number of clusters to use, it depends deeply on the context of the dataset. As such, it is advisable to produce multiple clusters solutions with different k and choose the most adequate one. In this choice, three criterions should be considered:

The intra & inter cluster variability, where a lower and higher variability values are desired, respectively. The dendrogram and the elbow graph are both useful tools in this context, since they allow for a graphical visualization of both metrics. The elbow graph illustrates the percentage of

variance explained, as the number of clusters varies. The number of clusters chosen is usually where the line graph forms an 'elbow', point where an increase on the number of clusters stops having an effective gain on the explained variance. In the Figure 11 it is possible to observe that the 'optimal' number of clusters is between 3 and five, region on the graph where an 'elbow' is formed, analogy that originated the name to this graphical tool;

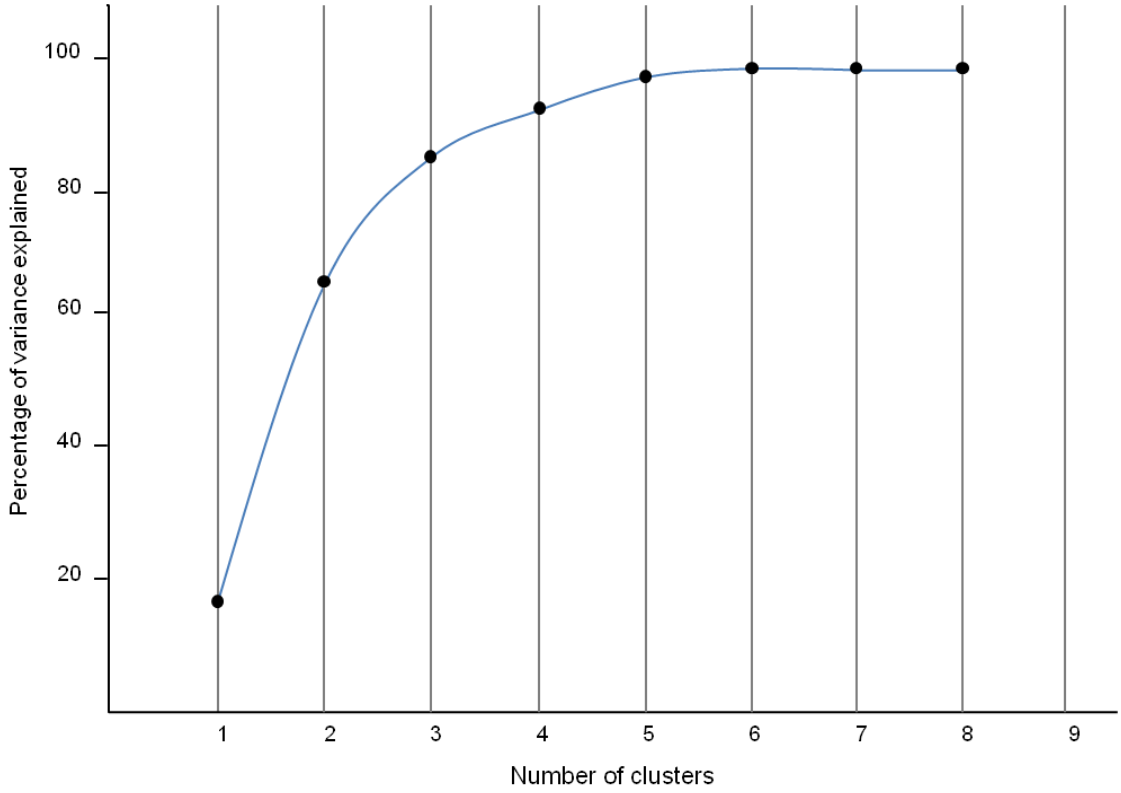

Figure 11 - Elbow graph example (Source: Author)

The conceptual evaluation of the clusters profile, in which the characterization of the clusters created, is compared to the needs of the context at hand. This criterion may seem of lower priority, but it is as fundamental as the other ones. Per example, it is known that the higher the k, the 'better' intra & inter variability values are obtained, making it so that k = n is the universal choice when setting the number of clusters to be used, but in reality is useless, because the clustering in itself would be fruitless, since the clusters would be the observations, meaning that the algorithm's output would be exactly the input;

The operational and externally imposed constraints, which may not have any kind of relevance to the algorithm or to the conceptual, has an big relevance for the practical use of the algorithm, since they are imposed restrictions that cannot be violated, regardless of being either for logistic reasons or fixed guidelines. Some considerations to have are that the number of clusters should be small enough in order to create a specific strategy and that each cluster has a large enough volume of observations/individuals that justifies the development of that same specific strategy;

A very important assumption to always have in mind when using k-means is that, it does not work properly with irregular shaped clusters, because its methodology always forms clusters of regular circular shape. This formation process also makes k-means incur in a problem in relation to its

dependence to the definition of the initial seeds, the initialization positions. This initial definition will make the algorithm build specific clusters accordingly, and throughout its iteration it may change immensely, but their relation to the initialization will remain and impact their final outputs, as it can be seen in Figure 12.

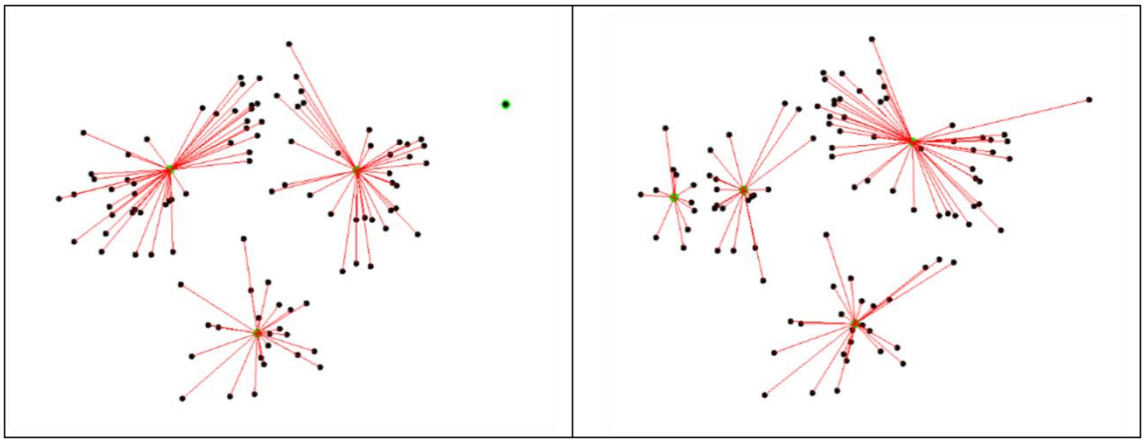

Figure 12 - Same dataset, different initializations, different results, example (Source[24])

These two sets of 4 clusters are the final output of two runs of the k-means algorithm on the same dataset, but with different initialization seeds, and as it is possible to observe, the discrepancy on the outputted clusters is evident, existing even in the case of the right, a cluster constituted by only one observation, its centroid. This shows how much of an impact the initialization can have on the result of the k-means algorithm. This can be mitigated by using multiple forms of initializations, being it different seeds or being it different initialization processes, besides the random one.

Another point to have caution about is the high volatility and sensibility that the k-means algorithm has in relation with the existence of extreme observations, observations that distinguish themselves from the remaining dataset, also known and referred as outliers. This can be justified by the way that the k-means calculates and obtains its centroids.

Having this being said it is possible to extrapolate that the k-means algorithm does not guarantee that its final output will be the global optimum, but guarantees instead that it is the local optimum in accordance with the initialization that it has and accordingly with the used dataset distribution (if it has as ideal clusters, regular shaped ones).

2.2.8. K-Means ++

K-means ++ is a variant of the previously presented k-means algorithm and tries to work around/mitigate one of its predecessors' greater limitation, its dependence on the initial seed definition, the initialization. All of its clustering logic is the same as the k-means, differing on the initialization step, not relying simply on the random selection but improving on it.

This improvement on the randomization of the initial selection comes in the form of a modification/fine-tuning in the chances of selecting observations to become seeds, accordingly with

[24] This image is constituted by the final part of two animations on the k-means processing process illustration, these can be found and viewed on, http://shabal.in/visuals/kmeans/right.gif (left side), and http://shabal.in/visuals/kmeans/left.gif (right side) (2018)

their respective distance to the already defined seeds. It fine-tunes these changes, making it so that, the observations that are further from the pre-existing seeds have a higher probability of being chosen to be a seed, and lowers it to the closer ones. This process can be defined as follows:

1. Selection of a purely random observation to form the first seed;
2. Select a new seed, by choosing an observation from the dataset with the probability:

$$\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$$

Where:

x: Observations belonging to the dataset

X: Dataset

$D(x)^2$: Squared distance of observation x to the closest seed (from the already defined ones)

The more distant the observation is from the closest seed, the better, since the probability of getting selected will be higher.

Equation 22 - K-means ++ seed selection probability

3. End when all the initial seeds have been created.

This is the only difference between the k-means ++ and its predecessor, being that, once terminated the initialization, the algorithm proceeds in the exact same fashion its operation/functioning. This might seem as a minor tweak, but as it is already known, the initialization is a major limitation in k-means, so this change can have, and in fact in the majority of the cases has, a very positive impact on the potential and efficiency of the algorithm, improving both the results' quality, the algorithm's processing time and the usage of computational resources[25].

2.2.9. K-Medoid

The k-medoid algorithm is very similar to the k-means' one, differing only on the clusters' center used in their process. It is already known that the k-means uses centroids as its cluster centers, but the k-medoid algorithm uses medoids. This is the only difference between them, having the remaining logic the same. This difference makes the more robust a resilient o the existence of outliers and noisy data, but as tradeoff comes an increase on the computational cost[26]

2.2.10. BIRCH

BIRCH, Balanced Iterative Reducing and clustering using Hierarchies, is an unsupervised algorithm that uses the hierarchical clustering logic. It has the ability to find 'good'[27] clustering solutions within a single scan of the dataset, making it a good option for when the processing of large datasets and/or streaming data is a necessity. It does so through the usage of a small set of summary statistics to represent a larger set of data, being this feature, one of BIRCH's distinction factors.

[25] (Arthur & Vassilvitskii, 2006)

[26] (Rajarajeswari & Ravindran, 2015)

[27] (Livny, Ramakrishnan & Zhang, 1996)

BIRCH can be said to have two main phases in its data clustering process, firstly building the cluster feature tree (CF tree), where the data is loaded into memory, and secondly, applies an existing clustering algorithm to the results of the first phase, the leaves of the CF tree. The first stage is the core of the BIRCH's algorithm mechanics, being at this stage where the summarization of the data will take form.

Before diving into the logic behind this phase, the introduction of some basic notes, in the context of the algorithm, are due. The term cluster feature (CF) is used to refer to the summary statistics which are an enough representative of a set of data. These summary statistics are namely, the number of observations in the cluster that the CF will represent (count); the sum of the individual coordinates (linear sum), so that a measure of the cluster's location is possible; and the sum of the squared coordinates (squared sum), to provide means of measuring the spread of the cluster. It is also possible to refer to the combination of the linear and squared sum, as mean and variance of the observation, since they are equivalents in this context. A CF can be represented as $\{n;\ LS;\ SS\}$, where the n is the count; $LS$is the linear sum; and SS is the squared sum.

There is a mechanism in the data clustering process logic of the algorithm that requires the merge of clusters under certain conditions. For this the BIRCH's algorithm states the Additivity Theorem[28], to sustain the merge of the clusters in the form of CF, by simply adding the items in their respective CF trees. This merging process can be represented as:

$$CF_{12} = \{nCF_1 + nCF_2;\ LSCF_1 + LSCF_2;\ SSCF_1 + SSCF_2\}$$

Where:

CF_{12}: Outputted cluster from the merging process of CF_1 and CF_2

nCF_1: Number of observations in the CF_1

nCF_2: Number of observations in the CF_2

$LSCF_1$: Linear sum of CF_1 (of each coordinate)

$LSCF_2$: Linear sum of CF_2 (of each coordinate)

$SSCF_1$: Squared sum of CF_1 (of each coordinate)

$SSCF_2$: Squared sum of CF_2 (of each coordinate)

Equation 23 - BIRCH's Merging Process

These operations will keep on happening, throughout the cluster feature tree building process. To better understand this process, a better understanding on the notion of cluster feature tree is due. A cluster feature tree is, as its name states, a tree structure composed of clusters features, being a compressed form of data which preserves the clustering structure, through the usage of the three elements, the branching factor B, which determines the maximum number of children allowed in a non-leaf node; the threshold T, which represents the upper limit of the cluster radius' in the leaf node; and the number of entries in the leaf node L.

As a side note, there are three elements in a cluster feature tree, a root node, a non-leaf node and a leaf node. A root node is, as its name says, the node that serves as basis for the entirety of the cluster feature tree, generating direct or indirectly all the other nodes; a non-leaf node is a node that is originated from the root node and gives origin to the leaf nodes, being an intermediary node; a leaf node is a node which is the final node in the CF tree structure, being the end result of the tree.

Now that the fundamental notions have been presented, the logic behind the clustering process can be explained. The algorithm starts by comparing the location of each record with the location of each

[28] (Harvard, 2014)

CF on the root node through the usage of the linear sum or mean and passing the incoming observation to the closest CF of the root node. Proceeding then, to the comparison of the location of the observation with the location of each non-leaf node CF, passing the incoming observation to the closest non-leaf CF. And lastly, compares the location of the observation to the location of each leaf node CF, tentatively passing the incoming observation to its closest leaf node CF. The algorithm depending on the respect of the threshold T by the radius of the chosen leaf (including the new observation),will either assign the new incoming observation and update the statistics of the corresponding leaf and its respective parent, to account for the new data, in the case that the threshold T has not been exceeded; or form a new leaf, consisting only of the incoming observation, and update all of its parents' statistics to account for the new data, if the threshold T was not respected by the radius. If the algorithm tries to create a new leaf, as stated in the previous sentence, and the number of number of entries in the leaf node has already been met, the leaf node will be split. This split is done, by using the most distant leaf node's CFs as leaf seeds, having the remaining CFs assign to whichever node is closer. If the parent node is also full, split the parent, being process throughout the whole parent hierarchy.

In this process logic each CF may be viewed as a sub-cluster with centroid $\bar{x}$ and radius R:

$$\bar{x} = \frac{\sum x_i}{n} \qquad R = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Where:

- $\bar{x}$: 'Centroid' (average of all data points belonging to the CF (mean))
- R: 'Radius' (variance of all data points belonging to the CF)
- x_i: Data point i of the CF
- n: Number of data points in the CF

Equation 24 - Centroid and Radius (Sub-cluster)

As such, the radius of a cluster can be calculated without knowing the data points as long as n, linear sum LS and squared sum SS are known:

$$\sum (x_i - \bar{x})^2 = SS - \frac{(LS)^2}{n}$$

Where:

- $\bar{x}$: 'Centroid' (average of all data points belonging to the CF (mean))
- x_i: Data point i of the CF
- n: Number of data points in the CF
- LS: Linear sum of CF (of each coordinate)
- SS: Squared sum of CF (of each coordinate)

Equation 25 - Cluster Radius

T allows the CF tree to be resized when B or L are exceeded, allowing for the observations to be assigned to the individual CFs, reducing the size of the tree.

Once this process ends, the final set of sub-clusters is obtained, and the second stage of the BIRCH's algorithm can start. Here any clustering algorithm may be applied, since the sub-clusters (CF leaf nodes) will be the observations used in it, and since there are many fewer sub-clusters than observations, the clustering will be much 'easier' to any clustering algorithm, improving both the results and the computational resources' usage[29].

[29] (Livny, Ramakrishnan & Zhang, 1996)

It is to be noted that, when a new data value is added, the summary statistics representing the larger set of observations is easily updated which allows for an efficient computation. The clustering solution may depend on the order of the data records due to the tree structure, but this can be overcome by running over different randomly sorted datasets and then looking for a consensus.

2.2.11. CURE

CURE, Clustering Using REpresentatives, is an unsupervised algorithm that uses the hierarchical clustering logic. Previously, this kind of logic was referred as a clustering discovery process that works by deriving, hierarchically, clusters from other clusters. This is generic description is a perfect fit to the CURE algorithm, all that remains is the criteria that it uses to derive the clusters. To do this, the CURE algorithm starts by selecting a constant number of well scattered points in the clusters, in such a way so that they capture the shape and dimension of the cluster to which they belong. The algorithm then proceeds to move these points towards the cluster's center by a predefined and fixed extent. These repositioned points are then used as representatives of the cluster.
Having the way to represent each of the existing clusters defined, it is possible to obtain the closest pair of representative points, and once obtained, the clusters to which they belong are merged. This is done at each iteration/step of the algorithm, until the final set of clusters is defined.

CURE can be viewed as a mix of the all-points and the centroid based approaches, but with the advantage of being a version that mitigates the short comings of both. The fact that CURE uses multiple representative points, allows for the discovery of the delimitation of non-spherical clusters, not being restricted to regular shaped clusters. It is also less sensitive to outliers, since the movement of the representative points, towards the cluster's center, mitigates their impact, by shortening their distance to the mean (cluster's center). Since this movement is done through the usage of a pre-defined and fixed parameter, throughout the processing of the algorithm, CURE creates the possibility to fine-tune the 'type' of clusters that it will reproduce and use. If the degree of movement is set to be 1, CURE becomes identical to a centroid based algorithm, if set to 0, turns into an all-points approach. By adjusting this parameter, between the values of 0 and 1, the kind of clusters to be formed will be defined, through degree similarity to one of the referred approaches.

Now that the criterion to derive clusters has been presented, and all of its inherent advantages, a crucial point has to be addressed, an essential characteristic of CURE, the ability to efficiently work with large datasets. This characteristic comes as a result of an input size reducing mechanism. This reduction can be achieved by efficiently implementing a random sampling procedure, through which a randomly picked subset of data will used in the CURE's algorithm instead of the entirety of the dataset. This will improve on the algorithm's execution time since the data that it will be using will decrease considerably. This reduction can also improve on another point, the outlier filtering, since the outliers are sporadic observations, by selecting a subset of the dataset, the probability of selecting an outlier is minimal, reducing the chance of them impacting the clustering process. This approach has some negative implications, it can be said that with the usage of only a subset of the dataset, information on certain clusters can be lost and missed, leading to inadequate defined clusters or even clusters left undiscovered. This argument is a valid one, saying that when using random sampling a tradeoff between efficiency and accuracy occurs, but in most of the datasets and with a moderate sized sample, the real impact on the accuracy of the clustering results is minimal,

being a very positive 'trade'. All that needs to be ensured for this to happen, is that the sample must contain more than fraction f of the total number of observations when:

$$f|u|, with\ 0 \leq f \leq 1$$

Where:

u: Cluster identifier

f: Fraction of observations/points

Equation 26 - Size of sample conditions

Concluding that the wanted fraction of known points in each cluster must be the same fraction of observations from the dataset, to be selected and included in the sample[30].

Having acquired all this knowledge on the Cure algorithm, it is possible to extrapolate in a reasoned fashion that with the advance of the algorithm's processing, the separation between clusters decreases and the observations' distribution, the ones that constitute each of the clusters, become more spread, so a need of dimensionally larger samples increases, so that it is possible to ensure that the distinction between the clusters can be made accurately. But as it was previously said, an unreasonable increase of the input data size has an impact on the clustering results, if the previously presented conditions are not followed. But there are cases where these conditions are not feasible, so a simple partition scheme is advisable in these cases.
The general logic of this partition scheme is to divide the sample dataset in a set number of partitions (p), in such a way that each partition has the same size has any other ($\frac{n}{p}$). A partial clustering procedure is then run in each partition until the final number of clusters in each partition is of ($\frac{n}{pq}$), where q is a constant larger than 1. The cluster merging process can also be stopped by another condition, if the distance between the closest clusters to be merged increases above a pre-defined threshold, this at each partition.
When the $\frac{n}{pq}$ clusters have been generated in each partition, another clustering procedure on the $\frac{n}{p}$ partial clusters, for all the partitions, is made.
With this, it is now possible formally define the CURE algorithm, being it defined as set of 6 steps:

1. Random sample selection;
2. Sample partition;
3. Partial cluster partition;
4. Outliers filtering/elimination;
5. Partial clusters' clustering;
6. Data labeling in disk.

2.2.12. RFM

R (Recency) F (Frequency) M (Monetary) is a particular clustering algorithm, that bases its logic on the three principles, and their respective assumption: recency, the more recently people bought something the more likely they are of repurchasing; frequency, the more often people bought something the more likely they are of repurchasing; monetary, the more money people have spent the more likely they are of repurchasing. Note that, since the context of this thesis' is AML/TF and not marketing, these principles will be adjusted, like, replacing the 'acquisition/bought' action by the 'suspicious' actions.

[30] (Guha, Rastogi & Shin, 1998)

There are two variants of this algorithm, the hard coding and the exact quintiles. The hard coding approach divides the categories based on reference values, which implies a costly implementation in terms of programming, since the categories also tend to change with time, and in order to accompany these variations, continuous changes on the code are a frequent must. Another consequence of this is the disparity of the cells' sizes, since the division is not adjusted to the data, the data will not be equally distributed across the categories. The exact quintiles approach starts by ordering the database from the most recent to the oldest, in terms of recency, and divides it into 5 equally sized segments, following the same process for the two remaining variables, frequency and monetary value, hierarchically. By hierarchically, is to be understood, that the frequency division, will order and divide each of the recency division's 5 segments into 5 each, resulting into 25 segments, and the monetary value division will follow the same principle, ending the process with a segmentation of 125 cells, of the same size (5*5*5).

2.2.13. Decision Tree

Decision tree is a supervised algorithm that can use both a regression and classification logic. A tree is a finite set of one or more elements, called nodes, which in themselves can be empty. Structure wise, a tree is composed, by a specially designated node called the root, which can be considered the starting point of the algorithm process, and by the set of the remaining nodes, partitioned into $0 \leq n$ disjoint sets $T_1, \dots, T_n$, being each of these sets a tree, more precisely, a sub tree of the root node.

The nodes will be structured in a parent-child relationship, in which an edge exists and two nodes are connected to it, the node that is closer to the root is the said to be the parent of the more distant one, and the most distant is the child of the closer, being that the parent node can eventually become the root node. The nodes that have all their children empty or that have zero children are referred as leaf nodes, while the nodes that have at least one non-empty child are referred as internal nodes. Having a tree with a sequence of nodes $n_1, \dots, n_k$, in such a way that n_i is the parent of n_{i+1} for $1 \leq i < k$, the sequence can be referred as the path from n_1 to n_k, with a length of $k - 1$.

In such a structure there are three interesting notions, depth, height and levels. The depth of a node is the length of the path from the root of the tree to the node, being the number of connections between the nodes 'leading' from node. The height of a tree is one more than the depth of the deepest node in the tree, being the number of levels of a tree. Each level is composed by all the nodes of the same depth.

As to illustrate such tree, refer to Figure 13.

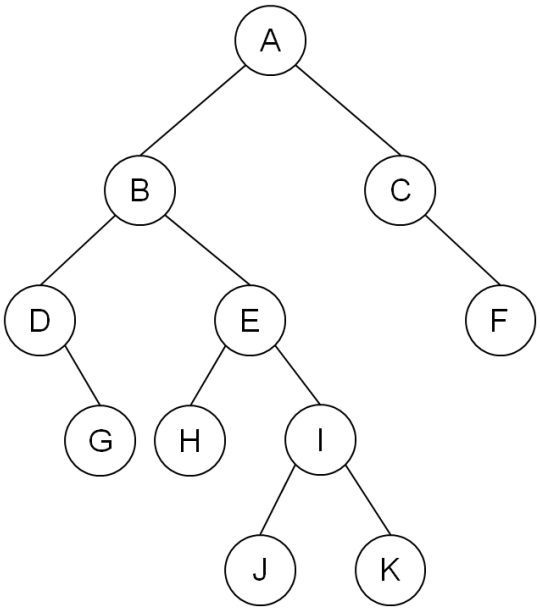

Figure 13 - Decision Tree (Source: Author)

In the previous figure, it is possible to observe a tree constituted by the nodes A, B, C, D, E, F, G, H, I, J and K. In its structure the node A is the root, with B and C as its children. The nodes C and F together form a sub tree, in which, C has two children, its left child is an empty tree and its right child is composed only by the F node. This is the case because the tree is binary[31], otherwise it could have been that the existence of empty children was impossible, it all depends in the tree at hand (type).

This tree has a height of 5, meaning that it has five levels, by which all its eleven nodes are distributed. The nodes G, H and I compose the entirety of the level 3, with a depth of three, and node A Level 0 with a depth of zero.

Structure wise, the tree has as its leaf nodes G, H, J, K and as its internal nodes A, B, C, D, E and I. In such case nodes A, B, E and I are ancestors of K, forming a path of length 4 from A to B to E to I to K.

A tree, in order to be a valid decision tree, must ensure that no path is cyclic, meaning, there cannot be more than one path leading to a node (Figure 14).

[31] Binary trees, are trees that have all of its nodes with at most two children.

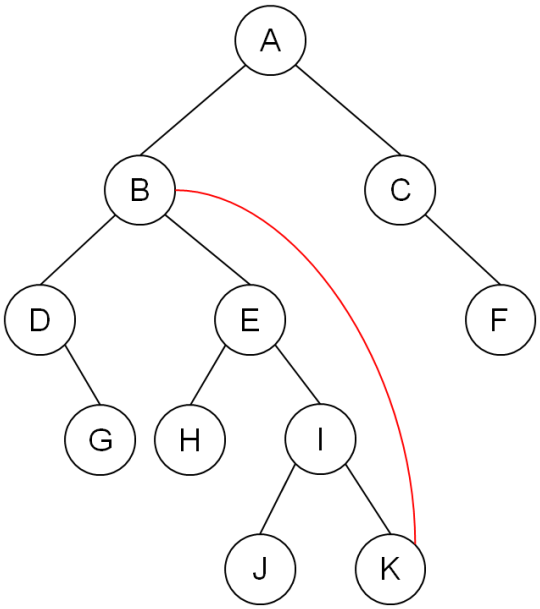

Figure 14 - Decision Tree Incorrect Structure (Source: Author)

There are times when a run through all the tree's nodes, in order to perform a specific action at each node, like the print of its contents, is a necessity. Such process is called transversal, and in the cases where this visit is done only one time per node is called enumeration. This process of visiting the nodes can follow different orders, being the most known ordering sequence: The preorder traversal, where any node is always visited before its children; the post order traversal, where any node is always visited after its children; the in order traversal, where for each node, first the left child, including its entire sub tree, is visited, then the node itself is visited, and finally the right child, once again including its entire sub tree, is visited.

Now that all the main specifications of the tree structure were introduced, the logic behind the decision tree can be explained in a sustained way. To begin with it is necessary to make a distinction between classification and regression decision trees. The difference resides on the type of data that is used and that is outputted by the algorithm, being that, if it has a categorical target it is a classification tree, and if it has a continuous target, it is a regression tree. In both types of trees, each node represents a decision, by which the data that it has been feed will be divided and 'sent' to further nodes.

A decision tree can be used as graphic representation tool, or as an intelligent algorithm. As a representation tool, it illustrates a set of results, and the decisions made in order to reach such result, having a fixed non-growing form. As an intelligent algorithm the tree estimation will depend on data, being that there are 3 criteria to be stipulated in order to define the tree estimation process, the splitting criteria, responsible for the identification of both the variable to be used in each decision, and the value of this variable in order to result in the division of the possible outcomes of the decision; the stopping criteria, that is responsible for the definition of when should the tree stop growing; and the assignment criteria, that has the responsibility of identification of the class to which the leaf node will belong to.

The splitting criteria have as its most basic function the evaluation of multiple splits and the selection of the best one. It can do this by making use of the notion of purity, characteristic of a highly

homogeneous group, and impurity, characteristic of highly heterogeneous group, these were topics already introduced in this thesis so to better understand them refer to the Literature Review chapter. The split chosen will then feature the higher reduction of the impurity possible, and in order to quantify the impurity the entropy can be used (other mechanisms of quantifying the impurities can be used, but for this case the entropy was chosen), to better understand entropy refer to the Literature Review chapter. With this the best possible split can be calculated and chosen, all in accordance with the metrics used.

The stopping criteria have the responsibility to stop the growth of the tree, or at least, define when it should be stopped. If it is not defined the tree will grow until no more splits are possible, meaning that each observation will have their own leaf node, which is not desirable, since in this case the tree will fit perfectly the data, fitting all of its noise and idiosyncrasies, and being a clear case of overfitting (Performing extremely well for the data that it used to be trained but it will not be efficient for other data sets). If the stopping criteria defined is to conservative, having the tree growth at a very preliminarily stage, the outcome will precisely the opposite, the tree did not grow into a complex enough structure so that it can model the data relationships, a case of under fitting. To correctly define the stopping criteria a middle term between these two notions is required, and preferably not just a middle term but most optimal point, and to do this there two main approaches: the gain delimiter, and the post-growth-prune approach. The gain delimiter approach required the stipulation of a gain threshold so that if the growth does not result in the gain value defined the growth process will come to an alt. The gain is obtained by subtracting to the parent's node (the top node) entropy, the entropy of the sub nodes that originated from it, taking into consideration the proportions of each sub node. The post-growth-prune, as its name states, is an approach that fully grows a tree and then prunes it to its up most optimal size, in accordance with the context needs. It does so by dividing the data set into two, a training set, that will be used to train the tree and for its growth, and a validation set, that will be used to define as the optimal size of the tree, being the first one larger than the latter, usually the proportions used are 70% and 30%, respectively. After the creation of the tree using the training set, and after running the created tree with the validation data, an elbow graph (for more detail refer to the Literature Review chapter, Elbow graph reference) must be built using the results of the run with the validation data, but instead of using as metric the number of clusters, it will be used the number of nodes. With this it is possible to identify where the optimal number of tree nodes is, proceeding then to the prune of the tree until that defined point.
The assignment criteria are very simple to specify, being based on the majority class in each leaf node, meaning that the class that is in majority in the final node will be the class attributed to the node, to that result.

2.2.14. Ensemble

Ensemble is an algorithm that can be seen as an aggregator of other algorithms. This is such because it uses the structure and outputs of other algorithms in other to develop a structure that is composed and/or was impacted heavily by them, being that it can be viewed as a combination of algorithms parts.
The ensemble starts by drawing various samples from the original data set, being that each sample can different both in terms of variables and observations selected. It then uses each of the drawn

samples to run an algorithm, using each of the resulting classification in an averaging mechanism to obtain an overall classification, the final output, the ensemble output.

Bootstrapping is a widely used and accepted sampling mechanism, which uses sampling with replacement on the original dataset. This means that a sample originated from this mechanism can have multiple replicated observations. Considering that the population of the original data is of n, the probability of an observation being sampled is of $1/n$, and consequently the probability of not being sampled is of $1 - (1/c)$. Taking n samples from the original dataset of size n, using bootstrapping, the fraction of observations not sampled is $(1 - (1/n))^{n}$, and as n moves towards infinity this is equal to 0.368 or (e^{-1}), having one observation 0.368 probability of not appearing in the bootstrapping sample, and consequently 0.632 probability of appearing. Having chosen the bootstrapping sample as the training set, and all of the remaining observations as the validation set, the combined error estimate is of 0.368 Error (Training) + 0.632 Error (Validation), which results in an attribution of a higher error to the validation set.

Having defined the samples to be used, and if possible, defining the optimal ones, the averaging mechanism will calculate the final output. There are three averaging mechanism that will be introduced in this thesis, the bagging, the boosting and the random forest, being these three mechanisms widely accepted and used.

The bagging mechanism starts by taking n samples from the original data, for this particular case we can consider n bootstrapping samples and will use each of these samples to be run in an individual algorithm. For the case of classification, the bagging uses the majority voting schema where let all of the built algorithms vote, being that each individual vote can have different weights depending on the accuracy of each individual algorithm (in case of a tie, the majority will be decided arbitrarily). For the regression case an averaging of the algorithms' outputs is made, being that the standard deviation and confidence intervals can be used in order to further fine tune the averaging. As seen the bagging mechanism uses averaging and majority logical processes in order to output its result, and as such it thrives in a context of instable algorithms[32].

The boosting mechanism builds classifiers using a weighted sample of the training data set, where a reweight of the training set is done iteratively, in accordance with the loss functions such as the classification error. The correctly classified observations will get a lower weight and the incorrectly classified ones will get a higher weight, shifting the mechanism attention towards the observations that are harder to classify. Being the final model a weighted combination of all individual classifiers. Multiple variants of this mechanism have sprung into existence, but one of the most popular is the Adaptive Boosting, also known as Adaboost[33].

The random forest mechanism starts an iterative procedure, where at the start of each iteration a n sized sample, like a bootstrap sample, is drawn from the original dataset (that has n observations and m inputs (a common choice for the m is 1 or 2)), meaning that a sample size is the same as the original dataset, but since it can have multiple replicated observations the sample does not have to be equal to the original dataset. The individual is then constructed, being that at each of its nodes

[32] (Breiman, 1994)

[33] (Freund & Schapire, 1997)

there is a random choice of m inputs on which is based the decision of spit. Considering all the splits of the m inputs, the best one is chosen, being that tree grown without pruning[34].

2.2.15. Artificial Neural Networks

Artificial neural networks can be both unsupervised and supervised neural network algorithms, being also able to model both, regression on continuous variables, as well as, classification on categorical variables. This algorithm got its name from its similarity to the structure of the brain, where there are multiple layers of neurons, that individually are very simple entities, but with the complex interconnectivity that exists between them and between their layers', they can output complex problem-solving solution, by f being feed simple inputs. The power of the neural networks comes from the connections that exist between the neurons (layers), and the connectivity that these layers have between themselves. Before diving in the artificial neural networks' intrinsic logic, two perspectives on them must be introduced, so that a sustained understanding of them is reached by the end of the chapter.

Firstly, a neural network can be viewed as a linear regression model, this is a view where a neural network of only one neuron is the same as a linear regression (including the identity transformation function). This view is based on the fact, that considering a linear regression model as the usage of, a dependent/target variable Y and dependent variables $X_1, \dots, X_n$; β_0 as the value of Y where all the $X_1, \dots, X_n$ are 0; The other $\beta_1, \dots, \beta_n$ as the values used to obtain the gradient; the objective of setting the parameters of β in such a way that the sum of the squared errors is minimized, resulting in a unique set of coefficients (where they are put in a $\beta^{\wedge} = (X^T X)^{-1} X^T Y$); the optimization using the standard errors (enabling the computation of the confidence intervals and the performance of hypothesis tests for the regression coefficients).[35] The parallelism with a central processing entity, which performs two tasks, firstly the calculation of the weighted sum of inputs ($X_1 * \beta_1, \dots, X_n * \beta_n$), with the addition of the β_0; and secondly, uses a transformation function (e.g. $f_{(z)} = z$) on the weighted sum; can be made.

Secondly, a neural network can be viewed as a logistic regression, and in this regard it is to be considered that the estimation will be of a classification model, per example of a binary target 0 or 1. In this case the linear regression would not work without a problem, since the residuals (deviations of the dependent variable observations from the fitted function) would not be normally distributed, and from the result would be unbound to the classes, different from values [0; 1]. As such a bound function ($f_{(z)}$) is used to set the value of the output (Y) to [0; 1], and in by doing so the calculation of the results probabilities would be possible (per example $f_{(z)} = \frac{1}{1 + e^{-z}}$). Now the usage of the linear model is possible, allowing the calculation of the probability of being each of the classes, being almost linear. It can become linear by using the natural logarithm. This way no distributional assumptions were made for the variables.[36] The parallelism with a central processing entity can also be made, where it performs two tasks, firstly the calculation of the weighted sum of inputs

[34] (Breiman, 2001)
[35] (Silverman, Tuncali & Zou, 2003)
[36] (Bergerud, 1996)

($X_1 * \beta_1, \ldots, X_n * \beta_n$), with the addition of the β_0; and secondly, uses a transformation function (e.g. $f_{(z)} = \frac{1}{1+e^{-z}}$) on the weighted sum. Viewing a neural network of only one neuron as the same as a logistic regression.

Having understood these perspectives on artificial neural networks, their different types of intrinsic logic can be understood and explored:

2.2.15.1. Multiple Layer Perceptron

As it was previously referred a neuron (perceptron) has two main functions/tasks, the computation of the input weighted sum and the insertion of its respective result in a transformation function (also known as activation function). A neural network is composed by a multiple neurons, and more specifically a multiple layer perceptron neural (MLP) network is composed by three interconnected layers, of intraconnected neurons: the input layer (composed by input neurons) has the multiple input variables; the hidden layer (composed by hidden neurons) has feature of extracting the features from the data, by combining the inputs in a particular way, considering the attributed weight of each connection between the neurons, and then passing the results to the output nodes(or to another layer of hidden neurons) for the optimal results; and the output layer (composed by the output neurons) has all of the targets metrics of interest of the outputs/results[37] (Figure 15).

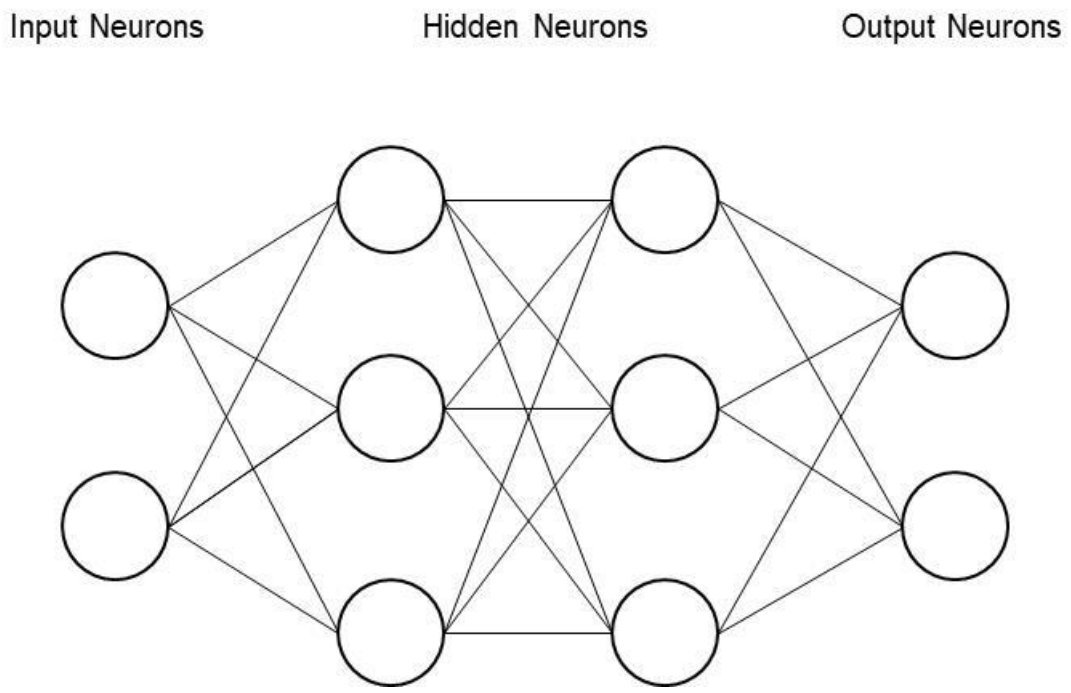

Figure 15 - Artificial Neural Network Schema (Source: Author)

The weights of the connections have a big impact on the algorithm and as such they must be attributed in an optimal way, and one of the most used methods to assign 'good' weights is the Back-

[37] (Ettaouil, Ghanou, Idrissi & Ramchoun, 2016)

propagation learning procedure. The method starts by randomly assigning small values to the weights and by running an observation through the network. The method then calculates the error of the output neuron (in accordance to an error function defined), being this error propagated to the preceding layers. Once propagated, the weights are recalculated using as basis the gradient descent[38] logic, whereas the error is minimized, evaluating the negative derivative of the error function towards the weight, subtracting to the old weight, the respective negative derivate times the learning rate. This process is then repeated until a convergence is obtained (in accordance to the convergence metric defined). The learning rate is a parameter that needs to be defined very carefully, since it has a huge impact on the weight calculation, if set to high, it will miss the minimum and start to diverge, and if set to low, it will make the learning process extremely low and not efficient. A good approach is to define it as an adaptive parameter, where it starts high, but as the calculations (iterations) are made, it decreases. The Back-propagation procedure can be run incrementally, updating the weight, observation per observation, or per batch, just after the whole dataset is read[39].

In similar fashion to other algorithms the definition of the optimal number of nodes, for this case the hidden neurons, is required, and similarly to other cases, the usage of a learning (create), validation (assess) and test (evaluate) dataset is advisable. The proportions used are usually, training > validation > test. Using different numbers of hidden neuros, in an incrementing fashion, the networks are trained using the training dataset, and are assessed in terms of performance using the validation dataset. In accordance with the assessment done on the optimal number of hidden neurons (e.g. with the usage of an elbow graph (for more detail refer to the Literature Review chapter, Elbow graph reference)), the corresponding network is run on the test dataset to be evaluated, in an independent fashion.

2.2.15.2. SOM- Konohen's Self-Organizing Maps

The Konohen's self-organizing maps (SOM) is an unsupervised learning algorithm, which enables clustering procedures on neural networks, enabling the visualization and the clustering of high dimensional data on neurons grids that have a lower dimensionality, usually two. SOM is composed by a two layered neural network, a neural network with an input layer and an output layer, where the output neurons are arranged so that they compose either a rectangular or a hexagonal two-dimensional grid. The difference in the form of the grid is the impact on the maximum number of neighbors for each neuron, for the rectangular can have up to four, and for the hexagonal can have up to six neighbors. Having this kind of structure, every input is connected to all output neurons with their respective weight, meaning that every output neuron will have their own combination of the input weights, which are initiated randomly.[40]

In SOM, when an observation is introduced to the algorithm, the weight vector (W_a) of the each neuron (a) is compared with the observation, with resource of the Euclidean distance (for more detail refer to the Literature Review chapter, Euclidean Distance reference), being the standardization of data advisable. Having now the notion of distance, it is possible to verify which of the existing neurons is the closest to the observation, being this neuron the best matching unit

[38] (Ruder, 2017)
[39] (Cun, 1988)
[40] (Miljković, 2017)

(BMU). The BMU and its neighbors will have their weight vectors updating, with the following learning expression:

$$w_i(t+1) = w_i(t) + h_{ai}(t)\,[x_t - w_i(t)]$$

Where:

x: Observation

t: Time index (training)

$\mathrm{h_{ai}(t)}$: Neighborhood[41] of BMU a, at t

Equation 27 - SOM Learning Expression

The neighborhood function is a non-increasing function of both the distance from the BMU and time. This function can take various forms, but for this particular case it is to be assumed as:[42]

$$h_{ai}(t) = L(t) + e^{\frac{\|r_a + r_i\|^2}{2\sigma^2(t)}}$$

Where:

$\mathrm{L(t)}$: Learning rate

r_a: Location of BMU

r_i: Location of neuron i

$\sigma^2(t)$: Radius

Equation 28 - Example of Neighborhood Function

It is advisable to make it so that both the radius and the learning rate decrease as time passes (as iterations are run), because such behavior will provide stability so that a coherent map is achieved after a certain period of training. This period, the period of training, is defined by putting it to a stop once the BMUs remain stable for a certain period of time, or if a certain number of iterations are achieved. By the end of the algorithm the neuron will had moved towards the observations, molding the network to the observations.

The SOM visualization interpretation is very rich in information (in this case using the Matrix U visualization), providing the average distance between the neurons and its neighbors (through a color gradient) and the weights between of each input and output neuron, allowing for an assessment of the contribution of each input to each output.

2.2.16. Association Rules

Association rules is one tool that focus on the mining of frequent patterns, meaning, it studies that variables that tend to stay together/are found combined with others, seeking to find the rules that form 'If premise, then consequence'. In order to mine association rules[43] two tasks are required, the calculation of the frequent item sets, and the mining action of the previously calculated item sets in order to achieve/uncover the association rules. For his explanation the A priori algorithm will be used[44]. Two notions must be introduced before diving into the algorithm per say, and these are the notion of support and confidence. Support is considered the proportion of observations that are

[41] Which defines the region of influence of that BMU at that given time.

[42] (Kohonen, 1990)

[43] Assuming homogeneous datasets

[44] (Deshpande & Harne, 2015)

contained on both X and Y, and Confidence the percentage of observations that are in Z containing X that also contains Y.

Having these notions introduced, the algorithm can be explained in a sustained fashion. Firstly, the algorithm will find all the frequent item sets, item set that has a frequency higher or equal to a specific threshold. In conjunction with this, the following notion is used, if an item set is not frequent then adding any item to this item set, will not change this status. Secondly, the association rules will be formed by, generating all the subsets of each frequent item set, considering that xx is a non-empty subset of x and that R is an association rule so that $(x \rightarrow xx), \ where \ (xx \rightarrow x)$ indicates x without xx, generating then R if it fulfills the minimum support and confidence requirements (doing this for every subset xx of x).[45]

2.2.17. Algorithms Final Intakes

This segment will serve as a summary of some of the practical conclusions to take from each of the previously presented algorithms.

Partition clustering algorithms, and in accordance with this thesis context, k-means, k-means ++ and k-medoid, are more suitable to work with small and medium sized datasets[46].

CURE and BIRCH work better in large datasets (in a generic case the hierarchical clustering algorithm have this particularity). In BIRCH it is not required to select the best choice of number of clusters as this is an outcome of the tree-building process, not being a pre-requirement of the algorithm. CURE has a time complexity that is no worse that the centroid based hierarchical algorithm, for datasets with lower dimensions, having in the worst case possible a time complexity of: [47]

$$O(n^2 \log n)$$

Where:

n: Input size

Equation 29 - Worse time complexity of CURE

RFM is a very simple algorithm, and its popularity derives from this. Having a very low cost for a behavior like customer classification, complemented by the capability to carry out tests in small groups representing each of its cells. In its original form, the RFM's algorithm would help companies, deciding to which customers they should be given select offers and promotional items; finding ways to increase customer spending; to target lost customers and give them incentives to purchase items; keep track of their customers and build relationship that can increase sales and productivity. But in accordance with this thesis' context, a direct parallelism of hypothesis can be made, helping deciding to which customers they should have more attention; finding the behaviors than are more suspicious; and keep track of their customers and differentiate if a behavior is suspicious or normal to a specific customer. The RFM's algorithm can only be used in customer files that contain historical information, being this point one of the algorithm's main constraints, but in the context of this thesis this can also be one of its key utility points.

[45] (Agrawal & Srikant, 1994)
[46] (Rajarajeswari & Ravindran, 2015)
[47] (Guha, Rastogi& Shin, 1998)

Decision trees as a standalone algorithm are an algorithm that is highly sensitive to the data that it used to be trained.

Ensemble, using the bagging averaging mechanism does not benefit from the usage if stable classifiers since they are robust in terms of the underlying data. The boosting mechanism having an 'attraction' towards the noisy data, using data that has a noisy target, such as data on fraud detection, might result in ineffective results. The random forest is very resilient, but outputs highly difficult results in terms of interpretation. All of these types of ensemble suffer from the same disadvantages and advantages of each other, since they all have a similar basis, differing only in the proportion of the impact (for more detail refer to the Literature Review chapter, Ensemble reference).

Neural networks have as its main constraints the existence of its multimodal error function, since it has multiple local minimums and there is no guarantee that the global optimum if the one reached; and the high level of parametrization that the algorithm has and requires in order to be optimized. On the other side the neural networks, are adaptive in their learning process, given the training data; self-organized; recognize patterns, given the training data; flexible in terms of the context where they found themselves, being able to be built regardless of the context; and in the worst case scenario they just as bad as classical statistical models.[48] More specifically for the SOM's case, the algorithm is extremely useful as a visualization tool being extremely good as a provider for initial insights on the data; automatically informs about the similarity between the neurons; has a tendency to underestimating high probability regions and overestimate low probability areas; has comparisons difficult since no objective function is defined; Does not required to have a predefined number of clusters; and requires an extensive experimental evaluation in order to determine its size [49]. Support vector machine, or SVM, is a supervised algorithm that uses the classification logic, which improves on some of the neural networks' problems (referred in the previous paragraph, from the first to the first half of the third line). If these problems happen to have a relevant impact for when the algorithms are implemented, it is advisable to explore this algorithm[505152].

Association rules are very useful in cases where there are very big datasets to be processed, but sadly incur on the curse of dimensionality, where the number of possible rules grows exponentially as the number of attributes increases, being that there are algorithms that mitigate this problem, like A priori.

As a final intake on this chapter the introduction and explanation of the 'No free lunch theorem' is mandatory. The 'No free lunch theorem', to be considered, states, as a conclusion, that if for a particular case, the algorithm X performs better than the algorithm Y, then for sure there is a case for which the algorithm Y performs equally better than the algorithm X[53]. The necessity of this knowledge is to understand that even if in theory or that for a particular case, a specific algorithm outputted an excellent result, that does not imply that for another similar case it will output an equally and/or similar result, concluding then that for each case there a necessity to test and validate.

[48] (Maind & Wankar, 2014)
[49] (Lobo, 2005)
[50] (Hears, 2017)
[51] (Chih-Chung, Chih-Jen & Chih-Wei, 2003)
[52] (Bhambu & Srivatava, 2005)
[53] (Macready 6 Wolpert, 1997)

3. METHODOLOGY/IDEA

In this chapter, the methodology used in this project will be introduced.

Firstly, an exploration and understanding of all the operating systems, terminology, and legal and operational requirements assessment is due. This will allow for a complete and efficient start, since the project context and fundamentals guidelines will be clarified and fixed.

Secondly, an analysis of the new system's capabilities and logic is done. This will allow for the definition of directives as of how and as of what is required for its correct and effective functioning.

Thirdly, a set of analysis to the customer's database, data and operational structure, and to the old system's logic and results is done. This in order to clarify, derive and define sustained guidelines and action plans for the sustained and correct implementation of the new system.

Fourthly, follows the implementation of the new system, and it's iteratively validation.

Lastly and continuously, a validation and optimization of the operational live functioning of the new system is due, as of to ensure its efficient and correct function. The definition of such validation and optimization methodologies/processes will be done transversally, after the clarification of the data available is accomplished, firstly analyzing the thematic that require such processes, secondly analyzing the data structure that is available, thirdly investigate a set of validation and optimization processes, fourthly select set of these processes taking into consideration their qualities and limitations and their degree of compatibility to the thematic and available data, and lastly implement and continuously work upon such processes.

4. PROJECT 'ENVIRONMENT'

In this chapter the context of the new system will be explained, in order to allow for a better understanding of its particularities. This will culminate in a complete understanding of the system, both in technical and theoretical side.

As it was previously said, this project is about the implementation of an AML/CFT system, more specifically, the transition from one AML/CFT system to another, but before diving into the all the details related to the change and the system in itself, a better understanding of the context/environment surrounding the system should be presented and understood.

This system will be playing a major role in the compliance office, and in all its subjacent departments, but mainly on the AML/CFT Monitoring department and Information Systems and Analytics department. This is such, because the both base their work on a set of implemented systems, whereas the first one interacts and analyze with their leads and outputs, and the latter, ensuring their maintenance and fine-tuning, as well as being the responsible for feeding them the proper inputs and settings.

This set is currently constituted by three main systems, one responsible for the market abuse analysis, other for the validation of the operations, and finally, the one that is the main topic of this thesis, the one responsible for the AML/CFT assurance.

The first system, being responsible for the market abuse detection, take as input the transactions made by the customers in the share market. It then uses a set of models and rules, that analyze these transactions in the context of the share market behavior for that given day and evaluates if they are suspicious of market abuse. If they are deemed as suspicious, an alert, that is to be analyzed by a specialist, is created. If this analysis, confirms the suspicions on this behavior, a report is filled by a jurist that is latter sent to the competent authorities[54].

The second is responsible for the filtering and validation in real time, of the account opening operations and transfers that are made though this banking institution. It does so, by comparing the attributes of the operation (beneficiary, country of origin, purpose, etc.), with a set of lists. Such lists are provided by official entities (e.g. Sanctioned countries, and Criminal organizations and individuals', lists); private certificated companies (e.g. PEP's lists); and internally built lists (e.g. Communicated entities lists). In accordance with a set of stipulated rules of attribute and perceptual match, the operations are stopped and an alerted is created. Such alerts must by validated as quick as possible, since the operations are in suspense until a decision is made, and as such, a specialist is attributed for its validation. If the alert is deemed a false positive, the operation is approved, if not, the operation is sent back to the source, and in this case, such attempt is reported by a jurist to the competent authorities[55].

[54] In order for such reposts to be sent, they must be approved in the compliance committee, which takes place one time each 15 days.

[55] In order for such reposts to be sent, they must be approved in the compliance committee, which takes place one time each 15 days.

The third, has its main objective identifying, in the most accurate and efficient way, the clients' behaviors that are suspicious of money laundering and/or terrorism financing. It does so by taking as input all the relevant transactions[56] and the customers' information, in order to profile and analyze if such behaviors are suspicious for the profile which the customer belongs to. If it deems a behavior as suspicious, the system will output an alert for the client in question, with the specific behavior's details, so that it can be analyzed by a specialist. If the suspicions are confirmed / validated by these specialists, a report is filled by a jurist, so that such case can be reported to the competent authorities[57]. These logics and processes will be further detailed in this chapter.

These systems appear to work in parallel without any interaction, and for the most part that is true, since they are responsible for distinct thematic, except for a special point in common, the fact that they are all impacted by the reported situations, regardless of the system that originated them. Both the first and third system will take into consideration these reports as a risk factor for the client and its inherent behavior, and the second will have its lists updated in accordance with the reported cases.

This way, all of the customer's base transactional behavior is validated and analyzed, ensuring that all of the duties which this financial institution is required to follow, and implement are put in action.

4.1. PROJECT OBJECTIVES

Having understood the context and the changes that this project will bring with his transition, it is hoped to see:

- An increase of flexibility and adaptability to new regulation requirements;
    - Adjustment of the current scenarios and definition of new ones.
- An increase of analytical capabilities;
    - Analyze behavior, and based on desired risk level, imply thresholds values, sustaining the decisions with data patterns and analyses.
- More efficiency on the investigation process.
    - Minimize the number of alerts generated;
    - More flexibility and optimization of scenarios and thresholds over time.

These objectives have a convergence onto the thesis objectives, denoting respect for the legal authorities requirements' and proposals'; adaptability of the system to the client behavior by developing analytical capabilities, so that the spotting of effective suspicious behaviors can achieved; And by providing the tools, the data, and the capabilities to support a continuous fine-tune of the system so that it can achieve its most optimal efficiency, across its life cyclic. Culminating this on an implementation of a fully functional and efficient behavior-based AML/CTF system.

[56] Refer to the data mapping chapter

[57] For such reposts to be sent, they must be approved in the compliance committee, which takes place one time each 15 days.

4.2. PROJECT TECHNICAL VIEW

The project technical view is a descriptive view of the technical processes that will serve as support for the correct operation of the system. The system will operate through a process with 4 stages (Figure 16):

Source Systems (The Bank)[58]

The Source Systems stage is constituted by the maintenance and forwarding of the required transactional data from the bank's databases to the next process stage.

Staging Area (SAS)[59]

The Staging Area is constituted by the load of the transactional data, from the bank's databases, on to SAS database. Once loaded, a transformation process comes into play, formatting and sending, the required data in the expected format, to the next process stage.

SAS Database (SAS)[60]

The SAS Database stage is constituted by a Core and a Knowledge database. The Core .db will receive the transformed and formatted data from the previous stage, processing it (AGP) and outputting the systems desired results, into the Knowledge .db.

'Application Front' (SAS)[61]

The 'Application Front' stage will then use the Knowledge .db in order to retrieve the required data, in order to present, in the front application view, all the required information to the user.

[58] Ensured by the Bank
[59] Ensured by SAS
[60] Ensured by SAS
[61] Ensured by SAS

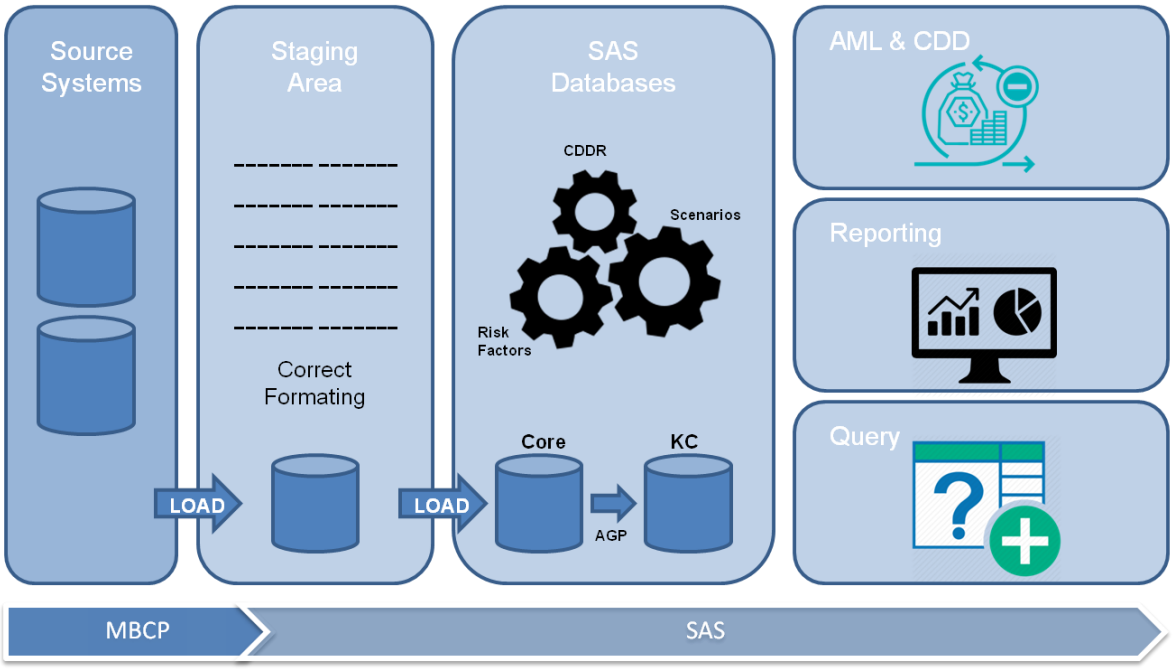

Figure 16 - Processment of the system's information (Source: Author, based on SAS documentation (2018))

The Core database, as previously induced, is used as storage for months of the required and correctly formatted transactional data, containing accounts, customers, households, transactions, and other data used in the generation of the alerts. Having this role, its structure is a dimensional one, which allows for a high level of indexation, data duplication (denormalized DBMS), data derivations and aggregations, and a high-volume of query access. Making possible, the data view through the main required measures, the transaction amounts, account, events, account, profile, party profile and trade amounts, and its sequential view/exploration throughout 5 dimensions, customer, accounts, transactions, banks, and external parties.

The Knowledge database is used to store and replicate transactions indefinitely, more specifically, defining scenarios, risk factors, risk classifiers, ETL jobs, and storage of generated alerts, cases and associated replicated facts. This database will have an Entity-Relationship type structure, which translates into a description of inter-related things of interest in a specific domain, being composed of entity types that specify the relationship that can exist between those entities. As such, this will allow for scenarios, risk factors, headers, configurations, alerts, cases, and SAR's interconnected relation.

4.3. Project Logical View

The project logical view is a descriptive view of the logical process that will be operating on top of the technical processes, which the system will follow in order to output the optimal required information. This logical process will be based on the evaluation of each client, each of the client's behavior, analyses of the suspicious client's behaviors, and validation of these processes.

The evaluation of each client is done by passing their information through a series of CCD rules and risk factors, which will classify each one, according with their characterization. The evaluation of the each of the client's behavior is done by passing their behaviors through a series of scenarios, with

specific parameters to match the client's classification and segmentation, which should and will identify the ones that fall into the suspicious category. The analysis of the suspicious behaviors is done manually by a group of specialists, which take the adequate measures and procedures, accordingly to the customer evaluation, to validate the veracity of the behavior flagged as suspicious. The validation of these processes is done via report and statistics evaluation, which will dictate possible changes or add-ons to the processes specifications.

It is to be noted that the CDD rules will be evaluating the risk of the customers in accordance with their information; the scenarios will identify the suspicious behaviors, in accordance with the context of the each client (e.g. population and risk level (attributed by the CDD rules)); and the risk factors, will evaluate the priority degree the suspicious behaviors detected, by the scenarios.

4.4. INTERNAL CONTROL

There are certain tasks/processes that due to its sensitivity and/or importance require some sort of internal control, being it either to ensure that it is done in compliance with the legislative constraints or that it is done in the delimited 'correct' way.

This is the case for the analysis of suspicious behaviors (alerts) generated by the AML/CFT system. These controls exist in order to ensure that each analysis is done correctly, not to misevaluate the suspicious behavior information leading to a misevaluation, and compliance with the imposed legislation constraints at the time of the analysis. This is done by applying the 'Four Eyes'' methodology, which dictates that each analysis must pass through a validation chain of command, obligating the analysis' validation by at least one hierarchical superior. This further increases the process transparency and efficiency, as well as ensuring an effective control and monitoring.

There is also, as alternative to the 'Four Eyes'' methodology, the sample evaluation technique. This technique has as its base concept the same idea as the 'Four Eyes'' methodology, the analyses' validation, but in contrast, instead of requiring the validation of each of the analyses, this one retrieves samples of the analyses done (from the perspective of all of the organization's elements validating the suspicious behaviors), validating each sample.

The sample evaluation technique can be particularly helpful in a context of a high suspicious behavior's count volume and low number of analyses and validation specialists, since it severely decreases the payload of each element. One possible down side of this technique is the possible miss of a misevaluation, since it is a sample that is going to be validated. This can be mitigated by ensuring that the retrieved sample is representative. But even so, the sampling risk can also fluctuate accordingly with the technique used to sample, statistical or non-statistical.

The Statistical methods outputs objective and mathematically sustained results employs the possibility of quantifying the misstatements that exist in the sample (sampling error) and ensuring the work consistency (since it is not based in empirical knowledge, it does not fluctuate accordingly with the sampling technician). On the other hand, the non-statistical method avoids the extensive exploratory population work required for the statistical method, but depends on the sampling technician knowledge degree, since it is based on it.

There are also internal audits done by a foreign department, to the compliance office, as to assure the veracity of its validations and control appliance.

This culminates into a very brief and simple conclusion, if the sampling is done correctly, the sample evaluation technique can bring only benefits to the organization, since it is able to gain reasonable

assurance regarding the entirety of the data/population. But if mal-conducted the organization ends up with an unreliable outcome.

4.5. PROJECT MULTITENANCY

The system will operate, having as its base the concept of multitenancy (Figure 17). This allows the use of the same configuration several, highly separated and independent business or organizational units. Each of the units is represented as a separated tenant has its own data schema, business configuration elements (scenarios, rules, screen extensions), business parameters, and daily operational cycle. In this project there will exist three tenants, the Portugal's tenant, with all the Portugal's bank non-employee customers' transactional behavior; the Macao's tenant, with all of the Macao's bank non-employee customers' transactional behavior; and the employee's tenant, with all of the bank's employee customers' transactional behavior.

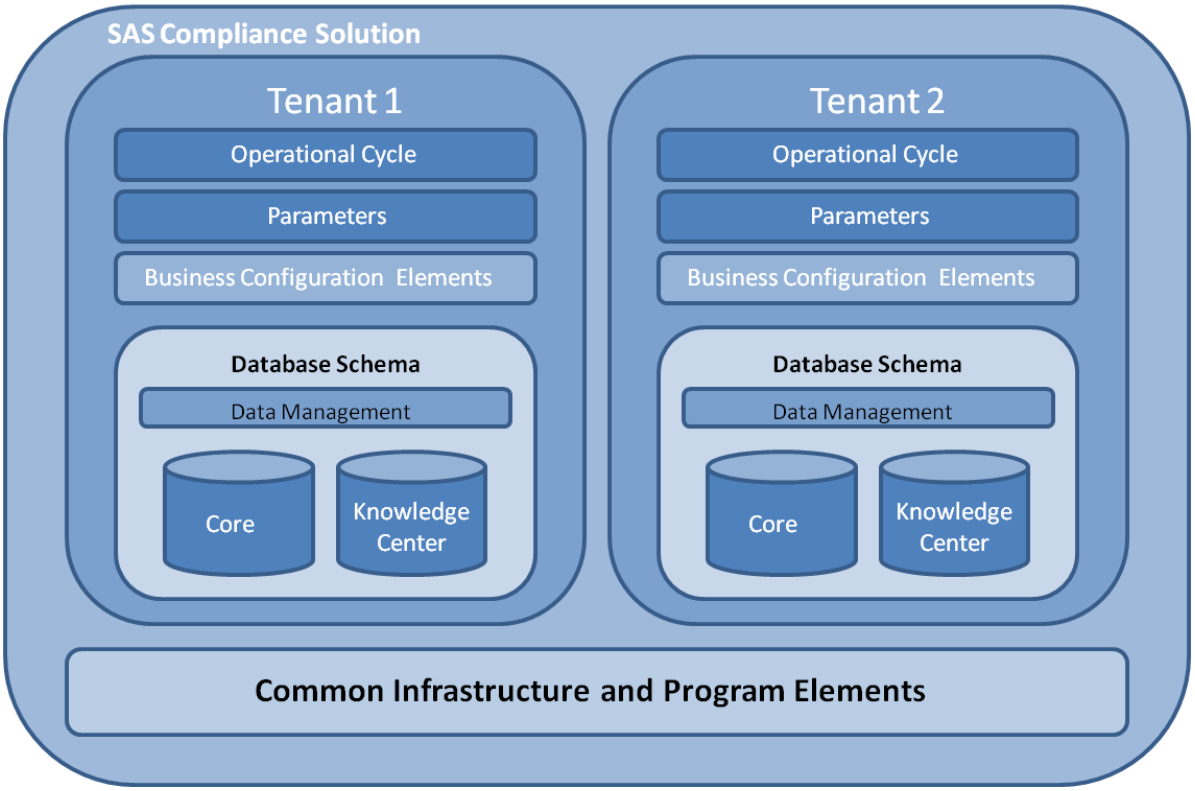

Figure 17 - System's Multitenancy (Source: Author, based on SAS documentation (2018))

4.6. PROJECT IMPLEMENTATION PLAN

The implementation process of an AML/CFT system is a complex and arduous path, being fairly easy to get lost tracked and ending up with a solution that is mismatched with the organization's needs, or even without any working solution. As to avoid this, an implementation plan was formulated (Figure 18). This plan is constituted by 6 main stages, Setup, Design, Build, UAT (User Acceptance Test), Deployment and Optimization, which impose the following:

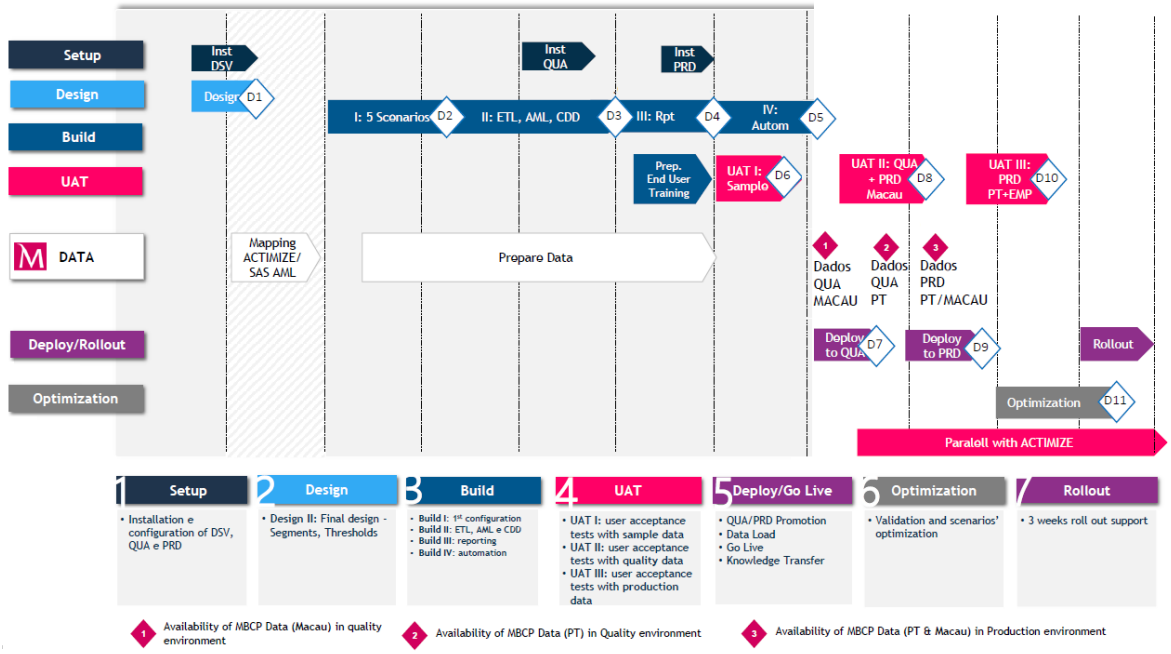

Figure 18 - Implementation Timeline (Source: SAS Documentation (2018))

Setup

The Setup stage is responsible for delimitation of the processes responsible for, the SAS AML software installation, the environments configuration (and sequential compliance with the SAS Best Practices), and the installation tests execution.

Design

The Design stage is responsible for the segments and thresholds final design, translating this into a segregation of the population into 5 main segments and a sequential organization them into three risk levels. The scenario threshold design is a process that will make use of the knowledge from the previous AML/CTF, for the cases where the scenarios have a direct equivalence from both systems. The remaining scenario cases will be set using empirical knowledge. The users' characterization is also attributed to this stage, where the assessment of the necessary permissions for each user, and the creation of the user profiles, according with the necessities unveiled. The final responsibility of this stage is the specifications design of the custom reports.

Build

The Build stage is responsible for the technical building process of the system, which still is a large and complex process (Table 8).

Build Phase 1	This phase is responsible for the configuration of the solution for a sample of 5 AML/CFT scenarios, serving as functional demo of the application.
Build Phase 2	This phase is responsible for the definition of ETL from staging to SAS AML, the mapping of parameters for the bank, the capture of risk scoring

	settings, the final settings of the AML scenarios and risk factors, the AGP definition, and the configuration and setting of CDD definitions on risk rating and rules.
Build Phase 3	This phase is responsible for the review of the out of the box VA reports, for minor changes, and for the creation of specific reports for alerts analyst analyses.
Build Phase 4	This phase is responsible for the definition of automation logic for batch processing

Table 8 - Four stage building phase

UAT (User acceptance test)

The UAT (User acceptance test) stage is responsible for the initialization of the system data processing procedure. It starts by feeding the system sample data, provided by SAS, ensuring that all the system's internal processes are working correctly and as expected. It then shifts into quality data, provided by the bank, in order to test the outputs of the system, ensuring that the system works with the organization's data information. Lastly, it changes into a production data, in order to test the system's robustness to the organization's normal transactional data.

Deployment

The Deployment stage is responsible for the process of shifting the system's orientation, from the building and testing stages/focus, to the live and operational system's cycle. This process starts, with a final transition of the data fed to the system, trading the quality oriented one for the live transactional data. This transition will not be promptly done, taking a couple of months in order to finish, ensuring that all the scenarios have enough historical data to sustain their correct functioning. Once the data transition is completed, the system goes live and into production mode. This is complemented by an introduction of the system to the end user, followed by an adjusted instruction, according with the role that the end user will have in the interaction with the system.

Optimization

The Optimization stage is responsible for the assessment, validation and optimization of the system's outputs/results. It does so, through an initial data validation where there is an insurance of its consistency among all the dimensions, and the alignment between the organization's objectives and the results' statistical analyses and derived conclusions. Making use of the data analyses to support the initial threshold setting. Complemented with the usage of descriptive statistics and clusters algorithms, in order to support the individual threshold configuration for each scenario, in case of necessity. Having ensured that the foremost optimization possible is reached, with the current available data, the system will enter its production life cycle.

This implementation plan has defined that the previous system and the soon to be implemented system will have a period of mutual operation, in order to assert the consistency of the results between both systems. This period will match the comprehended interval between the end of the UAT (User acceptance test) stage and the end of the Optimization stage.

5. DESCRIPTION OF THE TOOLS

In this chapter a breve introduction of the tools used and that have relevance for the thesis and/or project, is going to be made.

5.1. SQL SERVER MANAGEMENT STUDIO

SQL Server Management Studio is a relational database management system (RDBMS), which supports a wide variety of transaction processing, business intelligence and analytics applications.

5.2. MICROSOFT EXCEL

Excel is a spreadsheet, which features calculations, graphic tools, pivot tables and a macro type programming language (Visual Basic for Application), allowing for certain degree of data management, storing and manipulating.

5.3. R

R is a language and environment for statistical computing and graphics, providing a variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analyses, classification, clustering …) and graphical techniques, as well as a high extensible capacity (custom add-ons).

5.4. SAS® SOFTWARE

These are the AML/CFT system's components, which constitute this paper's main analysis focus, being that they are explained in a more detailed fashion in other points of this paper, so this serves only as introduction the software's main concept.

5.4.1. SAS® Compliance Solution

"*Compliance Solution (formerly SAS Anti-Money Laundering) represents the adoption of advanced analytic and investigative techniques to help compliance organizations apply a risk-based and cost-effective approach to money laundering and terrorism financing compliance. The web-based user interface (UI) supports the management, investigation, and reporting needs of anti-money laundering analysts and investigators*." (SAS)

5.4.2. SAS® Visual Analytics

"SAS Visual Analytics offers a leading business intelligence and analytics solution to visually discover relevant relationships in your data, create and share interactive reports and dashboards, and use self-service analytics to quickly assess probable outcomes and make smarter, data driven-decisions." (SAS)

5.4.3. SAS® Enterprise Guide

"SAS Enterprise Guide is an easy-to-use Microsoft Windows client application that provides an intuitive, visual interface to the power of SAS. Using SAS Enterprise Guide, you have transparent access to both SAS and other types of data. You can use this data in interactive task windows that guide you through dozens of analytical and reporting tasks. SAS Enterprise Guides enables you to export results to other applications and to the Web. By using SAS Enterprise Guide, you can analyze your data and produce great results without knowing how to write SAS programs. However, if you are a SAS programmer, SAS Enterprise Guide includes a full programming interface that you can use to write, edit, and submit SAS code. Finally, online Help, embedded context-sensitive help, and a Getting Started tutorial are all available to help you with your work." (SAS)

6. DATA AND EXPERIMENTAL SETTINGS – IMPLEMENTATION PHASE

6.1. PARALLELISM BETWEEN SYSTEMS

The parallelism between the currently operating AML/CFT system and the soon to be, is a very important concept in the course of the planning and implementation process of the latter. This is so, because there are logical points that are identical and/or complementary between systems, being possible to extrapolate certain points such as, threshold values and scenarios. Allowing for a more effective testing, not starting with just dummy or empirical setting values, but with some data sustained logic. There are also technical points that are identical, which can also be used to simplify the technical part of the implementation and built of the system, such as data mapping and connections.

The cases/implementation components where this will be used will have a short description of its objective and parallelism process used.

6.2. DATA EXPLORING

As in every conceptual and analytical process, knowledge of the current context is a necessity, so a data exploration is mandatory. This is usually the highest resource consuming stage, being it either of time or of computational resources. It is the first stage to be done, so that the project can be constructed from its conclusions, and usually continues to suffer modifications along the project implementation process.

The data exploration stage started with a side by side double analysis, from the client's and account's perspective, having started with a distribution of these across the populations' existent in the old system. These populations amounted to a total of thirty-five (Table 9).

Population	Description
CGV	Golden Visa
DINT	International Direction (Banks)
DINT-VL	International Direction (Banks) - Very Low Risk
DINT-L	International Direction (Banks) - Low Risk
DINT-ML	International Direction (Banks) - Medium Low
DINT-M	International Direction (Banks) - Medium
DINT-MH	International Direction (Banks) - Medium High
DINT-H	International Direction (Banks) - High
DINT-RG	International Direction (Banks) - Located on geographically risky areas
DINT-INV	International Direction (Banks) - Investigated
DINT-REP	International Direction (Banks) - Reported
EMP-H	Companies - High Risk
EMP-M	Companies - Medium Risk
EMP-L	Companies - Low Risk

EMP-LD	Companies - Large Dimension
EMP-MD	Companies - Medium Dimension
EMP-SD	Companies - Small Dimension
ENI-H	One person Company - High Risk
ENI-M	One person Company - Medium Risk
ENI-L	One person Company - Low Risk
ENI	One person Company
ENI-VW	One person Company - Very Wealthy
PART-H	Particular Individual - High Risk
PART-M	Particular Individual - Medium Risk
PART-L	Particular Individual - Low Risk
PART-INT	Particular Individual - Middle Man (e.g. Lawyer)
PART-VW	Particular Individual - Very Wealthy
PB-H	Private - High Risk
PB-M	Private - Medium Risk
PB-L	Private - Low Risk
PB	Private
PB-VW	Private - Very Wealthy
WU	Western Union
MACAO - EMP	Company
MACAO - PART	Particular Individual

Table 9 - Old Population Distribution

As it can be seen there is a high volume of populations, more specifically, of sub- populations, and in culmination with the specificity of the criteria necessary to belong to a sub population, the distribution of the two components in analysis will be impacted, causing it to be the following (For simplicity sake, only the segments that represent a volume of the total population higher than one, will be referred with their values):

In the client's distribution: the 'EMP-L' population represents 5.33% of the component; the 'EMP-MD' population 2.14%; the 'ENI' population 2.71%; the 'PART-L' population 88.22%; and the 'WU' population 1.11%.

In the Account's distribution: the 'EMP-L' population represents 1.37% of the component; the 'EMP-MD' population 1.83%; the 'ENI' population 4.27%; and the 'PART-L' population 75.43%.

This shows a clear uneven distribution, even more so, since each of the populations has a 'significant' part of the components, but there are a lot of sub-populations with near to no representability.

Volume does not necessarily translate into representability of the context that the population is supposed to cover, so in order to validate the context of these 'significant' populations, a set analysis was done:

Client wise, each of these populations were analyzed accordingly, through a set of descriptive variables distributions, that are as follows: Primary party identification document; Party type, if it is an organization or an individual; Party classification in the commercial context, which type of client it is; Age of client; Age of the bank and client relationship; Marital status; Occupation, activity of the organization/Individual; Citizenship country; Domicile Country; National residency; PEP identity; Number of SARs(reports); Internal risk, risk associated to the client accordingly with the bank criteria; and Employee of this banking institution.

Account wise, each of these populations will be analyzed with accordingly, through a set of distributions of the following variables: Account type, if it is an investment account, a credit card account, etc; Institutionally of the account; Account status, if it is dormant, active , etc; National residency; PEP identity; Internal risk, risk associated to the client accordingly with the bank criteria; Age of last update; Opening channel type, which channel was used in order to open the account; Initial source of funds, how the initial funding was made; and Mail type.

The conclusion that derived from these analyses was that the context of each population was indeed represented by the Individuals that belonged to it, both on the client and account perspective.

Having analyzed populations that are feed to the old system, namely the bank's clients and accounts, an analysis was made to the outputs of the old AML&TF's system, the alerts. This was firstly made through an alert volume distribution by population. In similarity to what was previously done, only the populations with more than one percent of the population representability will be referred with its values. They were as follows:

In the client's distribution: the 'CGV' population represents 1.73% of the total generated alerts; the 'DINT' population 1.07%; the 'DINT-VL' population 1.26%; the 'EMP-M' population 7.70%; the 'EMP-L' population 6.51%; the 'EMP_LD' population 6.10%; the 'EMP_MD' population 10.84%; the 'EMP_SD' population 7.55%; the 'ENI' population 5.54%; the 'PART_L' population 44.30%; and the 'PB' population 3.01%.

In the account's distribution: the 'CGV' population represents 1.72% of the total generated alerts; the 'DINT' population 1.48%; the 'DINT-VL' population 1.26%; the 'EMP-M' population 7.70%; the 'EMP-L' population 6.42%; the 'EMP_LD' population 6.22%; the 'EMP_MD' population 10.78%; the 'EMP_LD' population 7.50%; the 'ENI' population 5.54%; the 'PART_L' population 44.26%; and the 'PB' population 2.60%.

In these distributions we can see a huge similarity in the distributed volume of the alerts in each of its significant population, being these, mainly, the ones that were already significant in the context of the institution's clients, but it was also verified the appearance of other ones, mainly, sub-populations of the previously referred ones'. This is interesting, because even though there are populations with more individuals, they are producing less alerts, and those with less representatively are generating more alerts, in the total alert generation context. In order to better understand the reason as of why this is happened an analysis on the alerts generated in these populations was made.

In these analyses, a distribution of the rules that contributed for the alert generation was used, as well as, a distribution of the multiple rule combinations' that were the 'cause' the alerts, having as its universe, each individual population, to better understand the behavior being treated and validated in each population, and on the whole bank client's dataset, in order to visualize the real impact and usage of the rules and the universe of the alerts being generated. In addition to this, another layer of

information was added, the type of resolution of the alert to which the rules contributed to/caused, providing information on the veracity of their contributions.

Having all these sets of information regarding the alerts, the rules that contributed/'caused' them and their veracity (resolution of the alerts), a new layer of analyses was done, a set of analyses on the values of the rules parameters', the thresholds. This was done taking into consideration two-time intervals, one of three most recent moths and another of the two more recent years, being the latter used in order assess the behavior of the thresholds and the first to validate the up-to-datedness of the behavior (Figure 19).

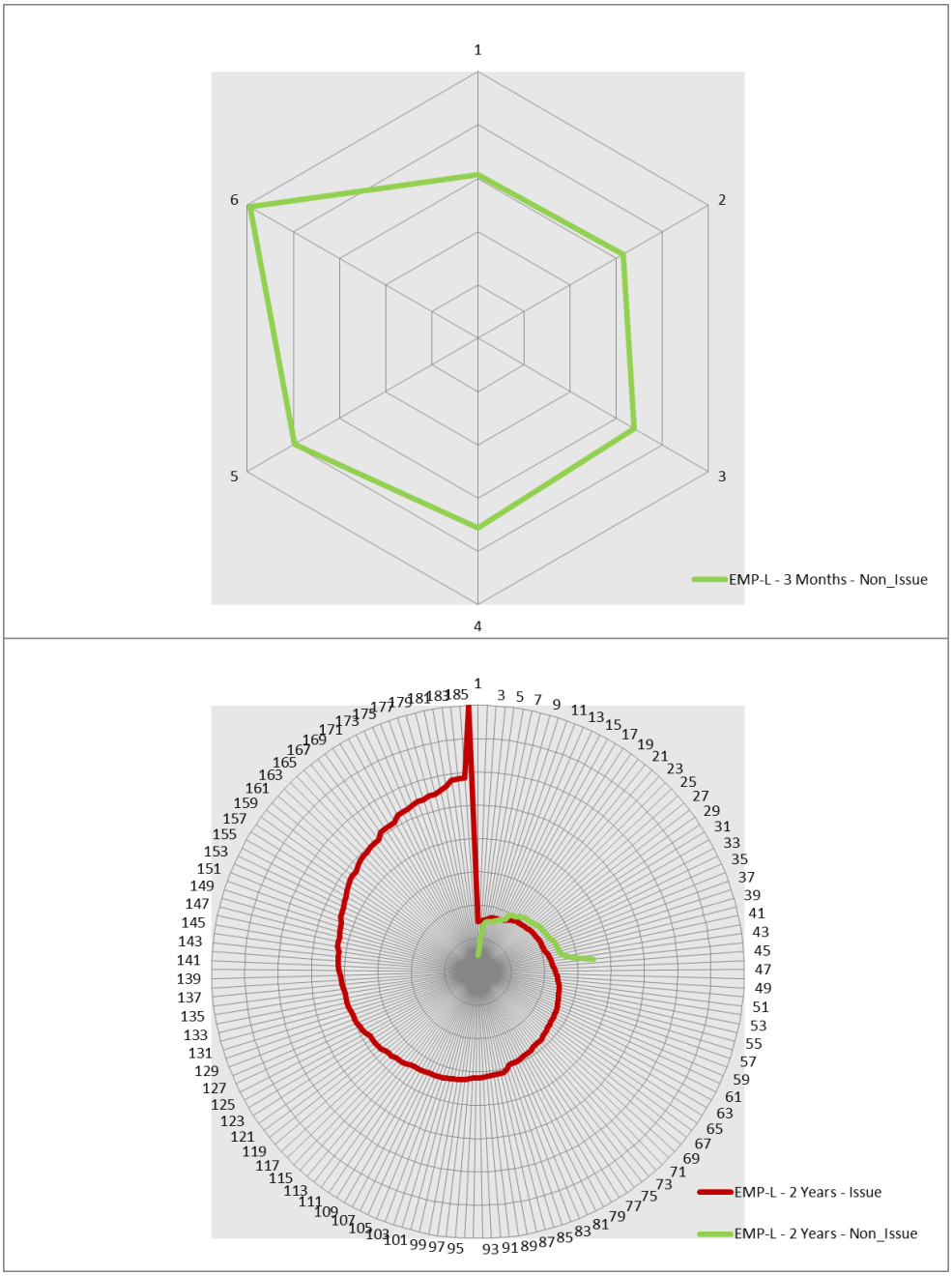

Figure 19 - Example of a threshold distribution (Source: Author)

As it can be seen in the figure above, in each distribution there are two lines, the line that represents the values of the threshold that contributed/'caused' an alerted that was resolved as one with issue, the red line, and another that represents the ones that were resolved as not having any issue, the green line. The optimal value point of the threshold is the value just below the two lines interception, being this value contextualized by the values on the three-month graph.

During this analysis phase there a few problems that were identified, being that first one, the fact that there were accounts and clients that were not attributed to any population, and the other, that existed client's profiles that had contradictory attributes, like contradictory occupation and professional situation (e.g. Lawyer and Student, respectively). As a result of these findings a data correction team was formed in order to resolve these situations.

6.3. DATA PARALLELISM

The analyses introduced in the previous chapter, proved to be reliable and provided sustained conclusions in order to be used as basis to the construction and definition of multiple aspects of the new AML&TF system. This chapter will detail which aspects are these, and how they will construct and sustained.

Following the order of the analyses on the previous chapter the first aspect to be introduced is the population thematic, which in the new system environment is be called segments. Taking into consideration and as basis, the distributions and conclusions presented, in the new system there will exist six segments. These segments will result in the aggregation by the significant population of the similar, behaviorally and context wise, populations. This resulted in the aggregation of the 'PART-H', 'PART-M', 'PART-L', 'PART-INT', 'PART-VW' and 'WU', creating the segment 'Particular'; the 'PB-H', 'PB-M', 'PB-L', 'PB', 'PB-VW' creating the segment 'PB & CGV'; the 'ENI-H', 'ENI-M', 'ENI-L', 'EN'I and 'ENI-VW', creating the segment 'ENI'; the 'EMP-H', 'EMP-M', 'EMP-L', 'EMP-MD' and 'EMP-SD', creating the segment 'SME'; the 'EMP-LD' in to the segment 'CORP'; and the 'DINT', 'DINT-VL', 'DINT-L', 'DINT-ML', 'DINT-M', 'DINT-MH', 'DINT-H', 'DINT-RG', 'DINT-INV' and 'DINT-REP', creating the segment 'DINT'. For the Macau's tenant the segments will remain two, equal to the populations in the old system, 'MACAO – EMP' as the 'CORP' and 'MACAO – PART' as the 'Particular' segment.

Regarding the rules, a parallelism between the similar rules of the old system and the rules of the new system, known as scenarios, was made, matching the rules that analyzed similar behaviors and that were also efficient in the detection of such, to the scenarios. This was made to all the scenarios that had some sort of similarity, being that it was not possible to sustain all the scenarios, since there are some that evaluated behaviors that were not considered in the old system. Taking into consideration these, when possible an assignation of values to the parameters of the scenarios was made, in accordance with the optimal values of the parameters of the old system's rules. These values were obtained in different forms, depending on the similarity of the rules and scenarios, and the on the parameters in themselves.

6.4. DATA MAPPING

The system will be feed mainly three set of data, the clients and account profile's information, the transactions and the transfers. Each of its composing data required to be assigned to its proper place

on the system, transformed into specific formats and matched to internally defined concepts. This process was very exhaustive and detailed, since the volume of data required to be mapped was very large and with a high dimensionality.

The client and account's profile had a very high volume and dimensionality, requiring also very specific formats that differ from the original data format, which implied the building of a set of ETL processes in order to sustain the continuous feed of the correctly formatted data.

The transfers also required the creation of an ETL process in order to be understood and readable to the system, being created automatisms that provided the necessary information for the system to use the necessary transfer data in its models.

The transactions required to be mapped into a set of predefined transaction types in the system. This meant that a contextualization and understanding of each internal transaction was to be understood in order to assignee it to the correct transaction type. This was a very important step, because it is this mapping that will dictate how the transactional behavior of a client will be viewed, if a transaction is wrongly set, it would mean that a behavior would be misinterpreted by the system. It these types have an important specification that sets them to be considered in the alert generation or to just be used as contextualization in the alert treatment, imposing an even more detailed layer of complexity since the utility of the transactions would also need to be validated.

This is a continuous work since the ecosystem of the three sets of data used is in constant change, growth and evolution.

6.5. POPULATIONS/SEGMENTS

The customer base from Portugal will be divided into 6 main segments, Particular, Private Banking & Clients Golden Visa (PB & CGV), Self Employed Entrepreneurs (SEE or ENI), Small and Medium Companies (SMC or SME), Corporations (CORP) and International Board, grouping clients according with similar behaviors. Within each one these, a new segmentation is made, this time according with the client's risk, in the context of its segment (Table 10). The customer base from Macau will be divided into 2 segments, Individuals and PB & CGV, also being further segmented by three levels of risk (Table 11). A special isolated population exists for the customers that are employed by the financial institution (Table 12).

	Particular	PB & CGV	ENI	SME	CORP	DINT
Low Risk	Segment 1	Segment 4	Segment 7	Segment 10	Segment 13	Segment 16
Medium Risk	Segment 2	Segment 5	Segment 8	Segment 11	Segment 14	Segment 17
High Risk	Segment 3	Segment 6	Segment 9	Segment 12	Segment 15	Segment 18

Table 10 - Portugal's clients' segmentation

	Individuals	PB & CGV
Low Risk	Segment 1	Segment 4
Medium Risk	Segment 2	Segment 5
High Risk	Segment 3	Segment 6

Table 11 - Macau's clients' segmentation

	Individuals
Low Risk	Segment 1
Medium Risk	Segment 2
High Risk	Segment 3

Table 12 – Bank's employees' segmentation

This initial segmentation was made accordingly with empirical knowledge gained from working with the previous AML&CFT system. Being that this segmentation will be validated in the validation stage and changed according with the analyses and results obtained.

6.6. SCENARIOS

The AML/CFT scenarios analyze transactional behaviors of customers or accounts, with the objective of detecting suspicious patterns. On these analyses, various attributes are taken into account, in order to allow for complex evaluations and detection logics.

Having right out-of-the-box, the following scenarios:

Scenario 1 – Large Cash Deposits

The large cash deposits scenario (scenario 1) checks for behaviors where the customers make one or more cash deposits, on a single day. The aggregated amount of these deposits must exceed the defined threshold in order to trigger this scenario.

Scenario 2 – Large Total Cash Transactions

The large total cash transactions scenario (scenario 2) checks for behaviors where customers make one or more cash transactions, of both credit and debit, through the course of several business days. The aggregated amount, of all the valid transactions, must be equal or higher than its threshold in order to trigger this scenario. By valid transactions, is to be understood, all the transactions that have their amount fall below the defined threshold; and the transactions that cumulatively, have an aggregated amount superior to the defined threshold, daily wise.

Scenario 3 – Structured Cash Deposits

The structured cash deposits scenario (scenario 3) checks for behavior s where the customer's the total amount of cash deposits consistently falls just below the daily Currency Transaction Report (CTR) threshold. It does so through the calculi of a structuring factor for each customer, translating into a value, the speed at which these funds are being made, in a period of several days. The scenario is triggered when the value of this factor is exceeded.

$$Structuring\ Factor = \frac{x^2 y^{1.9}}{(2 * CTR\ Limit - x)^{1.6} z^{1.4}}$$

Where:

x: Minimum daily transaction amount

y: Number of business days at or above the amount

z: Number of business days in the interval examined
CTR Limit: typically established by the regulatory entities

Equation 30 - Structuring Factor formula

Scenario 4 – Activity in an Inactive Account

The activity in an inactive account scenario (scenario 4) checks for behaviors where a customer's dormant account (account without any transaction in a delimited temporal interval) resumes activity, in such a way that changes its status into active one. This activity, composed by either credit or/and debit transactions, must have, collectively, an amount superior to the defined threshold, in order to trigger this scenario.

Scenario 5 – Transactions in Similar Amount

The transactions in similar amount scenario (scenario 5) checks for behaviors where the customer's account has many valid transactions, both credit and debit, in similar amounts over an extensive/extended period. By valid transactions, is to be understood, transactions that have their amounts superior than the defined threshold; and transactions that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval. The scenario is triggered when, the percentile difference between the value of the lowest transaction's amount and the highest ones is lower than the specified threshold, and when the volume of similar valid transactions is higher than the defined threshold.

Scenario 6 – Structured Withdrawals

The structured withdrawals scenario (scenario 6) checks for behaviors where the customer's account has multiple valid withdrawals over a short time period. By valid transactions, is to be understood, transactions that have their amounts superior than the defined threshold; and transactions that cumulatively, have an aggregated amount superior to the defined threshold, in the short temporal interval. This scenario is triggered, when the volume of valid transactions is higher than the defined threshold, in the defined temporal interval.

Scenario 7 – Multiple Location Usage

The multiple location usage scenario (scenario 7) checks for behaviors where the customer's account has multiple valid transactions, both credit and/or debit, at multiple locations on a single day. By valid transactions, is to be understood, transactions that have their amounts superior than the defined threshold; and transactions that cumulatively, have an aggregated amount superior to the defined threshold, in daily manner. This scenario will be triggered when valid transactions take place in more branches than those defined by the threshold, within the same day.

Scenario 8 – Structured Deposits across Locations

The structured deposits across locations scenario (scenario 8) checks for behaviors where the customer's account has been receiving valid deposits in a structured way, across one or more locations over a short time period. By valid deposits, is to be understood, deposits that amount to up to a ceiling, defined by a threshold; transactions that have their amounts superior than the defined threshold; and transactions that cumulatively, have an aggregated amount superior to the defined

threshold, in the defined temporal interval. This scenario will be triggered when valid deposits take place in more branches than those defined by the threshold, within the same day.

Scenario 9 – High-Risk Countries

The high-risk countries scenario (scenario 9) checks for behaviors where the customer's engages in valid transactions, both from and to, countries that are considered as high-risk (countries included the high-risk country list[62]). By valid transactions, is to be understood, transactions that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval.

Scenario 10 – Bidirectional Wires

The bidirectional wires scenario (scenario 10) checks for behaviors where the customer's account has both valid wires-in and valid wires-out transactions, in similar proportions and in a short chain reaction timestamp. By valid wires-in and valid wires-out transactions, is to be understood wires-in transactions that cumulatively, have an aggregated amount superior to the defined threshold, and wires-out transactions that cumulatively, have an aggregated amount superior to the defined threshold, of the percentage of the valid wires-in transactions amounts, both in the defined temporal interval.

Scenario 11 – Large Income Wires

The large income wires scenario (scenario 11) checks for behaviors where the customer receives valid wires-in transactions with large value amounts. By valid wires-in transactions, it is to be understood, wires-in transactions that have their amount up to a ceiling, defined by a threshold; wires-in transactions that have their amounts superior than the defined threshold; wires-in transactions that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval; and current day wires-in transactions that have their amounts superior than the defined threshold.

Scenario 12 – High Velocity Funds-Wires In

The high velocity funds-wires in scenario (scenario 12) checks for behaviors where the customer's account receives valid wired deposits that are quickly followed by valid withdrawals. By valid wired deposits, is to be understood, wired deposits that cumulatively, have an aggregated amount superior to the defined threshold. By valid withdrawals, is to be understood, withdrawals that have their amounts superior to the defined threshold; withdrawals that cumulatively, have an aggregate amount superior to the defined amount, in the defined temporal interval; withdrawals that cumulatively, have an amount superior to the defined threshold, of the percentage of valid wired deposits' amount, in the defined temporal interval; and withdrawals that cumulatively, have their aggregated amount up to a ceiling, defined by a threshold of the percentage of valid wired deposits' amount, in the defined temporal interval.

Scenario 13 – High Velocity Funds-Wires Out

[62] Annexes

The high velocity funds-wires out scenario (scenario 13) checks for behaviors where the customer's account receives valid deposits that are quickly followed by valid wired withdrawals. By valid deposits, is to be understood, deposits that cumulatively, have an aggregated amount superior to the defined threshold. By valid wired withdrawals, is to be understood, wired withdrawals that have their amounts superior to the defined threshold; wired withdrawals that cumulatively, have an aggregate amount superior to the defined amount, in the defined temporal interval; wired withdrawals that cumulatively, have an amount superior to the defined threshold, of the percentage of valid deposits' amount, in the defined temporal interval; and wired withdrawals that cumulatively, have their aggregated amount up to a ceiling, defined by a threshold of the percentage of valid deposits' amount, in the defined temporal interval.

Scenario 14 – Increased in Wire Activities

The increase in wired activities scenario (scenario 14) checks for behaviors where the customer's account as an increase, both in terms of amount and volume, of valid wired activity, both credits and debits, in relation of the current defined temporal interval and the previous ones. By valid wired activity, is to be understood, wired activity that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval; and wired activity that is composed by higher volume of wires that the defined threshold, in the defined temporal interval. This scenario will be triggered when there is a higher increase, in both terms of volume and amount, than those defined by their respective thresholds, in relation with the current defined temporal interval and the average of previous ones.

Scenario 15 – Early Termination of Multiple Loans

The early termination of multiple loans scenario (scenario 15) checks for behaviors where the customer closes multiple valid loans earlier than expected. By valid loans, is to be understood, loans that have their amounts superior than the defined threshold; and loans that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval. By early closed loans, is to be understood, loans that are closed with a temporal interval elapsed, higher than the defined by the threshold of percentage of time remaining until its termination. This scenario will be triggered when there is a higher volume of valid early closed loans, than the specified by the defined threshold.

Scenario 16 – Deposit Count in Excess of Expectations

The deposits count in excess of expectations scenario (scenario 16) checks for the behaviors where the customers' account receives a valid number of deposits that are unusually large, when compared to its normal activity. By valid number of deposits, is to be understood, volume of valid deposits that is higher than the defined threshold; and volume of valid deposits that cumulatively, have an aggregated volume higher than the defined by the threshold. This scenario will be triggered when there is a higher increase, on the valid volume of deposits, than the specified by the defined threshold, in relation with the current temporal interval and the previous ones.

Scenario 17 – Withdrawal Count in Excess of Expectation

The withdrawal count in excess of expectations scenario (scenario 17) checks for the behaviors where the customers' account receives a valid number of withdrawals that are unusually large, when

compared to its normal activity. By valid number of withdrawals, is to be understood, volume of valid withdrawals that is higher than the defined threshold; and volume of valid withdrawals that cumulatively, has an aggregated volume higher than the defined by the threshold. This scenario will be triggered when there is a higher increase, on the valid volume of withdrawals, than the specified by the defined threshold, in relation with the current temporal interval and the previous ones.

Scenario 18 – Account Activity in Excess of Expectations

The account activity in excess of expected scenario (scenario 18) checks for behaviors where the customer's account has an increase, both in terms of amount and volume, of valid activity, both credits and debits, in relation of the current defined temporal interval and the previous ones. By valid activity, it is to be understood, credit activities that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval, and debit activities that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval. This scenario will be triggered when there is a higher increase, on the valid volume of credit and debit activities, than the defined by the thresholds, in relation with the current temporal interval and the previous ones, and when there is a higher increase, on the aggregated amount of the valid volume activities, than the defined by the thresholds, in relation with the current temporal interval and the previous ones.

Scenario 19 – High-Velocity Funds in Excess of Expectations

The high-velocity funds in excess of expectations scenario (scenario 19) checks for behaviors where the customer's accounts receive an unusual amount of credits, and quickly after, receives valid debits in amounts like the credit ones. By unusual amount of credits, is to be understood, credits that cumulatively, have an aggregated amount superior to the defined by the threshold, in relation with the current temporal interval and the previous ones; and credits within which, there is at least one that has an amount superior than the defined by the threshold, in relation with the current temporal interval and the previous ones. By valid debits, is to be understood, debits that have their amounts superior to the defined threshold. By valid debits in amounts similar to the credits ones, is to be understood, valid debits that cumulatively, have an aggregated amount superior to the defined by the threshold, in relation with the current temporal interval and the previous ones; and valid debits that cumulatively, have an aggregated amount up to a ceiling, defined by the threshold, in relation with current temporal interval and the previous ones.

Scenario 20 – Correspondent Bank, N Internal Remitters, 1 External Beneficiary

The correspondent bank, n internal remitters, 1 external beneficiary scenario (scenario 20) checks for behaviors where multiple internal remitters wires a significant amount of money, to an external beneficiary over a short period of time. By significant amount, is to be understood, transactions that have their amounts superior to the defined threshold; and transactions that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval. By multiple internal remitters, is to be understood, volume of internal remitters is higher than the defined by the threshold.

Scenario 21 – Correspondent Bank, N External Remitters, 1 Internal Beneficiary

The correspondent bank, n external remitters, 1 internal beneficiary scenario (scenario 21) checks for behaviors where multiple external remitters wires a significant amount of money, to an internal beneficiary over a short period of time. By significant amount, is to be understood, transactions that have their amounts superior to the defined threshold; and transactions that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval. By multiple external remitters, is to be understood, volume of external remitters is higher than the defined by the threshold.

Scenario 22 – Correspondent Bank, 1 External Remitter, N Internal Beneficiaries

The correspondent bank, 1 external remitter, n internal beneficiaries' scenario (scenario 21) checks for behaviors where one external remitter wires a significant amount of money, to multiple internal beneficiaries over a short period of time. By significant amount, is to be understood, transactions that have their amounts superior to the defined threshold; and transactions that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval. By multiple internal beneficiaries, is to be understood, volume of internal beneficiaries is higher than the defined by the threshold.

Scenario 23 – Correspondent Bank, 1 Internal Remitter, N External Beneficiaries

The correspondent bank, 1 internal remitter, n external beneficiaries' scenario (scenario 20) checks for behaviors where one internal remitter wires a significant amount of money, to multiple external beneficiaries over a short period of time. By significant amount, is to be understood, transactions that have their amounts superior to the defined threshold; and transactions that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval. By multiple external beneficiaries, is to be understood, volume of external beneficiaries is higher than the defined by the threshold.

Scenario 24 – Cash Advance with Credit Card

The cash advance with credit card scenario (scenario 24) checks for behaviors where there are frequent valid cash withdrawals through ATM from credit card. By valid withdrawals, it is to be understood, withdrawals that cumulatively, have an aggregated amount superior to the defined threshold. By frequent valid withdrawals, is to be understood, volume of valid withdrawals higher than the defined threshold.

Scenario 25 – Frequent cash Advance with Credit Card

The frequent cash advanced with credit card scenario (scenario 25) checks for behaviors where there are frequent valid cash withdrawals through ATM from credit card. By valid cash withdrawals, it is to be understood, withdrawals that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval. By frequent valid withdrawals, is to be understood, volume of valid cash withdrawals that is higher than the defined threshold, in the defined temporal interval.

Scenario 26 – Credit Card Payment

The credit card payment scenario (scenario 26) checks for behaviors where there are frequent valid cash payments on credit card. By valid cash payments, is to be understood, payments that

cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval. By frequent valid cash payments, is to be understood, volume of valid cash payments that is higher than the defined threshold, in the defined temporal interval.

The objective behind the mutual existence of these multiple scenarios is the existence of a protective net, which filters all the behaviors, only catching the suspicious ones. In order to have a generalized view of the universe that is being filtered, the following tables and analyses were done:

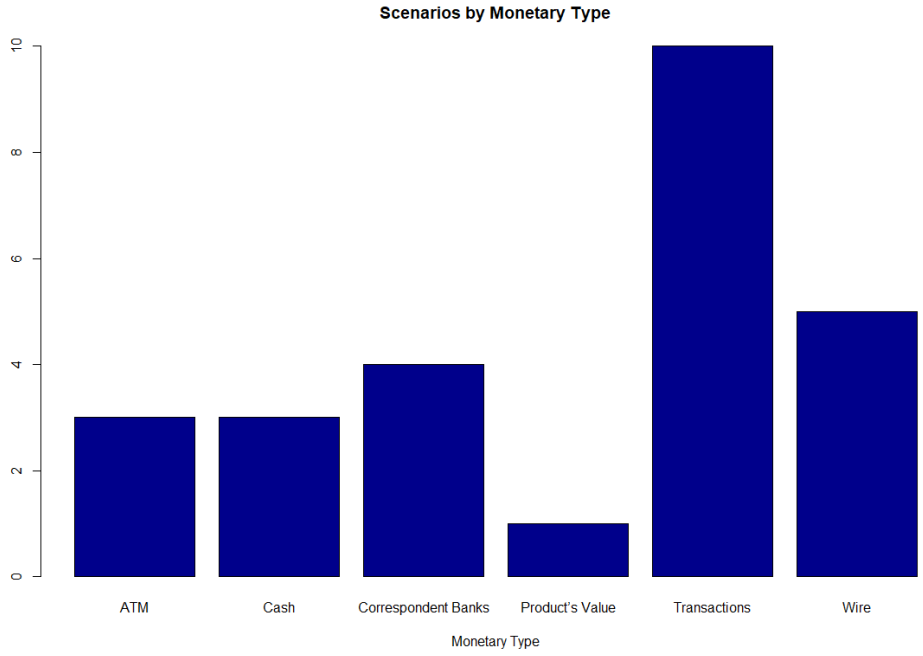

Figure 20 - Scenario Distribution by Monetary Type (Source: Author)

The previous figure, the figure 20, illustrates the distribution of the number of scenarios across the 'Monetary Type' categories, which represents the nature of the value evaluated in each scenario. It is possible to visualize that the number of scenarios is well distributed across all the 'Monetary Type' categories, except for the 'Product's Value' and 'Transactions'. 'Product's Value' category only has one scenario to verify its associated behaviors, but this volume should be satisfactory since it is a specialized case for one product value. The 'Transactions' category has a very high number of scenarios to validate its behaviors, but this is due to the innumerous suspicious behaviors that this type can have.

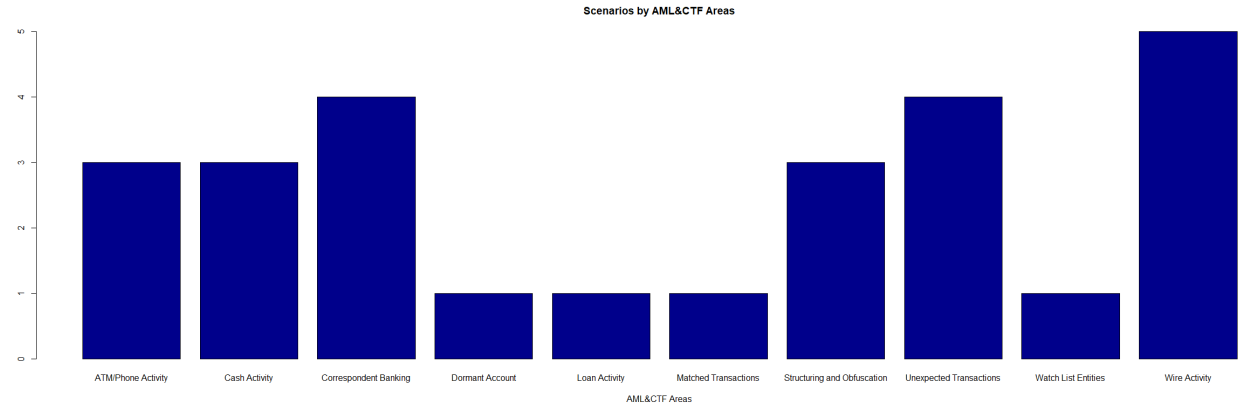

Figure 21 - Scenario Distribution by AML/CTF Areas (Source: Author)

The figure 21 illustrates the distribution of the volume of scenarios across the 'AML/CTF Areas', which represents the areas of interest of the AML/CTF. It possible to visualize that the volume of scenarios is well distributed across the 'AML/CTF Areas', having some, a significantly low number of scenarios but these areas in themselves do not justify a higher volume, since they represent very specific theme (Matched Transactions, Watch List Entities and Loan Activity).

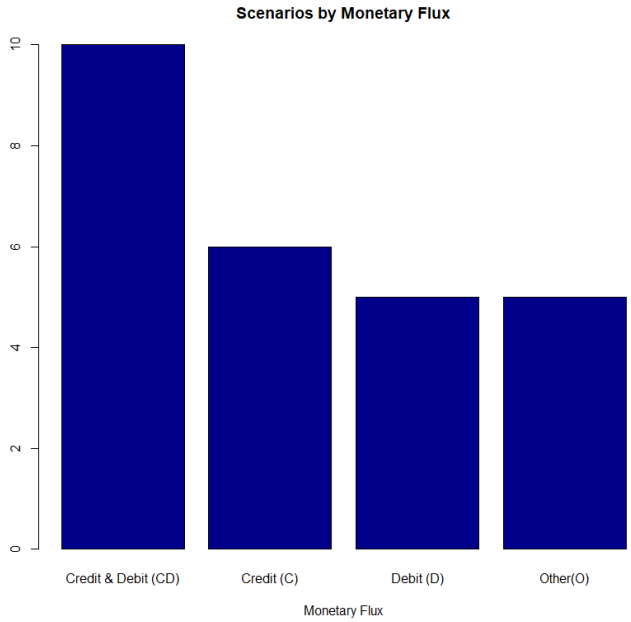

Figure 22 - Scenario Distribution by Monetary Flux (Source: Author)

In the figure 22 it is possible to visualize the distribution of the number of scenarios across the 'Monetary Flux', which represents the circulation direction of the value taken into consideration by scenario. It is possible to visualize that the scenarios are well distributed across the 'Monetary Flux' types, being that there are five scenarios that do not match any of the traditional types, since the value which is analyzed in those cases is intrinsic to the activity/product itself and cannot exist outside that context.

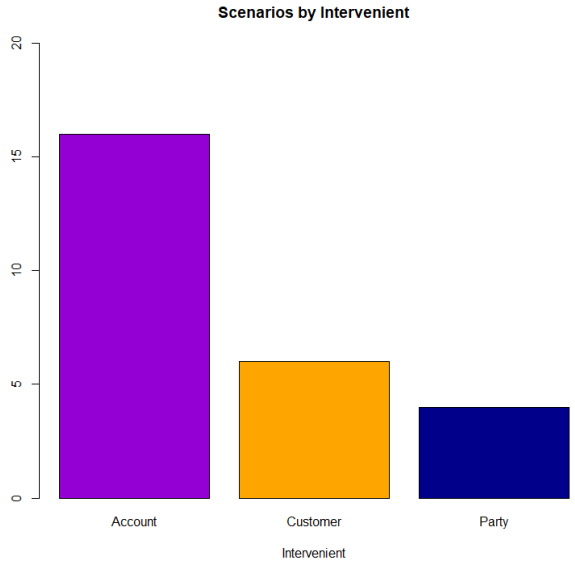

Figure 23 - Scenario Distribution by Intervenient (Source: Author)

The figure 23 illustrates the distribution of the volume of scenarios across the possible intervenient contained by it. The distribution shows a scenario tendency, to focus more on the account behaviors.

Scenario ID(Frequency)	Cash	Wire	ATM	Transactions	Product's Value	Correspondent Banks
Cash Activity	1(C); 2(D); 3(C)	-	-	-	-	-
Dormant account	-	-	-	4(CD)	-	-
Matched Transactions	-	-	-	5(CD)	-	-
Structuring and Obfuscation	-	-	-	6(D); 7(C); 8(C)	-	-
Watch List Entities	-	-	-	9(CD)	-	-
Wire Activity	-	10(CD); 11(C); 12(CD); 13(CD); 14(CD)	-	-	-	-
Loan Activity	-	-	-	-	15(O)	-
Unexpected Transactions	-	-	-	16(C); 17(D); 18(CD); 19(CD)	-	-
Correspondent Banking	-	-	-	-	-	20(O); 21(O); 22(O); 23(O)

ATM/Phone Activity	-	-	24(D); 25(D); 26(D)	-	-	-

Table 13 - Scenario logical classification

The table 13 illustrates the culmination/aggregation of the figures n, n, n and n. This translates into a distribution of the scenarios' respective 'Monetary Flux' type and Intervenient, across the 'AML/CFT Areas' and 'Monetary Type' categories. It is possible to see that the scenarios belonging to the 'Monetary Type' category 'Transactions' are the only ones that do not focus in only one of the 'AML/CFT Areas'. This may derive from the fact that this 'Monetary Type' category is very broad, being used across a multitude of areas. In regards with the 'Monetary Flux', each group of scenarios represents the totality of its universe. Intervenient wise, the scenario groups tend to be homogeneous, representing thematic analyses focused on a certain intervenient.

Accordingly, with the results obtained in the validation and report stages, and along the operation lifetime of the system, new scenarios can and will be added to check risk behaviors have yet to be validated.

6.7. RISK FACTORS

In a similar fashion to the AML/CFT's scenarios, the risk factors analyze transactional behavior of customers or accounts, with the objective of detecting risk patterns. A wide range of attributes thematic are considered in those analyses, allowing for the implementation of complex evaluations and detection logics.

The following risk factors are the ones that are going to be implemented and used right out-of-the-box:

Risk Factor 1 – High Velocity Funds

The high velocity funds risk factor (risk factor 1) checks for patterns where valid funds flow into and out of an account over a short time period. By valid funds, is to be understood, funds that cumulatively, have an aggregated amount superior to the defined threshold, in the defined temporal interval. By flow into and out over a short period of time, is to be understood, that the calculated velocity factor is higher than the defined threshold, in the defined temporal interval.

$$Velocity\ Factor = \frac{Max - Avg}{Man - Min}$$

Equation 31 - Velocity Factor formula

Risk Factor 2 – Transactions in Round Amounts

The transactions in round amounts risk factor (risk factor 2) checks for patterns where a customer's transactions' amounts are dominated by round values. By round values is to be understood, values that have several trailing zeros higher than the defined threshold. By round values dominated, is to be understood, the volume of round valued amounts is be higher than the defined threshold, in accordance with the totality of values transacted.

Risk Factor 3 – ATM Deposits at Multiple Locations

The ATM deposits at multiple locations risk factor (risk factor 3) checks for patterns where an account has ATM deposits at multiple unique locations on any single day, during a specific time period. By multiple unique locations, is to be understood, volume of ATM's locations used is higher than the defined threshold, in a single day. This risk factor will be activated when the number of single days that shown activity is higher than the defined threshold.

Risk Factor 4 – Multiple Branch Usage

The multiple branch usage risk factor (risk factor 4) checks for patterns where an account repeatedly conducts transactions at multiple branches over a short period of time. By multiple branches, is to be understood, volume of branches is higher than the defined threshold, in a consecutive manor, meaning, within a defined temporal interval and with a time frequency between them, lower than the defined threshold. This risk factor will be activated when multiple branches are used within the defined temporal interval.

Risk Factor 5 – Activity Dominated by Wires

The activity dominated by wires risk factor (risk factor 5) checks for patterns where a large percentage of a customer's total transaction amount is wire transfers. By large percentage of wired transfers, is to be understood, volume of wired transfers is higher than the defined threshold, in accordance with the totality of transactions made, in the defined temporal interval.

Risk Factor 6 – Foreign Wire Activity

The foreign wire activity risk factor (risk factor 6) checks for patterns where customers exceed a given number of foreign wire transactions. By exceeds a given number of foreign wire transactions, is to be understood, volume of wired transactions is higher than the defined threshold, in the defined temporal interval.

Risk Factor 7 – New Customer

The new customer risk factor (risk factor 7) checks for patterns where a customer has been with a financial institution for a short period of time. This risk factor will be activated, when the age of the relationship with the financial institution bellow the ceiling defined by the threshold.

Risk Factor 8 – Politically Exposed Personal (PEP) Indicator

The politically exposed personal (PEP) indicator risk factor (risk factor 8) checks for a customer's risk rank increase due to already being a politically exposed person (PEP). Meaning, this risk factor will be activated when the customer has the PEP flag tagged. It is to be noted that this risk factor does not determine if a customer is PEP or not.

Risk Factor 9 – Recent Suspicious Activity Report (SAR)

The recent suspicious activity report (SAR) risk factor (risk factor 9) checks for a customer record that is marked, as recently having a suspicious activity report (SAR).

This risk factor will not be used, since this AML/CFT's suspicious activity reporting tools will not be used for the bank's context (the reporting action is already accounted for, in internal client's risk).

Risk Factor 10 – High Account Turnover

The high account turnover risk factor (risk factor 10) checks for a customer's total transaction amount that is unusually large, when compared with its balance. By unusually large total transaction amount, is to be understood, minimum average turnover of the accounts is higher than the defined threshold, in accordance with the totality of turnover in the account, in the defined temporal interval.

In similarity to the scenarios' objective, the mutual existence of these multiple risk factors is the existence of a risk net, which filters risk patterns in order to correctly assess the true risk value inherent to the customer. In order to have a generalized view of the universe that is being filtered, the following tables and analyses were done:

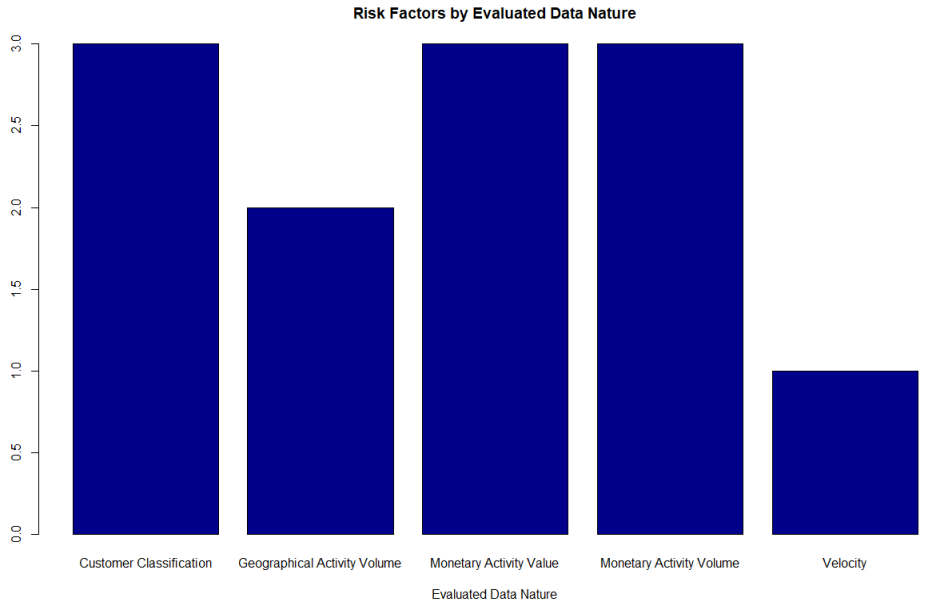

Figure 24 - Risk Factor Distribution by Evaluated Data Nature (Source: Author)

The figure 24 illustrates the risk factors' distribution in regards of the nature of the data that is analyzed by them. It is possible to visual that there is no focus on a specific data nature, being the scenarios well distributed across all the data. The exception to this is the 'Velocity' nature, which is justified since it is a very specific case. Thematic wise, it is possible to extrapolate 3 main groups, the volume ('Monetary Activity Value' and 'Velocity'), the value ('Monetary Activity Volume' and 'Geographical Activity Volume') and the classification ('Customer Classification') related one. With this aggregation, a greater focus on the volume's related group becomes evident.

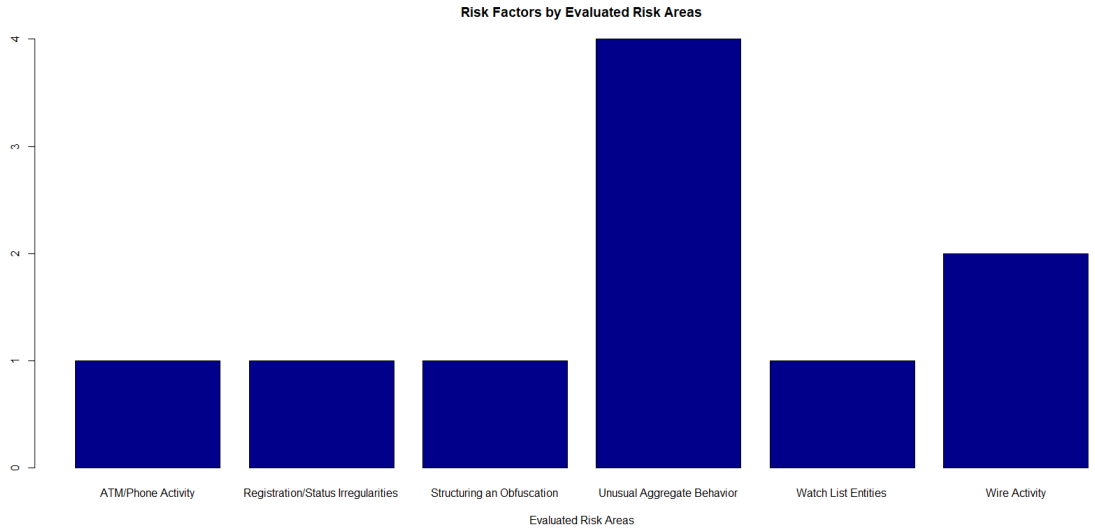

Figure 25 - Risk Factor Distribution by Evaluated Risk Areas (Source: Author)

The figure 25 illustrates the distribution of the volume of risk factors, in accordance with the evaluated risk area, showing a tendency of greater focus on the risk area 'Unusual Aggregative Behaviors'. This derives from the wide ranged nature of this risk factor, which justifies the more detailed attention granted to it.

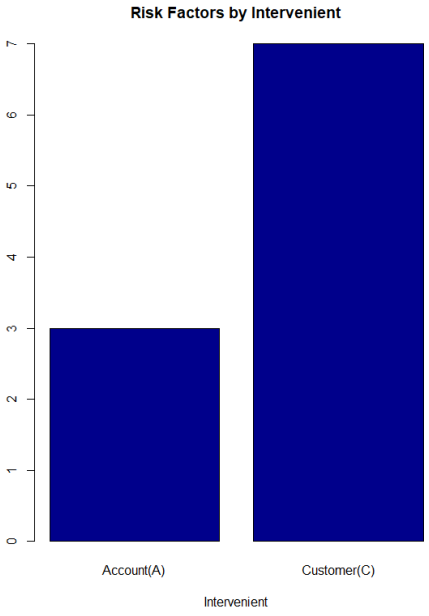

Figure 26 - Risk Factor Distribution by Intervenient (Source: Author)

The figure 26 illustrates the distribution of the number of risk factors in regards of its intervenient. This distribution shows a clear focus on the customer, which may serve as a balance achiever for the focus of the scenarios (Previous chapter) on the account analyses.

Risk Factor ID(Intervenient)	Monetary Activity Value	Monetary Activity	Velocity	Geographical Activity	Customer Classification

		Volume		Volume	
Unusual Aggregate Behavior	1(A); 2(C); 10(A)	2(C)	1(C)	-	9(C)
ATM/Phone Activity	-	-	-	3(C)	-
Structuring an Obfuscation	-	-	-	4(A)	-
Wire Activity	-	5(C); 6(C)	-	-	-
Registration/Status Irregularities	-	-	-	-	7(C)
Watch List Entities	-	-	-	-	8(C)

Table 14 - Risk Factor logical classification

The table 14 illustrates the culmination of the figures n, n and n. This translates into a distribution of the risk factors and their respective 'Intervenient', across the 'Evaluated data nature' and the 'Evaluated risk areas' categories. It is possible to visualize that the risk factors do not tend to form groups, spreading across both the 'Evaluated data nature' and the 'Evaluated risk areas' categories. The exceptions of this are the risk factors' 1, 2 and 10 group, and the risk factors' 5 and 6 one. This first group and the risk factor 4 are the only group of risk factors that, majority wise, focus on the pattern analyses of accounts.

Accordingly, with the results obtained in the validation and in the report stages, and along the operation lifetime of the system, new risk factors can and will be added to check risks that have yet to be validated, by the previous ones.

6.8. THRESHOLDS

The threshold definition is a sensible theme, because their definition is what will dictate way that the system will work, since, even if the risk factors and the scenarios are correctly designed and include all the suspicious behaviors' and risk patterns' methodologies, if their respective threshold setting is misadjusted to the population context, their efficiency and effectiveness will be nil.
The initial setting followed the logic referred in the Data Parallelism chapter. The future setting, that will be a continuous work, will follow the logic referred in the Threshold Optimization chapter.

6.9. CDDS

The CDD scoring model concept resolves around the individual's attributes evaluation, having its risk model as the one responsible for definition of the business logic to determine the risk level of each individual, either client or related person, based on set of attribute values stored on them.
The model is built up in a hierarchical schema:

1. Risk Level, is the top level, representing the final categorization of the risk;

2. Overall Score, is the level that supports the final categorization, storing the overall score value;
3. Categories, is the multi-categorical level that supports the overall score value calculation, storing the score of each individual category;
4. Attributes, is the multi-attribute level that supports each categorical score, storing the score and value for each attribute, related to its respective category;
5. Lookup tables are the record tables that support each attribute score, storing table of records related to its respective attribute.

The CDD rules that are going to be implemented/used from the get-go are:

CDD 1 – Registration

The Registration customer due diligence rule (CDD 1) is a Non-personal country rule. It looks up the risk score accordingly with the country where the organization is registered in.

CDD 2 – Primary Business Address

The Primary business address customer due diligence rule (CDD 2) is a Non-personal country rule. It looks up the risk score accordingly with the country of the organization's primary address.

CDD 3 – Beneficiary Owner Residence

The Beneficiary owner residence customer due diligence rule (CDD 3) is a Non-personal country rule. It looks up the risk score accordingly with the organization's beneficial owner's country of residence.

CDD 4 – Residence

The Residence customer due diligence rule (CDD 4) is a Personal country rule. It looks up the score accordingly with the country of the customer's residence.

CDD 5 – Citizenship

The Citizenship customer due diligence rule (CDD 5) is a Personal country rule. It looks up the risk score accordingly with the customer's country of citizenship.

CDD 6 – Industry

The Industry customer due diligence country rule (CDD 6) is a Non-personal entity rule. It looks up the risk score accordingly with the organization's industry.

CDD 7 – Legal Entity Status

The Legal entity states customer due diligence rule (CDD7) is a Non-personal entity rule. It looks up the risk score accordingly with the organization's legal entity status.

CDD 8 – Non-Profit Organization

The Non-profit organization customer due diligence rule (CDD 8) is a Non-profit organization attribute rule. It looks up the risk score accordingly with the organization type, if it is a non-profit organization or not.

CDD 9 – Money Service Business (MSB)

The Money service business (MSB) customer due diligence rule (CDD 9) is a Non-personal entity rule. It looks up a risk score accordingly with the organization type, if it is MSB related or not

CDD 10 – Engaged in Internet Gambling

The Engaged in internet gambling customer due diligence rule (CDD 10) is a Non-personal entity rule. It looks up a risk core accordingly with the organization type, if it is internet gambling related or not.

CDD 11 – Trust Account

The Trust account customer due diligence rule (CDD 11) is a Non-personal entity rule. It looks up a risk score accordingly with the organization/account type, if it is a trust account or not.

CDD 12 – Foreign Consulate Embassy

The Foreign consulate embassy customer due diligence rule (CDD 12) is a Non-personal entity rule. It looks up a risk score accordingly with the organization type, if it is foreign consulate embassy or not.

CDD 13 – Issue Bearer Shares (Out of Scope)

The Issue bearer shares customer due diligence rule (CDD 13) is a Non-personal entity rule. It looks up a risk score accordingly with the organization type, if it is bearer shares emit related or not.

CDD 14 – Politically Exposed Person

The Politically exposed person customer due diligence rule (CDD 14) is a Non-personal rule. It looks up a risk score accordingly with the organization type, if it is a politically exposed organization or not´.

CDD 15 – Negative Media News Search (Out of scope)

The Negative media news search customer due diligence rule (CDD 15) is a Non-personal entity rule. It looks up a risk score accordingly with the negative opinion that exists in the media news, about the organization.

CDD 16 – On Boarding Risk Score (Custom)

The On boarding risk score customer due diligence rule (CDD 16) is a Non-personal entity rule. It looks up a risk score accordingly with the risk assessment of the customer's on boarding.

CDD 17 – Occupation

The Occupation customer due diligence rule (CDD 17) is a Personal entity rule. It looks up a risk score accordingly with customer's occupation.

CDD 18 – Politically Exposed Person

The Politically exposed person customer due diligence rule (CDD 18) is a Personal entity rule. It looks up a risk score accordingly with the customer type, if it is a politically exposed customer or not.

CDD 19 – Negative Media News Search (Out of Scope)

The Negative media news search customer due diligence rule (CDD 19) is a Personal entity rule. It looks up a risk score accordingly with the negative opinion that exists in the media news, about the customer.

CDD 20 – SAR's Count

The SAR's count customer due diligence rule (CDD 20) represents the number of SARs filed rule. It looks up a risk score accordingly with existence of SAR's filed on the name of the customer.

CDD 21 – On Boarding Risk Score (Custom)

The On boarding risk score customer due diligence rule (CDD 21) is a Personal entity rule. It looks up a risk score accordingly with the risk assessment of the customer's on boarding.

CDD 22 – Product

The Product customer due diligence rule (CDD 22) is a Product attribute rule. It looks up a risk score accordingly with the customer's accounts associated product names.

In similarity to the scenarios' and risk factors' objectives, the mutual existence of these multiple customer due diligence rules is the existence of a classification net, which filters risk attributes in order to correctly assess the true risk value inherent to the attribute and sequential impact on the respective customers. In order to have a generalized view of the universe that is being filtered, the following tables and analyses were done:

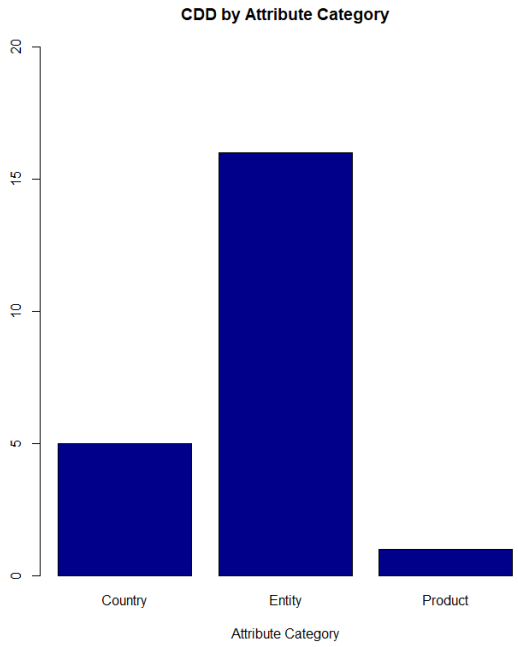

Figure 27 - CDD Rule Distribution by Attribute Category (Source: Author)

The figure 27 illustrates the distribution of the customer due diligence rules in regards of the 'Attribute Category'. This distribution is very uneven, existing a very disproportional degree of focus between the three existing categories. The main focused one is the 'Entity' category, having roughly three quarters of the rule universe population converge on it. The least focused, is the Product one, having only one rule converging on it.

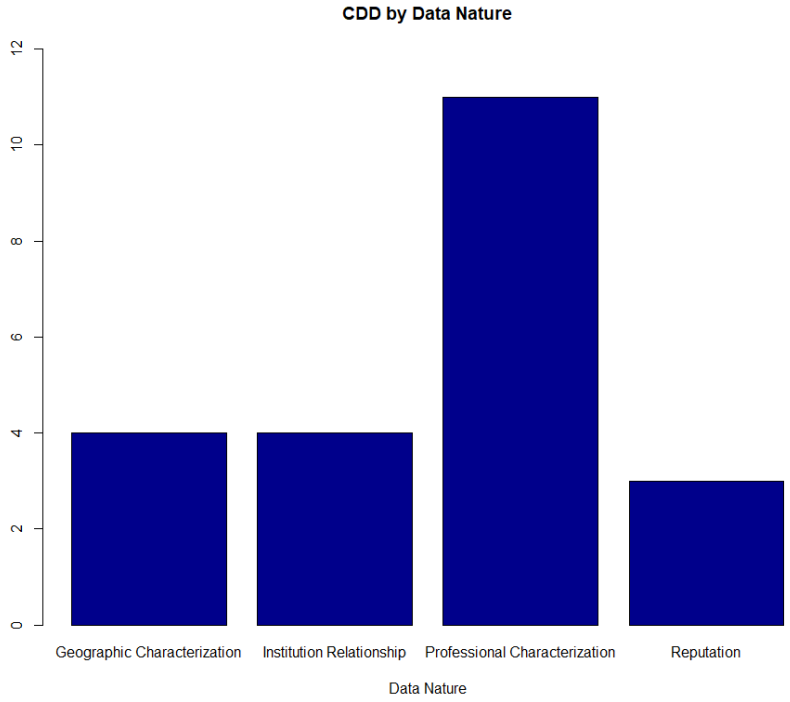

Figure 28 - CDD Rule Distribution by Data Nature (Source: Author)

The figure 28 illustrates the volume of the distribution of the customer due diligence rules, in accordance with the nature of the data that it's feed. This distribution is even between three of the four 'Data Nature' areas, being the fourth, the 'Professional Characterization' one, more focused on, having roughly half of the rule universe converging on it.

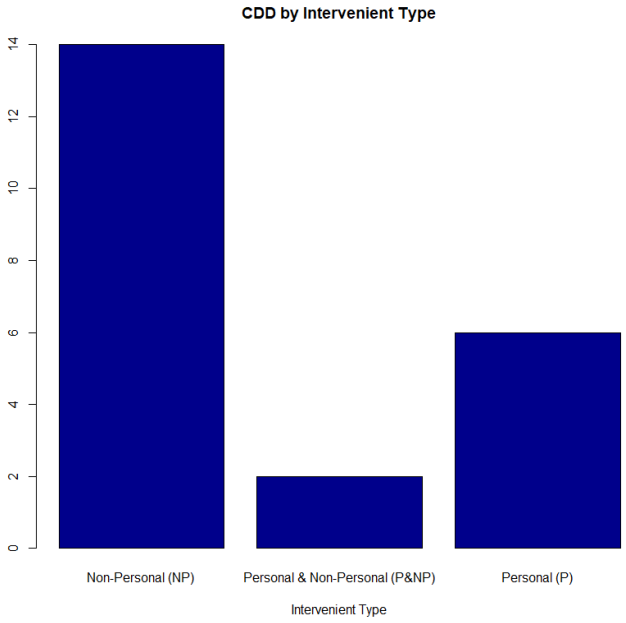

Figure 29 - CDD Rule Distribution by Intervenient (Source: Author)

The figure 29 illustrates the distribution of the customer due diligence rules in regards of its Intervenient. This distribution shows a clear focus on the Non-personal intervenient, being that this may be due to the higher standardization of the Non-personal attributes, in comparison with the Personal ones. By standardization, is to be understood, common existence of clear, reliable and updated data.

CDD ID(Intervenient Type)	Country	Entity	Product
Institution Relationship	1(NP)	16(NP); 21(P)	22(P&NP)
Geographic Characterization	2(NP); 3(NP); 4(P); 5(P)	-	-
Professional Characterization	-	6(NP); 7(NP); 8(NP); 9(NP); 10(NP); 11(NP); 12(NP); 13(NP); 14(NP); 17(P); 18(P)	-
Reputation	-	15(NP); 19(P); 20(P&NP)	-

Table 15 - CDD logical classification

The table 15 is the illustrated the culmination of the data from the figures 27, 28 and 29. This translates into a distribution of the customer due diligence rules and their respective intervenient type, across the 'Data nature' and 'Attribute Categories'. It possible to visualize that there is a focus, by the customer due diligence rules, to converge on the entity data professional characterization category. Regarding the intervenient, it is possible to assess that the groups created are balanced, if it is taken into consideration the proportional volume of each respective intervenient.

Accordingly, with the results obtained in the validation and in the report stages, and along the context changes during the operation lifetime of the system, new CDD can and will be added.

6.10. PRACTICAL USAGE OF THE SYSTEM

In this chapter it is possible to visualize and gain a better understanding of the way that system will function, user wise; the roles that the users can assume; and the impact and capabilities of each role. Depending on the internal control chosen, the usage process will vary, and as such the 'four eyes policy' must be considered the default control in action, for the analysis in this chapter (unless stated otherwise).

Taking this into consideration we can start by analyzing Figure 30.

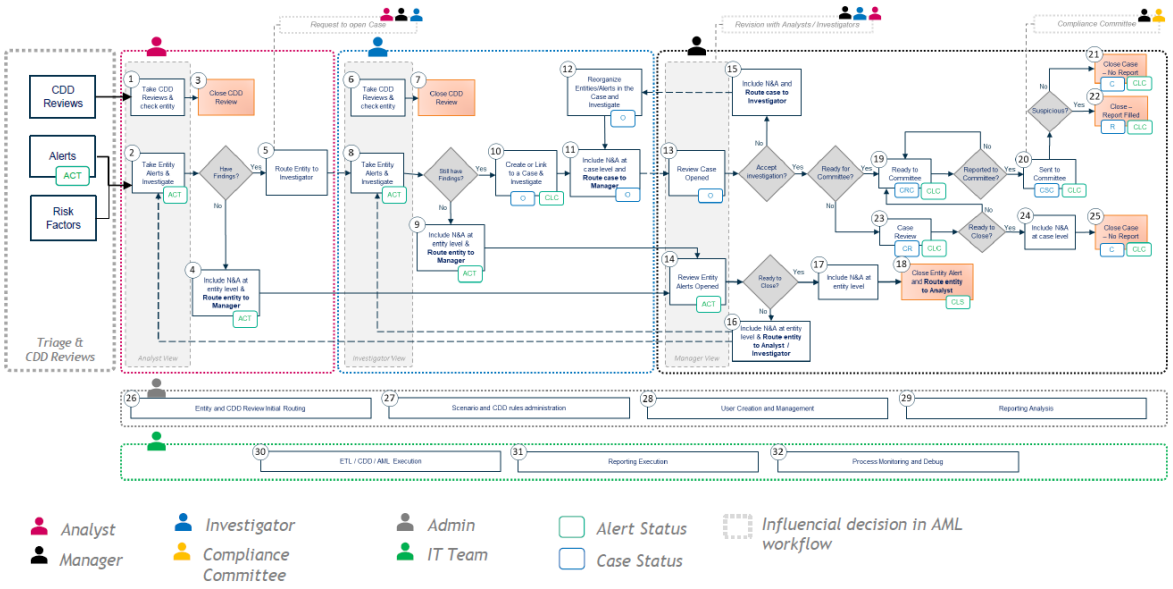

Figure 30 - System's Operation Flowchart (Source: SAS documentation)

In this flowchart it is possible to observe the various stages of interaction that the multiple categories of the compliance personal can have with and/or as a consequence of the system. The figure should be read with a left to right flow, being that in this context there are four main layers.

Before diving into the layers of interaction with the system, a crucial notion is required, the notion of the Triage and CDD review stage. This stage can be viewed as that central point of interaction between the users and the system, being here that the analysis of the alerts/reviews begins, having the various alerts and made available to interact with, in accordance with the profile of each user. The interactions that are made will then go through the following layers:

The first layer of users can be seen as the first stage where the work load will start to be treated, analyzed and processed by the intervenient to whom the work was assigned, being composed solely by the Analyst role. This intervenient role will be responsible for the CDD rule reviews and entity alerts' investigation. With regards of the capabilities of the role in terms of how much it can do in each of its processes, this role has the full autonomy on the CDD review process, being able to close the reviews. In regards of the entity alerts' investigation process, the role only has two options, they

either find evidences that corroborate the veracity or either the falsity of alert, in their perspective. If their opinion is of the veracity of the alerts, they will route the entity to the second layer of users, if on the contrary, their opinion is of the falsity of the alerts they include a note stating so and route the entity to the third layer of users.

The second layer of users is a second layer of validation and analysis, being responsible for the CDD rule review process and for the suspicious entity investigation process. The intervenient role responsible for these duties is deemed as the investigator. In relation with the CDD rule review process, the investigator capabilities will be the same as the analyst ones. When it comes to the suspicious entities' investigation process, the investigator will start by only investigate the entities that were already deemed as suspicious by the analysts, having then to find evidences that corroborate the veracity or the falsity of alert, in their perspective. If their opinion is of the veracity of the alerts, they will have to create a new case and link the related entity to it, or, depending this on the nature of the behavior in question, link to an already defined case. After this linkage, the investigator will have to leave a note of review on the case to which the entity/entities were linked, and route it to the third layer of users. In case of their opinion being of the falsity of the alerts they include a note stating so and route the entity, also to the third layer of users. It is to be noted that the organization/reorganization of the entities/alerts in a case is also a competence of the investigator, accordingly with the validation and decision of the third layer users.

The third layer of users is the last layer of triage on the suspicious entities, being constituted by the manager intervenient role. The manager is responsible for giving continuation to the investigation processes done by the previously introduced user layers. In regards of the entities that have been deemed suspicious in an invalid way by the investigators and/or analysts, the manager will decide, based on their investigations and analysis, if the review process is ready to be closed. If it is not, the manager will route the entity back to the first or second layer of users, for them to proceed with a more thorough approach. If it is deemed as ready to close, the manager will proceed to indicate the closure without a problem of the review process and close the entity alerts, routing the entity back to the first- or second-layer users. In relation with the entities that have been deemed as suspicious in a valid way and that now constitute a case, the manager will decide if the case should be accepted or not. If rejected, the manager will route it back to the second layer of users to further investigate the entities and/or add of remove investigations considering the context of the case at hand. If accepted, the manager will have to further decide if it the case is ready for committee (by ready for committee it is to be understood, if its contents have/show evidences that support suspicious behaviors). If it is not ready for committee the manager will close the case and leave a note stating that that the entities/behaviors contained in it were not reported. If on the other hand, the case is ready for committee, this one will be flagged as such, and sent to the forth layer of users.

The forth layer of users (indirect users) can viewed as the final stage of the suspicious behaviors, being on this layer where the process final decision will be made. This layer's intervenient role is the compliance committee which is constituted all the compliance directors and by the compliance head. It is to be noted that in order to achieve a decision on the committee, a consensus regarding a case must be reached, being that the compliance head has the veto power to shut down any consensus achieved. This final decision can only take two forms, either the case is to be reported or not. In case

of being reported, signifying that the consensus achieved in committee was of validating the suspicions, the information regarding the case will be sent to a jurist so that a report can be formulated and sent to the responsible authorities, and the case is closed with a note indicating its report. If on the contrary, it was decided to not report the case, the case will be closed with the information that was not reported.

The flowchart also has a layer of transversal users, being these transversals in the sense that they do not interact with the system in an orderly way, like previous layers, they interact with the application with a structural and management optic. In this layer there are two main groups of users, the Admin and the IT Team.

The Admin is responsible for: The Entity and CDD Review initial routing, which is only done by this personal only at the initial stage of the objects, since once assigned to a user, the objects will flow through the previously explain layers. This can be viewed as an introduction to the system's user treatment process; the administration of the scenario and CDD rules, which will allows for the parametrization of the thresholds, the creation of new rules, the deactivation of rules, and many others rule settings; the user creation and management, that can go as far as redefine the capabilities of a user profile in its interaction with the system; and the report analysis, that gives the user full access to the reports produced by the application, providing the to the user a more sustained way of taking decisions and have a view of the application functioning and user interaction efficiency centrally .

The IT team is responsible for the more technical support to the function of the system, being responsible for the ETL, CDD and AML execution, with all of the changes and adjustments necessary for the correct, both technically and logically, function of the system; the reporting administration, which evolves that maintenance of the already defined reports, the creation of new ones and all of the necessary adjustments to deliver the most up to date and crucial information; and the process monitoring and debug, so that all of the system's functionalities are working as they should.
It must be noted that one individual can have multiple roles within the system.

6.11. REPORTING (VA)

In this chapter, the reporting capabilities of the chosen tool will be showcased, explaining what the reporting tool can do and what in fact will be done in order to answer the needs of the organization.
The default reporting tool of the AML/CTF system is the *SAS Visual Analytics*, also known as VA. The VA tool is a business intelligence tool that is connected to the system's Knowledge and Core databases, having full access to the information that is on the system, such as its variables, parameters, rules and alerts. This tool will be used in a functional optic, meaning, it will be used to ensure and validate the correct function of the system, the performance of the individuals using the system, analyze the nature and context of the system's outputs, and much more (detailed later in the chapter).

VA provides a self-service data discovery process with its interactive experience with the reports, showing what the user wants, through the usage of filters and calculations, and with the detail that

he so desires, through the usage of hierarchies that allow for a drill-down and/or up, on the data demonstrated o report initially. Such interactions can be done with and through a vast range of analytical visualizations, from box plots, to decision trees, having this tool the capability of identifying the best visualization to explain the data that is wanted. In regards with the variables that it uses, the tool can use the variables that it is feed, to do an ETL process, being able to calculate new variables in order to complement the already existing ones so that an entirely complete view of the wanted context is possible to achieve. In addition to this the VA tool allows for the importation of external data/processing sources, such as *Youtube* videos, data from social networks, such as *Facebook* and *Twitter*, in order to do some text and sentiment analysis, and *ERSI ArcGIS* & *OpenStreetMap,* so that geographical visualizations and analysis are enabled.

When it comes to accessibility, the tool allows for a categorization of the users in accordance with reports to which each they should have access, meaning that it is possible to only propagate the reports to the people to who should have access. This access is also possible to be done in a multitude of forms, via web (Figure 31), app, pdf exportations (loses the interaction capabilities in this form), and others.

It must be noted that there are users with the capability of accessing the reports and interact with them, but in addition to these, there are also users with the ability to create and modify reports.

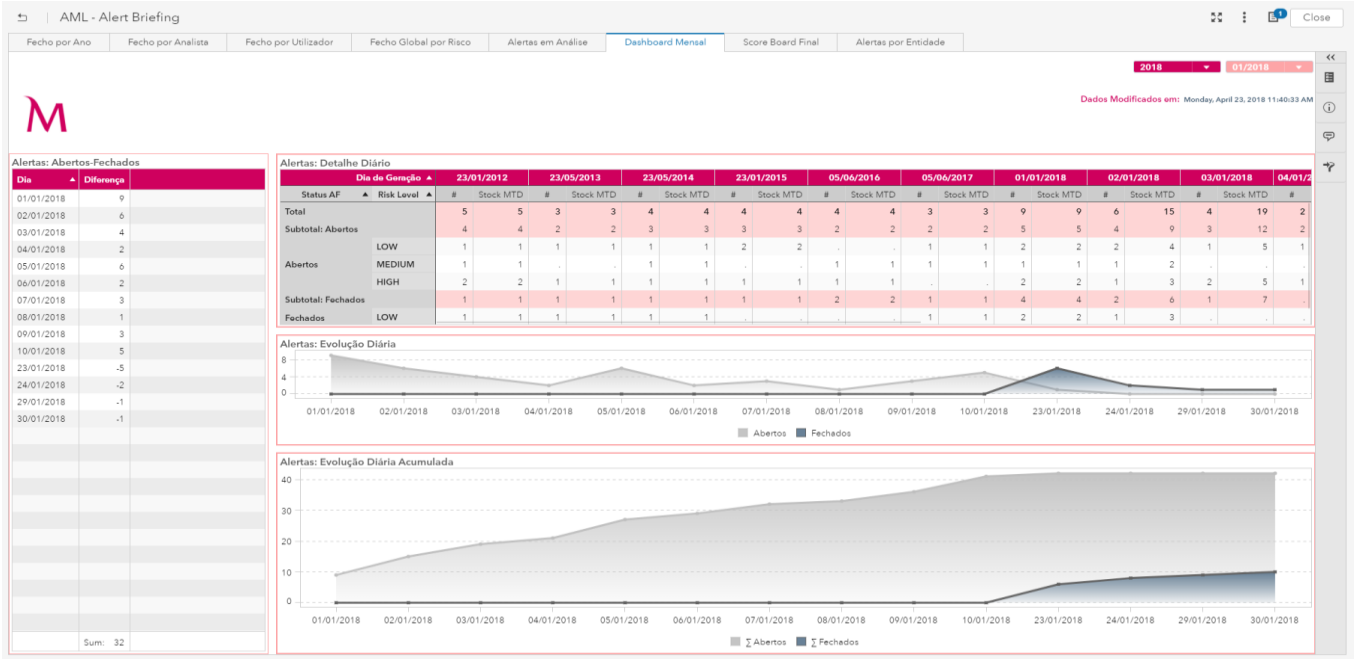

Figure 31 - Dashboard of the generated and analyzed alerts (Source: SAS-Bank's Solution)

The default reports, the reports implemented as the system starts, will cover four main thematic, the alert information, the scenario administration, the system administration and the performance. The alert information will be represented by a set of five reports that will allow the visualization of the alert metrics, the age of active alerts, the duration of the alerts, the alerts created across the time and the details of the alerts, providing a view on the alerts from all the possible perspectives, facilitating a better comprehension of the alert population, outputted by the system. The scenario administration will be supported by two reports, where the active thematic of the scenarios and risk factors will be show and detailed, so that a gain on the understanding the behaviors being analyzed

by the system can be achieved. The system administration will allow for the visualization of the system's statists, viewing its users by their capabilities and responsibilities, their group and role, and the batch execution specifications and statistics, which as a whole will provide key operational technical details necessary for the administration of the system (uncover needs of complementation on certain roles, capabilities, etc.). The performance is explained and represented by a set of seven reports, having, the alert closing process explained in relation with the year, the intervenient, the user and the risk; the details of the alerts still in analyses/processing illustrated; a monthly dashboard, where all the key performance indicators are shown, such as alerts generated, alerts closed, their respective context, and much more, this in a monthly temporal series, so that a view on the evolution of the performance can be gained; and a final scoreboard report that will provide view on the alerts in a segmented way, providing visualizations on the alerts, in accordance with their specifications, so that a better understanding of the impact of these characteristics have on their treatment and handling.

6.12. SYSTEM'S ASSUMPTIONS/RESTRICTIONS

Until now, each of the AML&CFT system's component, thematic, mechanic, logic, intervenient and view, have been introduced and explained, but always as an independent factor, a factor that would not have impact on another factor, but this is not true, some will be impacted by others, and as of this, the current chapter will dictate assumptions/restrictions made to factors, so that the impacts would be minimized.

Two main groups of restrictions were necessary, referring the first group to restrictions/assumptions that take form from the interactions/combinations of the customer's population (segment) with the scenarios and risk factors, and the second from the scenarios with the customer's type and risk level.

Regarding the first group, the restrictions/assumptions were made from the customer's population perspectives, since there were scenarios that did not fit in the context of certain customer's populations, being even some scenarios created specifically for certain populations. The restrictions/assumptions that resulted of this were:

For the customer populations Particular, PB & CGV, ENI, SME and CORP, the risk factor 1 and the scenarios 18 and 19 will not considered. The risk factor is not adjusted since it evaluates the transactional risk behavior of high velocity funds and there are scenarios that validate this thematic so a redundancy of this is unnecessary. The scenarios 18 and 19, which does not have much logic since the population of individuals in question is Particular.

On the customer population DINT, the only scenarios that will validate it will be the 1, 20, 21, 22 and 23, being these exclusive of this population, since they evaluate behaviors characteristic of the population.

In relation with the second group of restrictions/assumptions, the perspective adopted was the scenarios. This way, scenarios that did not require certain specifications or details, or that were specific to certain types of customers, thanks to their logical nature, will be simplified and directed to their optimal form. They were:

For the scenario 16, the customer risk level will not be considered since it compares the customer profile with their own behavior, existing this only for the parameter that refers to the frequency of deposits.

For the scenario 17, the customer risk level will not be considered since it compares the customer profile with their own behavior, existing this only for the parameter that refers to the frequency of withdrawals.

For the scenario 18, the customer risk level and type, will not be considered since it compares the customer profile with their own behavior in regards of unexpected activity.

For the scenario 19, the customer risk level and type, will not be considered since it compares the customer profile with their own behavior in regards of unexpected high velocity funds activity in the account level.

For the scenario 20, the segmentation logic was implemented, but since the entity to be shown in the triage will be an external entity, an entity that is not a client of this bank, and since the system will not have enough information to characterize it, all the entities will be characterized as PART.

For the scenario 22, the segmentation logic was implemented, but since the entity to be shown in the triage will be an external entity, an entity that is not a client of this bank, and since the system will not have enough information to characterize it, all the entities will be characterized as PART.

In addition to the two previously presented groups, another restriction/assumption group exists. This group was not presented in conjunction with other two because it does not derive from the impact of multiple factors coexisting, like the others, but from the non-existence of supporting capabilities in order to ensure the correct usage and/or calculation of these factors. In this group are included the CDD 13, 15 and 19; and the scenario 20 and 22.

7. ONGOING PHASES

The current stage at which the project is, is the start of production, meaning, the system has already been correctly implemented, tested in a controlled environment, and been fed segments of the regular transactional data, in order to validate all of its functionalities, being now the start of the feeding of the production transactional data (real data). At this stage, none of the validation and optimization algorithms can be tested, since the data to be feed to then has only just started to be produced, so the methodology to be used in this chapter is of advisement and planning on what to do/use/implement in order to validate and optimize the various key points of the system. This will be done having as basis the theoretical explanations of the algorithms and their inherent characteristics (refer to the Literature Review chapter, at the algorithms' respective reference), and the specifications of the key points in question. It is proposed to use the RFM algorithm results, in the context of the key point, as a transactional like attribute calculator. It is also to be noted that in each implementation, it is advisable to do a data exploration/analysis phase, whereas the business needs are understood; the data prepared, by means of imputing missing & invalid values and filtering outliers; and explored the combinations of various variables and their transformation.

7.1. POPULATIONS VALIDATION

The populations' validation has as objective the assessment of the segmentation of the customers into segments implemented, and as it was previously referred, the one implemented was defined as a result of a mix of small analysis on the previous AML&CFT system's segmentation and business empiric knowledge. A new segmentation will be calculated, without the old system's premises, in order to validate the one implemented. So, the input that will be used is the large customers' database but using the variables that are available as of the entering of a new customer, which consist of a low-medium dimensionality.

Taking into consideration this validation objective, and the characteristics of the data to be used, the algorithm proposed is unsupervised, on the field of clustering, namely the CURE and the BIRCH.

7.2. POPULATIONS OPTIMIZATION

The populations' optimization main objective is the fine tuning of each of the previously presented population, by means of sub populations' creation. This translates into; the customers associated to each population being the input of the optimization, having the optimization of the population a lower volume of data than the one used in the validation, but requiring the algorithm to be run as many times as there are populations (at least). On the other hand, the optimization will have a higher dimensionality than the validation, because the optimization will use transactional behavior variables in order to cluster the customers.

Taking this into consideration, the SOM algorithm will be used firstly, in order to obtain some insights on the context of the transactional behavior. After, the algorithms proposed for the population validation, CURE and BIRCH, can used, but using the insights regarding the transaction behavior taken by using the SOM, as a visualization tool.

7.3. SCENARIO VALIDATION

The scenario validation has as its main objective the identification of scenarios that are creating more noise that effectively contributing to the detection of suspicious behaviors. This analysis can be made with a simple frequency analysis, to when it comes to an individual scenario, but in the detection of combinations of scenarios that are being triggered together, this approach fails. As such it is advisable to use/implement an Association Rules algorithm, just like the A priori one. This will help identify scenarios that are just incrementing the alert count by only generating alerts when other scenarios are triggered, the validation of the same suspicious behaviors.

The main approach to correct this is inactivating such scenarios, but in the 'Conclusion/Future Work' another approach will be referred.

7.4. THRESHOLD OPTIMIZATION

The thresholds' optimization main objective is the determination of the optimal value which to assign to the thresholds of a specific scenario, so that it will alert efficiently to suspicious behaviors. This means that the data to be used in the optimization, would be the transactional behavior that contributed to the generation of a specific scenario alert, as well as the type of resolution that was made on the alert (if the suspicious behavior was indeed suspicious or not). This means that both the volume and dimensionality of data will be medium like (It will depend on the scenario at hand).

Taking this into consideration the algorithm to be implemented can be both supervised and unsupervised, but always with a regression like logic, focused on the values, so as a proposition the implementation of artificial neural networks, SOM, SVM, Decision Trees and Ensembles is advisable.

7.5. RISK FACTORS VALIDATION

The risk factors' validation objective is to ensure that the risk factors cover the correct characterization of the priority risk of the alert entity. In order to do this, a clustering algorithm will be used, creating clusters of customers in order to characterize individuals in accordance with the effective conclusion of the alert. This will be done using the customers' transactional data, within the risk factors' contextualization, which will translate into both a medium volume and a sized dimensionality. Taking all this into account, it is advisable to implement the SOM and Decision Trees algorithms in order to assess the both the identity and the weight of the risk factors that in fact should considered in order to rank the alerts.

7.6. CDDS VALIDATION

The CDDs' validation has as its main objective the assessment of the risk factors that are being considered. This validation will be made, by comparing the results of a clustering algorithm with three clusters (High, Medium and Low risk) in regards with the risk factors' components that characterize the clusters, with the ones that are currently being used. This will be done in accordance with the type of intervenient, so the data to be used is the customers' data by type, having a

medium-large volume of records, and since the context is of CDDs' the dimensionality is medium sized.

Considering all this, the algorithms to be implemented must be of clustering, with the capability of limiting the number of clusters, being advised to implement the K-means (and its variants). If the results are not the desired, being it due to the computational complexity, or result accuracy, then it is advised to pass onto the remaining algorithms, BIRCH, CURE and SOM.

8. CONCLUSIONS

The Conclusion chapter will serve as the projects' present situation assessment point. In this regard, and according with the defined objectives, the implemented AML&TF system is fully functional, as assessed by the user tests, meeting the requirements imposed by legal authorities; it validates the client's transactional behavior and not just a view/part of the transactional operations, being more adjusted to spotting of suspicious behaviors by analyzing the client as the main entity and not their accounts alone; and it ca be assessed by usage of validation and optimization methodologies, on the system's key components. At the current stage of the project, it has been implemented a fully functional and efficient behavior-based AML/CTF system, with capabilities and delimited strategies, in order to enable a continuous fine-tuning and optimization across its life cycle.

As of to sustain even more what was previously said, a series of questioners/interviews[63] were made. These questioners/interviews where made with five main building blocks in mind, the first one referring to the characterization the interviewee; the second to the system's response to the legal requirements; the third to the behavior analysis' capabilities of the system; the forth to the optimization capabilities of the system, and finally, the fifth, refers to the future work to be done.

The interviews were conducted in a population with an average of more than twenty years of experience in the banking sector and an average compliance experience of five to ten years. It is possible to conclude from this that the population interviewed is highly experienced and as such their answers will be relevant and sustained. It is also to be noted that the population covered the entirety of the intervenient roles, achieving a full view of the system.

8.1. FEEDBACK

When questioned in regards of their knowledge of the legal requirements imposed to the compliance office, the unanimous response was of full knowledge. The interviewees were also asked on the respect of such requirements by the system, and similarly to the previous answer, the majority validated the system's respect for the requirements, classifying this respect as high and highly adaptive to future changes on these requirements. The remaining population did not affirm this, but also did not affirm the contrary, that the system did not respect, they simply did not answer because they felt like they had no knowledge in this regard, so they did not answer.

When inquired about the system's capability to analyze the client's suspicious behaviors, and sustain the decision-making process about such behavior, the unanimous response was affirmative. The reasons given, as of why were, the fact that is uses tailored scenarios for the currently practiced ML/TF behaviors; the fact that it provides useful information that allows for the detection of the true ML/TF activity; the quality of the data that it has on the risk of the client; the combination of information about the transactions, risk, products in question; The adjustment of the data on the alerted behavior in order to sustain a decision; the realization of accurate analysis, based on different types of populations and risks; the quality and quantity of demographic information on the customer,

[63] Formulary used during the interview on the Appendix Chapter

and its transactional information details; the variety and cover area of the scenarios used; the availability of different roles for different profiles.

In regard to the change of system, and its inherent view on the notion of behavior, every single interviewee answered that the change was extremely beneficial. The given reasons were all around the increment of the adequate information in regards to the analysis of the customer behavior; the way that such information was presented and displayed, being in more easy to read and comfortable format; the fact that it now has an open model, which allows for a continuous tailoring of the scenarios; the increment of transparency; the change on the vision, now by client, allowing for a better notion of its complete behavior and profile.

In relation to the system's optimization capabilities, the large majority of the population did not answer, this is due to the centralization of this knowledge in single team, being that the analyses of this part of the interviews will be restricted to this part of the population.

When asked about the system's flexibility in order to support its logic optimization and validation across its life cycle, the unanimous opinion was of total agreement, with the requirement of being frequently calibrated and complemented with the usage of tools from SAS or Open Source in order to develop such processes. It was also referred that the ability to create new scenarios and/or revise the existing ones; the easiness of update on the scenarios, with new thresholds and new populations; and the capability of testing the new scenarios or new parameters with production data, are key points in the continuous sustained optimization of the system.

In regards to the impact of system change, once again the unanimous opinion was that it was extremely beneficial, since now there is the capability of using the complementary tools in order to enrich the processes; the existence of a richer approach on transactional analysis; the capability to map different type of information, based on the lookup tables; the availability of more information in order to improve the quality of the analyses; and the fact that all the information necessary is stored and revised to raise the alerts, being ready to use as is.

The interviewees when inquired about the future work to be done in the system, noted that all the thematic surrounding the system should be a work to be done in a continuous way, as of to ensure the correct functioning of it. It was then asked to give an order of priority to the themes, and the results were as follows: Population Optimization; Population Validation; Threshold Optimization; Scenario Validation; Risk Factors Validation; CDD Validation; System's Usability; and System's Analysis[64].This can be used as an order for the implementation process of the algorithms proposed in the Experimental / Theoretical Results chapter.

The system was also assessed by a team of consultants (international consultants), and their findings deemed the implemented system compliant with the best practices red flags (in accordance with their audit experience and the demands of over 30 regulators, regulatory bodies and standard setters, like Egmont group, FAFT, ESA, FCA, WOLFSBERG); that it is aligned with the logical best practices (in accordance with a top down analysis), from the transaction coverage to the threshold calibration. Concluding that the system's coverage seems to be about 90% of the total standards setters' scenarios, being more efficient, assertive, flexible, customizable, and a generator of more

[64] Ordered form higher priority to lower.

assertive alerts (fewer ‘false-positive’ alerts).The alert generation is aligned with peers, which indicate an appropriate calibration, but must be done continuously.

8.2. FINAL STATEMENTS

As of all that was introduced and stated it is possible to conclude that this project was and is of extreme importance to the financial institution where it toke place, touching on highly sensitive and important topics, that in case of bad management could impact the whole organization. As such its correct and efficient implementation was mandatory, and its current and future validation and optimization are fundamental for the well-being of the organization, being the work developed will sustain such approach. Personally wise, the project was extremely useful and interesting, given its large dimension, meaning, the project covered a multitude of topics whereas the knowledge acquired during the academic course were employed, having the opportunity to experience and make use such knowledge in a real world context. At this stage, some of this knowledge was employed in a planning manner, but being that it is now starting to be put into practice. The academy, with this thesis, gained transversal knowledge on the implementation of an efficient and functional AML/CTF system, and on its intrinsic logic and methodology. There were also some referred themes that may be of investigation interest.

9. FUTURE WORKS

This chapter will have as objective, the delimitation of the next steps to be done in accordance with the project's scope. These next steps can be aggregated into two groups, the implementation and the investigation group. The implementation group refers to the implementation of the algorithms referred in the Experimental / Theoretical Results chapter, in the order of priority dictated by the system's results typology. The Investigated group is constituted by possible investigation work that can be done in order to optimize the methodologies presented. One of these is the investigation of Genetic Programming, more specifically, Geometric Semantics Genetic Programming[65] in the context of scenario optimization and aggregation (per example by typology).

[65](Castelli, Manzoni, Vanneschi, Silva & Popovič, 2016)

10.BIBLIOGRAPHY

10.1. PAPERS & BOOKS

Agrawal, Rakesh & Srikant, Ramakrishnan, 1994, An overview of gradient descent optimization algorithms

Anil, K. Jain, 2010, Data clustering: 50 years beyond K-means q, *Pattern Recognition Letters*

Aone, Chinatsu & Larsen, Bjornar, 1999, Fast and Effective Text Mining Using Linear-time Document Clustering

Arthur, David & Vassilvitskii, Sergei, 2006, k-means++: The Advantages of Careful Seeding

Association of Certified Anti-Money Laundering Specialists, 2016, Anti-Money Laundering Compliance Program, Chapter 4

Association of Certified Anti-Money Laundering Specialists, 2016, Implementing an Effective Anti-Money Laundering System

Batistakis, Yannis, Halkidi, Maria & Vazirgiannis, Michalis, 2001, On Clustering Validation *Techniques, Journal of Intelligent Information Systems*

Bergerud, A. Wendy, 1996, Introduction to Logistic Regression Models, (British Colombia)

Bhambu, Lekha & Srivatava, K. Durgesh, 2005, DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE

Breiman, Leo, 1994, Bagging Predictors

Breiman, Leo , 2001, Random Forests

Castelli, M., Manzoni L., Vanneschi, L., Silva, S., & Popovič A., 2016, Self-tuning geometric semantic Genetic Programming, Genetic Programming and evolvable machines

Chih-Chung, Chang, Chih-Jen, Lin & Chih-Wei, Hsu, 2003, A Practical Guide to Support Vector Classification

Counter-Terrorism Implementation Task Force, 2009, CTITF Working Group Report Tackling the Financing of Terrorism

Cun, le Yann, 1988, A Theoretical framework for Back-Propagation

Deshpande, S.D. & Harne, R. Pooja, 2015, Mining of Association Rule- A Review Paper, *International Journal of Engineering Technology, Management and Applied Sciences*

Edmonds, Tim, 2018, Money Laundering Law

Ettaouil, Mohamed , Ghanou, Youssef , Idrissi, Janati Amine Mohammed & Ramchoun, Hassan, 2016, Multilayer Perceptron: Architecture Optimization and Training, *IJIMAI*

European Investment Bank Group, 2018, Anti-Money Laundering and Combating Financing of Terrorism Framework

Flynn & Jain, Murty, 1999, Data Clustering: A Review. ACM Computing Surveys

Freund, Yoav & Schapire, E. Robert , 1997, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, *Journal of computer and system sciences 55*

Galeano, Pedro, Joseph, Esdras & Lillo, E. Rosa, 2013, The Mahalanobis distance for functional data with applications to classification

Gao, Xuedong, Li, Zhongmou, Liu, Yanchi, Wu, Junjie & Xiong, Hui, 2010, Understanding of Internal Clustering Validation Measures

Ghosh, Joydeep, Mooney, Raymond & Strehl, Alexander, 2000, Impact of Similarity Measures on Web-page Clustering

Guha, Sudipto, Rastogi, Rajeev & Shin, Kyuseok, 1998, CURE: an efficient clustering algorithm for large databases

Hears, A. Marti, 2017, Support vector machines

Hoque, Ziaul Mohammed, 2009, Measures against Money Laundering and Terrorism Financing: An analytical Overview of Legal Aspects

Kohonen, Teuvo, 1990, the Self-Organizing Map

Livny, Miron, Ramakrishnan, Raghu & Zhang, Tian, 2014 , BIRCH: An Efficient Data Clustering Method for Very Large Databases

Lobo, V., 2005, Sistemas de Apoio à Decisão– Redes de Kohonen (SOM), *NOVAIMS*

Macready, G. William & Wolpert, H. David , 1997, No Free Lunch Theorems for Optimization

Maind, B. Sonali. & Wankar, Priyanka, 2014, Research Paper on Basic of Artificial Neural Network, *International Journal on Recent and Innovation Trends in Computing and Communication*
Miljković, Dubravko, 2017, Brief Review of Self-Organizing Maps

Rajarajeswari, A. & Ravindran, Malathi R., 2015, A Comparative Study of K-Means, K-Medoid and Enhanced K-Medoid Algorithms, *International Journal of Advance Foundation and Research in Computer (IJAFRC)*

Raphael, Benny , Saitta, Sandro & Smith, F. C. , 2007, A Bounded Index for Cluster Validity, *MLDM*
Rendón, et al., 2008, ORGANIZATIONAL ASSESSMENT OF PROCUREMENT MANAGEMENT PROCESSES
Ruder, Sebastian, 2017, An overview of gradient descent optimization algorithms

Touil, Karima, in the name of Association of Certified Anti-Money Laundering Specialists, 2016, Risk-Based Approach, Understanding and Implementing, challenges between risk appetite and compliance

10.2. WEBSITES

Banco de Portugal, 2018, Page of Banco de Portugal (URL: https://www.bportugal.pt/en/page/branqueamento-de-capitais-e-financiamento-do-terrorismo, consulted on 07-01-2018)

CE MATH, 2018, Page of CE MATH (URL: http://www.wseas.us/e-library/conferences/2011/Mexico/CEMATH/CEMATH-26.pdf, consulted on 14-01-2018)

Comissão de Mercados de valore Imobiliários, 2018, Page of Comissão de Mercados de valores Imobiliários (URL: http://www.cmvm.pt/en/Cooperacao/cmvm_int_activity/Pages/internacional_act_CMVM.aspx, consulted on 07-01-2018)

Comissão de Mercados de valore Imobiliários, 2018, Page of Comissão de Mercados de valores Imobiliários (URL: http://www.cmvm.pt/en/The_CMVM/Overview/Pages/Who-We-Are.aspx , consulted on 07-01-2018)

Corporate Finance Institute, 2018, Page of Corporate Finance Institute (URL: https://corporatefinanceinstitute.com/resources/knowledge/accounting/audit-sampling/, consulted on 10-01-2018)

Financial Action Task Force, 2018, Page of Financial Action Task Force (URL: http://www.un.org/en/terrorism/ctitf/pdfs/ctitf_financing_eng_final.pdf, consulted on 06-01-2018)

Harvard, 2014, Page of Havard (URL: http://www.math.harvard.edu/~lurie/281notes/Lecture17-Additivity.pdf , consulted on 20-02-2018)

No-Free-Lunch, 2018, Page of No-Free-Lunch (URL: https://www.sas.com/en_us/software/visual-analytics.html#m=sas-viya-features, consulted on 13-01-2018)

SAS, 2018, Page of SAS (URL: https://support.sas.com/en/software/visual-analytics/7-4.html, consulted on 18-01-2018)

SAS, 2018, Page of SAS (URL: https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-visual-analytics-on-sas-viya-108779.pdf, consulted on 14-01-2018)

SAS, 2018, Page of SAS (URL: https://www.sas.com/en_us/software/visual-analytics.html#m=sas-viya-features, consulted on 18-01-2018)

Silverman, G. Stuart , Tuncali, Kemal & Zou, H. Kelly, 2003, Correlation and Simple Linear Regression

Stanford, 2018, Page of Stanford (URL: https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html , consulted on 14-01-2018)

Statistics How To, 2018, Page of Statistics How To (URL: http://www.statisticshowto.com/mahalanobis-distance/, consulted on 13-01-2018)

United Nations Industrial Development Organization, 2018, Page of United Nations Industrial Development Organization (URL: http://www.cmvm.pt/en/The_CMVM/Overview/Pages/Who-We-Are.aspx, consulted on 10-01-2018)

Whittington & Associates, 2018, Page of Whittington & Associates (URL: https://www.whittingtonassociates.com/2012/12/audit-sampling/, consulted on 10-01-2018)

11.ANNEXES

Original Date:	
Date Reviewed:	

AML/CFT SYSTEM'S QUESTIONNAIRE

All questions contained in this questionnaire are strictly confidential
and will used in the writing of an academic thesis.

Name *(Last, First, M.I.):*	☐ M ☐ F	**Function:**
Banking Experience (Years):	☐ Exp < 2 ☐ 2≤ Exp < 5 ☐ 5≤ Exp < 10 ☐ 10≤ Exp < 15 ☐ 15≤ Exp < 20 ☐ Exp ≥ 20	
Compliance Experience (Years):	☐ Exp < 2 ☐ 2≤ Exp < 5 ☐ 5≤ Exp < 10 ☐ 10≤ Exp < 15 ☐ 15≤ Exp < 20 ☐ Exp ≥ 20	

FUNCTION SPECIFICATIONS

Professional Area:	☐ AML/CTF Monitoring Department ☐ Information Systems and Analytics Department ☐ Other	
Functional Role:	☐ Analyst	☐ Administrator
	☐ Investigator	☐ IT Team Member
	☐ Manager	☐ Compliance Committee Member
Additional Information		

LEGAL REQUIREMENTS

ALL QUESTIONS CONTAINED IN THIS QUESTIONNAIRE ARE OPTIONAL AND WILL BE KEPT STRICTLY CONFIDENTIAL.						
Knowledge	☐ No knowledge in regards of the requirements imposed					
	☐ Knowledge in regards of the requirements imposed but not on the system's response to them					
	☐ Knowledge in regards of the requirements imposed and on the system's response to them					
System	Does the system meet the actual legal requirements?				☐ Yes	☐ No
	Does the system have adaptive capability to answer to changes in the requirements?				☐ Yes	☐ No
	If the answer was "Yes", then classify the adaptability level:					
	Adaptability Level	☐ High	☐ Medium	☐ Low		

BEHAVIOR ANALYSIS

ALL QUESTIONS CONTAINED IN THIS QUESTIONNAIRE ARE OPTIONAL AND WILL BE KEPT STRICTLY CONFIDENTIAL.
Does the system analyze the clients' suspicious behavior, sustaining the decisions with data patterns and analyses? Explain.
Was the change of system, and its inherent view on the notion of behavior, beneficial? Explain.

SYSTEM'S OPTIMIZATION

ALL QUESTIONS CONTAINED IN THIS QUESTIONNAIRE ARE OPTIONAL AND WILL BE KEPT STRICTLY CONFIDENTIAL.
The system has flexibility on the optimization and validation of its logic, supporting its life cycle? Explain.
Was the change of system, an improvement of on the adaptability, optimization and validation of the system with the transactional context? Explain.

FUTURE WORK - THEMATICS

ALL QUESTIONS CONTAINED IN THIS QUESTIONNAIRE ARE OPTIONAL AND WILL BE KEPT STRICTLY CONFIDENTIAL.		
☐ Population Validation	☐ Population Optimization	☐ Scenario Validation
☐ Threshold Optimization	☐ Risk Factors Validation	☐ CDD Validation
☐ System's Usability	☐ System's Analysis	☐ Other:

High Risk Country List:

ISO-3	Name
AFG	Afghanistan
AND	Andorra
AIA	Anguilla
ATG	Antigua And Barbuda
NAM	Namibia
ABW	Aruba
BHS	Bahamas
BRB	Barbados
BHR	Bahrain
BLZ	Belize
BLR	Belarus
BOL	Bolivia
BIH	Bosnia and Herzegovina
BRN	Brunei Darussalam
CPV	Cape Verde
CHN	China
CRI	Costa Rica
CUB	Cuba
DJI	Djibouti
DMA	Dominica
EGY	Egypt
ARE	United Arab Emirates
ERI	Eritrea
ETH	Ethiopia
FSM	Micronesia, Federated States of
PHL	Philippines
GMB	Gambia
GIB	Gibraltar
GRD	Grenada
GTM	Guatemala
GUY	Guyana
GNB	Guinea-Bissau
HTI	Haiti
HND	Honduras
GUM	Guam
KIR	Kiribati
NIU	Niue
TKL	Tokelau

PCN	Pitcairn
SPM	Saint Pierre and Miquelon
NFK	Norfolk Island
SGS	South Georgia and the South Sandwich Islands
TUV	Tuvalu
BMU	Bermuda
CCK	Cocos (Keeling) Islands
COK	Cook Islands
GGY	Guernsey
FLK	Falkland Islands (Malvinas)
FJI	Fiji
MDV	Maldives
MNP	Northern Mariana Islands
MHL	Marshall Islands
UMI	United States Minor Outlying Islands
CXR	Christmas Island
PLW	Palau
SLB	Solomon Islands
SJM	Svalbard and Jan Mayen
TCA	Turks and Caicos Islands
VGB	Virgin Islands, British
VIR	Virgin Islands, U.S.
IDN	Indonesia
IRQ	Iraq
IRN	Iran, Islamic Republic of
JAM	Jamaica
JOR	Jordan
KWT	Kuwait
LBN	Lebanon
LBR	Liberia
LBY	Libyan Arab Jamahiriya
MYS	Malaysia
MMR	Myanmar
MDA	Moldova, Republic of
MCO	Monaco
MSR	Montserrat
NRU	Nauru
NGA	Nigeria
PAN	Panama
PAK	Pakistan

PYF	French Polynesia
PRI	Puerto Rico
QAT	Qatar
YEM	Yemen
CAF	Central African Republic
VUT	Vanuatu
COD	Congo, the Democratic Republic of the
GIN	Guinea
PRK	Korea, Democratic People's Republic of
RUS	Russian Federation
ASM	American Samoa
WSM	Samoa
SHN	Saint Helena
LCA	Saint Lucia
KNA	Saint Kitts and Nevis
SMR	San Marino
STP	Sao Tome and Principe
VCT	Saint Vincent and the Grenadines
SYC	Seychelles
SGP	Singapore
SRB	Serbia
SYR	Syrian Arab Republic
SOM	Somalia
LKA	Sri Lanka
SWZ	Swaziland
SDN	Sudan
OMN	Oman
TON	Tonga
TTO	Trinidad and Tobago
TUN	Tunisia
TKM	Turkmenistan
TUR	Turkey
UKR	Ukraine
UZB	Uzbekistan
VEN	Venezuela
ZWE	Zimbabwe
AGO	Angola
DZA	Algeria
AZE	Azerbaijan
BGD	Bangladesh
BWA	Botswana

BDI	Burundi
CMR	Cameroon
KHM	Cambodia
KAZ	Kazakhstan
TCD	Chad
CYP	Cyprus
COM	Comoros
SLV	El Salvador
ECU	Ecuador
EST	Estonia
GAB	Gabon
GHA	Ghana
GNQ	Equatorial Guinea
HKG	Hong Kong
CYM	Cayman Islands
IMN	Isle of Man
JEY	Jersey
LAO	Lao People's Democratic Republic
LVA	Latvia
LIE	Liechtenstein
LTU	Lithuania
MDG	Madagascar
MWI	Malawi
MLI	Mali
MLT	Malta
MUS	Mauritius
MRT	Mauritania
MEX	Mexico
MOZ	Mozambique
NPL	Nepal
NIC	Nicaragua
NER	Niger
PNG	Papua New Guinea
PRY	Paraguay
KEN	Kenya
KGZ	Kyrgyzstan
MAC	Macao
COG	Congo
DOM	Dominican Republic
SLE	Sierra Leone
SSD	South Sudan

CHE	Switzerland
TJK	Tajikistan
TGO	Togo
UGA	Uganda
URY	Uruguay