NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

Identificação de segmentos de risco numa Instituição Financeira

Aplicação de técnicas de Data Mining a dados de Risco Operacional

André Miguel Alves Rodrigues Godinho Sobral

Proposta de Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação, Especialização em Business Intelligence

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

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

IDENTIFICAÇÃO DE SEGMENTOS DE RISCO NUMA INSTUIÇÃO FINANCEIRA

Aplicação de técnicas de Data Mining a dados de Risco Operacional

por

André Miguel Alves Rodrigues Godinho Sobral

Proposta de Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação, Especialização em Business Intelligence

Orientador: Rui Alexandre Henriques Gonçalves

Novembro 2015

DEDICATÓRIA

Dedico esta dissertação de mestrado aos meus pais. Em primeiro lugar, quero agradecer-lhes toda a paciência, dedicação e amor que sempre manifestaram por mim, cada um à sua maneira, claro. Quero agradecer-lhes pela confiança que sempre depositaram em mim, e pelos ensinamentos que todos os dias me transmitem. Ensinamentos que vão desde a noção mais básica de como ser uma pessoa consciente e decente neste mundo por vezes injusto e cruel, até à mais ríspida palestra sobre deveres e afazeres, muitas vezes aborrecida e difícil de digerir. Todos estes ensinamentos fizeram de mim a pessoa que hoje sou e são estas as bases que levo comigo no futuro. Em segundo lugar, quero agradecer a possibilidade que me deram de prosseguir os meus estudos até ao ponto que eu próprio achei ser suficiente, sem nunca me colocarem nenhum tipo de condição a não ser a de que eu desse o meu melhor.

A ti mãe, a ti pai, obrigado, espero que fiquem orgulhosos.

.

AGRADECIMENTOS

Começo por agradecer novamente aos meus pais por me terem possibilitado todas as condições para a conclusão este trabalho.

Quero agradecer também à minha namorada. Rita, obrigado por seres a minha melhor amiga, obrigada pelo carinho e pela força que todos os dias me transmites. Obrigado também pelas dicas, sugestões e eventuais puxões de orelhas, acredita que tudo me ajudou a chegar até aqui.

Por último, quero agradecer ao meu orientador, Prof. Rui Gonçalves, por ter aceitado trabalhar comigo, mesmo quando eu ainda nem tinha uma ideia concreta sobre o que iria fazer. O professor acreditou em mim e no trabalho que me propus a desenvolver, ajudando-me quando necessitei, obrigado.

RESUMO

Nas últimas duas décadas, o risco operacional assumiu-se como uma disciplina de máxima importância por si só. Nesse sentido, as instituições no seu geral, e em particular da área financeira, começaram a atribuir-lhe uma especial atenção, para além da que sempre deram a outros tipos de riscos mais antigos e conhecidos, nomeadamente o de mercado e o de crédito. Uma das armas que as instituições financeiras têm ao seu dispor para se precaverem contra este tipo de risco centra-se na forma como conseguem desenvolver processos analíticos que potenciem dois dos seus maiores activos: por um lado o aumento do poder computacional a um menor preço e, por outro, as quantidades continuamente crescentes de dados que são produzidas e armazenadas. Desta forma, os dados são inquestionavelmente vistos como o recurso mais abundante nas organizações, contudo, é a capacidade de analisar esses dados e deles extrair informação que possa gerar conhecimento, a força motora responsável pelo aumento da produtividade, da inovação e das vantagens competitivas destas mesmas organizações. Esta necessidade originou o desenvolvimento do processo de *Knowledge Discovery on Databases (KDD)* e do seu componente analítico mais complexo e relevante, o Data Mining. Este último surge com base em diversas áreas como a estatística, a inteligência artificial ou a ciência computacional, tendo nas últimas duas décadas ganhado bastante relevância tanto a nível académico como a nível organizacional, pela forma como: (i) possibilita a exploração e a respectiva análise de grandes quantidades de dados; (ii) através da análise possibilita a descoberta de relações escondidas e de padrões de semelhança entre os dados; (iii) permite ao analista inferir conclusões sobre essas mesmas relações e padrões, conferindo novo conhecimento preditivo capaz de potenciar actividades de negócio e daí retirar proveitos em relação à concorrência. No caso dos bancos, estas ferramentas proporcionam uma melhor análise dos seus desempenhos, da sua realidade e, ajudam na tomada de decisões, como por exemplo na concessão, ou não, de créditos. Por outras palavras, analisando o passado e o presente, o Data Mining descobre as bases para prever o futuro.

PALAVRAS-CHAVE

Risco Operacional; Dados; Data Mining; KDD; Instituições Financeiras

ABSTRACT

During the last two decades, operational risk has become a subject of great importance. As so, institutions in general, and those within the financial sector in particular, started to pay more attention to it, in addition to the attention they were already paying to other long-standing and well-known types of risk, namely, those of market and credit risk. One of the many tools used by financial institutions to prevent against operational risk concerns the way these can develop analytical processes that strengthen two of their biggest actives: on the one side, the intensification of computer power at a lower cost, and, on the other side, the increasingly growing quantity of data that is produced and stored. This way, data is unquestionably seen as the most plentiful resource in organisations. However, it is the capacity to analyse data, and to extract information that carries some knowledge, that is the driving force responsible for the productivity growth, innovation and competitive advantages of organisations. This necessity initiated the development of the *Knowledge Discovery on Databases* (KDD) process, and its most complex analytical component, Data Mining. The latter has its origins in various subjects, such as statistics, artificial intelligence, or computer science. Data Mining has gained major relevance in the last two decades, both academically and at the organisational level, having this mainly to do with: (i) how it allows to explore and analyse large quantities of data; (ii) through the analysis, how it allows to uncover hidden relationships and similarity patterns amongst data; (iii) how it allows the analyst to withdraw conclusions about these relationships and patterns, achieving new predictive knowledge capable of increasing business activities and, as a consequence, displaying more advantages over the competition. When it comes to the banking sector, these tools allow for a better performance analysis of the reality and help in decision-making processes, for instance, whether a credit granting should be or should not be attributed/permitted. In other words, due to its capacity to analyse both past and present, Data Mining is able to find the basis that allow to predict the future.

KEYWORDS

Operational Risk, Data, Data Mining, KDD, Financial Institutions

ÍNDICE

1. INTRODUÇÃO 1
   1.1. Contextualização e Relevância do tema 2
   1.2. Problema de Estudo e Objectivos da Tese 6
   1.3. Estrutura do Trabalho 6
2. REVISÃO DA LITERATURA 8
   2.1. Risco Operacional 8
      2.1.1. Características distintas do Risco Operacional 10
   2.2. Knowledge Discovery in Databases (KDD) 13
   2.3. Data Mining 16
      2.3.1. Análise de *Clusters* 20
3. METODOLOGIA DE INVESTIGAÇÃO 25
   3.1. Dados 25
   3.2. Tratamento de Dados 26
4. ANÁLISE DE RESULTADOS 30
   4.1. Análise Exploratória 30
      4.1.1. Eventos Iniciados e Descobertos 30
      4.1.2. Eventos Iniciados não Descobertos 31
      4.1.3. Eventos Iniciados e Terminados 32
      4.1.4. Eventos Iniciados não Terminados 33
      4.1.5. Eventos Descobertos e Terminados 34
      4.1.6. Eventos Descobertos após terem sido Terminados 35
      4.1.7. Eventos Descobertos não Terminados 35
      4.1.8. Processos de Risco 36
      4.1.9. Categorias de Risco 38
   4.2. Análise da Informação Trabalhada 39
      4.2.1. Análise Descritiva da Informação Trabalhada 40
   4.3. Análise de *Clusters* 45
5. CONCLUSÕES 55
6. LIMITAÇÕES E INVESTIGAÇÃO FUTURA 62
7. BIBLIOGRAFIA 63
8. ANEXOS 67

ÍNDICE DE FIGURAS

Figura 1 - Processo *KDD* segundo (Fayyad, Piatetsky-Shapiro, & Smyth, 1996b)……..15

Figura 2 - Processo k-Means segundo (Mooi & Sarstedt, 2011) : topo esquerdo 1ª fase (i), topo direito 2º fase (ii), fundo direito 3º fase (iii), fundo esquerdo 4º fase (iv)……23

Figura 3 - Análise da variável Frequência……30

Figura 4 - Análise da variável Severidade……30

Figura 5 - Análise da variável Severidade……31

Figura 6 - Análise da variável Frequência……31

Figura 7- Análise da variável Severidade……31

Figura 8 - Análise da variável Frequência……31

Figura 9 - Análise da variável Severidade……32

Figura 10 - Análise da variável Frequência……32

Figura 11 - Análise da variável Frequência……33

Figura 12 - Análise da variável severidade……33

Figura 13 - Análise da variável Frequência……33

Figura 14 - Análise da variável Severidade……34

Figura 15 - Análise da variável Frequência……34

Figura 16 - Análise da variável Severidade……35

Figura 17 - Análise da variável Frequência……*36*

Figura 18 - Análise da variável Severidade……37

Figura 19 - Análise da variável Frequência……37

Figura 20 - Análise das quantidades perdidas por intervalos……38

Figura 21 – Análise da duração dos eventos por intervalo……39

Figura 22 – Variáveis utilizadas na análise de *Clusters* pelo *software Miner*……43

Figura 23 - Escolha do número de *Clusters (Software SAS Miner)*……44

Figura 24 - Distribuição dos eventos por *cluster*……*44*

Figura 25 - Visualização das distâncias entre os *Clusters (Software SAS Miner)*……*45*

Figura 26 - Perfil *dos Clusters* em árvore *(Software SAS Miner)*……*45*

Figura 27 - *Segment Profile do primeiro* cluster (Software SAS Miner)……47

Figura 28 - Perfil *dos Clusters* em árvore *(Software SAS Miner)*……*47*

Figura 29 - *Segment Profil*e do segundo *cluster (Software SAS Miner)*……*48*

Figura 30 - *Segment Profile* do terceiro *cluster (Software SAS Miner*) *Software SAS Miner*)……49

Figura 31 - Perfil dos *Clusters* em árvore (*Software SAS Miner*)……49

Figura 32 - *Segment Profile* do quarto *cluster (Software SAS Miner*)……50

Figura 33 - Perfil dos *Clusters* em árvore *(Software SAS Miner)*……*51*

Figura 34 - *Diagrama do SAS Enterprise Miner*……*63*

Figura 35 - Código *SAS* para realizar o Gráfico do Cotovelo……..63

Figura 36 - Perfil em árvore dos *Clusters* (*Software SAS MINER*)……*64*

Figura 37 - *Segment Profile* dos *Clusters*……*65*

ÍNDICE DE TABELAS

Tabela 1 - Casos de estudo referidos em (Jednak & Jednak, 2013)……..4

Tabela 2 - Categorização do Risco Operacional segundo (Jednak & Jednak, 2013)……..10

Tabela 3 - Variável criada referente aos montantes perdidos……..27

Tabela 4 - Variável criada referente à duração de um evento……..27

Tabela 5 - Detecção de um evento antes e após a sua data de termo……28

Tabela 6 – Análise da Detecção de um evento antes e após a sua data de termo……39

Tabela 7 - Comparação da coluna " Perdas" com as restantes colunas criadas……40

Tabela 8 - Comparação da coluna " Duração" com as restantes colunas criadas……40

Tabela 9 - Comparação da coluna " Detecção" com as restantes colunas criadas……41

Tabela 10 - Comparação de Processos e Categorias com as restantes colunas criadas……42

Tabela 11 - Importância das variáveis para a construção dos *Clusters*……*46*

Tabela 12 - Caracterização do primeiro *cluster*……*46*

Tabela 13 – Caracterização do segundo *cluster*……*48*

Tabela 14 – Caracterização do terceiro *cluster*……*49*

Tabela 15 – Caracterização do quarto *cluster*……*50*

LISTA DE SIGLAS E ABREVIATURAS

KDD *Knowledge Discovery on Databases*

CRM *Customer Relationship Management*

OLTP *Online Transaction Processing*

IT Tecnologias de informação

1. INTRODUÇÃO

O conceito de risco tem vindo a assumir um papel crescente no seio da sociedade e das organizações ao longo dos anos, com particular destaque para o risco operacional, no caso das instituições financeiras. Segundo Gonçalves (2011) é o mais antigo tipo de risco que estas instituições enfrentam e encontra-se presente em todas as suas actividades. Cummings e Embrechts (2006) afirmam que estas instituições têm sido sujeitas a riscos operacionais ao longo da sua história, no entanto, apenas na última década tem sido dada a devida importância a esta problemática, quer por parte das entidades reguladoras, quer por gestores ou até mesmo investidores. Buchelt e Unteregger (2004) indo mais longe, admitem que, não sendo novidade, o risco operacional é mais antigo do que o risco de crédito e de mercado, visto que está presente desde o início do funcionamento do sistema financeiro.

O perigo do risco operacional, segundo Jednak e Jednak (2013), reside no facto de ser heterogéneo e muito complexo, tendo a capacidade de produzir perdas de valor incalculável, totalmente desproporcionais à realidade financeira das empresas. Isto acontece porque se trata de um risco que facilmente interage com outros riscos, multiplicando assim o seu efeito nefasto. Nos dias de hoje, o maior desafio do mundo da banca, tal como dos negócios em geral, é a implementação de sistemas de gestão de risco com capacidade de identificar, medir e controlar a exposição das empresas ao próprio risco, seja qual for a sua natureza (Koyuncugil & Ozgulbas, 2008b). Atendendo às palavras de Beroggi e Wallace (1994) foi necessário um esforço comum por parte das organizações para que fossem desenvolvidos sistemas de apoio à decisão em tempo real, capazes de ajudar a prevenir ou até mesmo de mitigar estas consequências indesejáveis provenientes do risco operacional.

Segundo Valverde (2015), o desenvolvimento de sistemas/processos de apoio à tomada de decisão como forma de mitigar riscos ou de criar conhecimento preditivo que, tendo como base o passado e o presente permite-nos antever e analisar o futuro, remete-nos para a questão dos dados. Nas últimas duas décadas temos assistido a um fenómeno de computorização da nossa sociedade, apoiado na evolução e na expansão das tecnologias de informação por todas as organizações independentemente do sector. Esta evolução transformou a forma como as organizações operam, tornando-as globais, dotando-as de inovadoras formas de fazer negócio, quebrando as barreiras físicas e territoriais e conferindo-lhes, por um lado, dimensão e por outro, novas oportunidades de gerar lucro. Este enquadramento evolutivo teve como consequência directa, por um lado, o aumento exponencial do número de dados produzidos por todas as organizações, assim como, a criação de reservatórios capazes de os alocar. Chen, Han e Yu (1996) afirmam que estes reservatórios são utilizados para a recolha de dados provenientes das mais diversas áreas, passando da gestão de negócios e da administração governamental aos diferentes ramos da ciência. Por outras palavras os dados constituem a imagem digital de uma actividade empresarial e continuarão a crescer de forma significativa no futuro. Contudo, e apesar de estes constituírem uma importante oportunidade de aprendizagem e compreensão das diferentes actividades de negócio, este recurso tem sido pouco explorado. Segundo Bisson, Stephenson e Viguerie (2010) embora seja expectável que o volume de dados criados aumente cinco vezes mais nos próximos cinco anos, na melhor das hipóteses apenas 10% dessa informação criada será correctamente organizada e posteriormente utilizada como um activo para as organizações, sendo que este número tenderá a descer à medida que a sua produção aumente. Nos dias de hoje, vivemos na era da informação, mas mais do que isso, vivemos na era dos dados, sendo neste momento possível afirmar que os dados são o recurso mais importante ao dispor

de uma empresa que pretenda ser inovadora numa era de tão grande competitividade, em que as margens de manobra entre concorrentes e recursos são tão escassas. Em suma, o grande objectivo de qualquer organização passa por tornar esta contínua quantidade crescente de dados num "*driver*" para o aumento da sua própria produtividade e inovação.

Fayyad, Piatetsky-Shapiro e Smyth (1996a) afirmam que a humanidade está obrigada a criar sistemas informáticos analíticos parcialmente autónomos, que sirvam de auxílio à descoberta de padrões, estruturas ou relações entre estes massivos volumes de dados. Na mesma linha de pensamento, Han, Pei e Kamber (2012) admitem ser imperativo criar técnicas e modelos tecnológicos que, de forma inteligente e automática, consigam processar e transformar estes dados em informação útil e, posteriormente, em conhecimento. Este enquadramento que relaciona a abundancia de dados multidimensionais, provenientes de múltiplas áreas e a necessidade de criar uma nova geração de técnicas, métodos e de ferramentas que permitam a análise desses mesmos dados, com o objectivo de encontrar padrões robustos capazes de gerar novo conhecimento preditivo de forma a auxiliar a tomada de decisão, está na base da criação do processo de *Knowledge Discovery on Databases (KDD)*. Este processo, em que o Data Mining constitui a mais importante etapa, tem ganhado cada vez mais aceitação e importância tanto na comunidade académica como a nível empresarial e científico, tendo como ideia base descobrir conhecimento preditivo de alto nível, ou seja, potenciar diferentes e mais avançadas interpretações de dados ao dispor, levando a análise para patamares analíticos superiores às da comum análise que a mente humana por si só é capaz de executar.

A relação entre o *KDD* e o Data Mining será igualmente abordada neste trabalho. Contudo podemos distinguir *KDD* como o processo geral de descoberta de conhecimento a partir dos dados Fayyad et al. (1996a), ou, segundo Han et al. (2012), o processo de extracção automático, ou não, de padrões indiciadores de futuro conhecimento que se encontram armazenados em *data warehouses* (repositórios centrais de dados de uma empresa), na própria internet, ou em repositórios massivos de informação. Por outro lado, o Data Mining é visto como a mais importante, distintiva e complexa etapa desse processo, sendo também considerado como uma área multidisciplinar segundo Han et al. (2012), devido ao facto de esta absorver conhecimento de diferentes áreas do saber, nomeadamente, da estatística, das tecnologias de base de dados, da inteligência artificial, da visualização de dados ou do reconhecimento de padrões. Exactamente por isso, grande parte da literatura, seja pela importância que tem no processo ou simplesmente por uma questão de simplificação e de interpretação, generaliza todo o processo e apelida-o de Data Mining.

1.1. Contextualização e Relevância do tema

A capacidade de assumir risco é desde há muito tempo considerado como um dos maiores *drivers* que permitem a qualquer instituição financeira maximizar a sua criação de valor e potenciar a sua própria rentabilidade. Desta forma, a gestão dos principais factores de risco a que cada organização se encontra exposta assume uma importância vital, garantindo o cumprimento de objectivos traçados, assim como a sua própria competitividade e a sua sobrevivência. Até ao fim da década de 90, quer os reguladores quer os próprios mercados, por razões associadas à difícil identificação, medição e até mesmo devido à pouca consciência que tinham do seu impacto nocivo, foram negligenciando a análise e a gestão do risco operacional. Moosa, Li e Naughton (2015) afirmam que, dado o secretismo com

que as organizações tratam as suas perdas operacionais torna-se difícil adquirir dados com qualidade e, em segundo lugar, é difícil modelar o risco operacional dada a sua natureza heterogénea, variando desde fraudes, processos judiciais, ou um simples incêndio. No entanto, o risco operacional, apesar de muitas vezes negligenciado, sempre foi um risco ao qual os bancos estiveram expostos. Exposição essa que aumentou significativamente à medida que as organizações foram crescendo graças ao fenómeno da globalização. Este fenómeno foi o responsável pela desregulação dos mercados financeiros e pelo aumento da complexidade dos seus produtos e respectivas operações financeiras. Durante a última década, estas instituições têm modificado os seus produtos e processos internos a um elevado ritmo, numa tentativa de acompanhar os avanços tecnológicos, mudanças de comportamento de clientes e a actividade dos seus concorrentes directos, de forma a aumentar a sua exposição ao risco operacional (Di Renzo et al., 2007).

A verdade é que nos dias de hoje o risco operacional é considerado uma área de estudo e de relevo por si só. Diferentes autores apresentam as suas razões para a crescente importância concedida ao risco operacional. Por exemplo, Geiger (2000) considera como factores: (i) um melhor entendimento do impacto dos riscos operacionais; (ii) a constatação da insuficiência de recorrer a abordagens meramente quantitativas de risco de crédito e de mercado como forma de captar alguns tipos de risco, assim como, um reconhecimento de que a gestão de risco operacional deveria ser considerada uma área de análise e de estudo por si só; (iii) a inclusão dos riscos operacionais na problemática da gestão global de risco; (iv) um maior interesse, quer ao nível da gestão, quer ao nível dos requisitos de capital por parte das entidades reguladoras em relação ao risco operacional. Por outro lado, Cummins e Embrechts (2006), admitem diferentes razões como: (i) a globalização dos mercados financeiros internacionais, que aumentou a complexidade dos serviços prestados pelas respetivas firmas aumentando desta forma a exposição a ocorrências de risco operacional; (ii) uma maior consolidação das indústrias financeiras, tornando-as maiores e mais complexas, mas aumentando também o risco de falhas ou incompatibilidades entre sistemas após as suas integrações/migrações serem concluídas; (iii) o aumento da exposição dos bancos a possíveis perdas e o consequente contágio para os seus parceiros devido à partilha de acções e a investimentos/empréstimos conjuntos, (iv) o aumento da dependência nos sistemas informáticos quer a nível de alocação de recursos quer ao nível dos processos internos e externos das organizações, deixando-as mais vulneráveis a possíveis falhas. Por último Buchelt and Unteregger (2004) argumentam que apesar de ocorrências externas ou riscos de possíveis fraudes terem sido uma constante ao longo da história, foram sem dúvida, os progressos tecnológicos os catalisadores que levaram ao aumento do risco operacional. Estes avanços proporcionaram uma rápida inovação financeira, assim como, a multiplicação de produtos financeiros fortemente dependentes de sistemas expostos em demasia a riscos operacionais.

Sob outra perspectiva, alguns académicos admitem que o interesse demonstrado por este tema foi aumentando como consequência directa de vários escândalos financeiros públicos, e tremendamente dispendiosos, que ocorreram no fim da década de 90 e durante a primeira década do século XXI. Na opinião de Power (2005) o verdadeiro "inventor" do risco operacional foi Nicholas Leeson que, sozinho e de forma fraudulenta, conseguiu levar o maior banco de investimento Inglês, Barings, à falência em 1995. Este e outros escândalos do mesmo género trouxeram para a ribalta o problema da ineficaz regulamentação do risco operacional, assim como a incapacidade de quantificação os seus "estragos". Esta constatação apoia os argumentos de De Fontnouvelle (2007) que considera que a frequência com que estas grandes perdas operacionais ocorrem, o seu impacto,

o seu poder de alcance e as consequências reputacionais desastrosas que delas advêm, são razões mais do que suficientes para obrigar as instituições financeiras a adoptarem melhores instrumentos de medição, monotorização e mitigação do risco operacional assim como a precaverem-se com maiores reservas de capitais de forma a colmatar estas perdas. Na mesma linha de pensamento, Buchelt e Unteregger (2004), Jednak e Jednak (2013) e Cummins, Lewis e Wei (2006) tentam ir mais longe afirmando que, ao nível organizacional, no seu todo, deve ser criada uma consciência de risco acompanhada de uma implementação de medidas de gestão diárias, optando por uma postura proactiva e não reactiva em relação ao risco. É então possível concluir que o risco operacional facilmente se associa a eventos que se "desviam" dos habituais processos de uma empresa, mas também do normal comportamento dos seus funcionários. Tais particularidades vêm dificultar a quantificação da exposição ao risco operacional por parte das empresas, assim como a sua previsão ou modelação. Na tabela 1 podemos verificar alguns dos mais conhecidos casos de estudo e a dimensão das suas perdas apresentados por Jednak (2013). Importante referir que, para além dos custos financeiros (directos) associados, a reputação destas instituições foi bastante afectada (custos indirectos).

Instituição	Ano	Perdas
Sumitomo	1996	49 Milhões de euros
Barings	1995	691 Milhões de dólares
SocGEn	2008	2.6 Milhões de dólares
AlliedIrishBank	2002	1.3 Milhões de dólares

Tabela 1 - Casos de estudo referidos em (Jednak & Jednak, 2013)

A indústria da banca tem vindo a sofrer mudanças estruturais tremendas na última década, sendo que, como consequência directa, a quantidade de dados produzidos e armazenados por estas instituições tem aumentado a grande velocidade. Bhambri (2011), assume esta perspectiva e afirma que desde a década de 90 a indústria da banca tem sofrido profundas alterações na forma como executa as suas operações devido ao desenvolvimento das novas tecnologias de informação, dando alguns exemplos: a computorização das operações financeiras apoiada em *softwares* automáticos de suporte, a centralização de informação em grandes repositórios de dados, o aparecimento das transacções *online* por todo o mundo, assim como, uma nova metodologia de trabalho orientada para cada cliente, enquanto entidade individual. Por conseguinte, Moin e Ahmed (2012) afirmam que as técnicas de análise de dados baseadas em técnicas estatísticas têm sentido dificuldade em gerir todo este volume de dados e, como tal, são necessárias novas ferramentas/técnicas para analisar dados e para que deles se possa extrair a chamada "informação escondida". Fayyad et al. (1996a), procura quantificar o tamanho das bases de dados e a forma como estas se têm vindo a desenvolver, admitindo que o seu crescimento acontece segundo dois vectores: i) por um lado crescem em número de *N* objectos, ou registos, admitindo que é comum encontrar bases de dados com N=10^9 objectos ou mais; ii) por outro lado crescem em número de *D* atributos associados a cada objecto, o que pode variar entre 10^2 ou 10^3 por objecto. Tendo em conta esta realidade, o autor defende que é impossível fazer algo construtivo a partir de milhões de registos, cada um com centenas de atributos, admitindo que este trabalho não pode estar somente a cargo dos humanos. Deverá ser, no mínimo, parcialmente informatizado. Bhambri (2011) corrobora esta tese e afirma que produzir e armazenar um grande volume de dados é extremamente útil, mas que de pouco servirá se não se conseguir capitalizar perfeitamente os benefícios que daí possam advir. Moin e Ahmed (2012) tornam-se

perentórios ao afirmarem que os dados são um dos recursos mais valiosos para o sector bancário, mas apenas se este souber como deles extrair o previamente mencionado "conhecimento escondido", acrescentando que, a partir do Data Mining é possível extrair conhecimento de dados históricos/armazenados e, assim, prever situações futuras, permitindo também otimizar as decisões de negócio, aumentar o valor de cada cliente, melhorar a comunicação e, por consequência a satisfação dos mesmos.

Chitra e Subashini (2013) e Kuonen (2004) admitem a importância estratégica do Data Mining para a indústria bancária nos dias de hoje. Segundo os autores, é reconhecida a forma como as grandes organizações tentam imitar o tipo de comportamento e a orientação de negócio que as pequenas empresas sempre tiveram, isto é, criar relações de um para um com os seus clientes. Qualquer indústria, ou companhia com ambição, reconhece que uma abordagem individual para com os seus clientes é da maior importância dando alguns exemplos: i) conhecer individualmente cada cliente, saber quais as suas preferências, o que procuram ao certo ou as razões pelas quais fazem negócio consigo e não com a concorrência. No fundo, extrair conhecimento de interacções passadas para que no futuro os possa servir de forma mais personalizada e eficaz; ii) por outro lado, estas organizações procuram dotar-se de um maior grau de conhecimento sobre o *lifetime value* de cada cliente para assim saber a quais devem dedicar mais tempo e recursos ou quais os que simplesmente podem deixar cair; iii) por último, desenvolver processos de *CRM (Customer Relationship Management)*, uma estratégia que vai ao encontro do que já foi mencionado. De acordo com os autores, estes processos têm como base mudar o foco das empresas, ou seja, em vez de estas se focarem apenas e só na aquisição de clientes o foco passará a incidir também sobre a retenção dos seus clientes garantindo que tempo, dinheiro e recursos são canalizados em quantidade suficiente para cada um individualmente. Esta estratégia permite não só estabelecer relações duradouras com os clientes como também potenciar os proveitos vindos de cada um. Para realizar este tipo de acções, os bancos recolhem dados através do chamado *OLTP (online transaction processing)*, sistemas que se encontram implementados por todo o mundo registando toda e qualquer actividade que um comum cidadão faça no seu dia-a-dia. Esta informação só poderá ser considerada como uma oportunidade de aprender algo sobre um determinado cliente se após a sua produção esta for armazenada e organizada de forma metódica num *data warehouses,* caso contrário não passará de um conjunto de dados armazenados para um qualquer propósito operacional. Segundo Kuonen (2004), um *data warehouse* confere memória a uma organização, memória sobre os comportamentos dos seus clientes. No entanto a memória tem pouca utilidade sem inteligência. Neste caso, a inteligência é conferida pelo Data Mining e pela sua capacidade de analisar esta quantidade de dados, tentando encontrar padrões, aplicar regras ou fazer previsões sobre o futuro. De certa forma é possível dizer que o Data Mining e as suas ferramentas, quando utilizadas com critério e conhecimento de causa, conferem inteligência a um *data warehouse.*

Pivk, Vasilecas, Kalibatiene e Rupnik (2013) consideram que uma abordagem assente em técnicas de Data Mining por parte dos bancos será extremamente útil para obter conhecimento sobre comportamentos e hábitos dos clientes, tal como as suas necessidades ou desejos. Os autores afirmam ainda que esta abordagem será útil para reduzir os riscos, aquando da obtenção de créditos com o objectivo de minimizar o número de *defaults,* ou para identificar possíveis acções fraudulentas. Por exemplo, quando se pretende entender os padrões que no passado levaram a situações de fraude com cartões de crédito, cheques ou situações de reivindicação abusiva/fictícia de direitos para com seguradoras. Fazendo a ligação com o último ponto, praticamente toda a literatura refere que o Data

Mining é uma potente ferramenta na luta contra a fraude bancária. Por sua vez, Hormozi e Giles (2004) admitem que os bancos podem utilizar bases de dados de terceiros e analisá-las com técnicas de Data Mining numa tentativa de encontrar possíveis padrões de fraude. Posteriormente, esta informação é cruzada com as suas próprias bases de dados para tentar encontrar situações semelhantes, ou então proceder a análises directamente nos seus próprios reservatórios de dados. Em suma, e dando uso às palavras de Koyuncugil e de Ozulbas (2008a), de acordo com o que foi exposto ao longo deste capítulo, a descoberta de conhecimento e a identificação de factores de risco provêm da capacidade de cada um para identificar e analisar as relações entre variáveis financeiras e variáveis operacionais.

1.2. Problema de Estudo e Objectivos da Tese

O objectivo principal deste trabalho consiste em **identificar factores que podem ter impacto na *frequência* e na *severidade* do Risco Operacional dentro de uma Instituição Financeira.**

Para atingir este objectivo, irão ser aplicadas técnicas de Data Mining a uma Base de Dados de Risco Operacional proveniente de uma instituição financeira.

No sentido de alcançar o objectivo principal foram definidos três objectivos específicos:

(i) Análise exploratória da base de dados de Risco Operacional.
(ii) Organização da informação de forma a construir um conjunto de análises que permitam um maior conhecimento sobre os eventos de Risco Operacional.
(iii) Desenvolvimento de segmentos capazes de identificar factores que possam vir a ser indicadores da evolução das variáveis *frequência* e *severidade*

No âmbito da dissertação de mestrado que se pretende desenvolver, torna-se necessário **clarificar o problema de estudo e o consequente acréscimo de conhecimento que este trabalho irá trazer para o mundo científico/académico**. Após exaustiva compreensão da revisão literária, optou-se pela elaboração de um trabalho a nível nacional. De facto, as instituições financeiras Portuguesas têm desenvolvido muito cálculo de risco operacional, existindo inclusivamente estudos sobre a quantificação do risco que os bancos incorrem nas suas operações. No entanto, pouco se tem feito com o objetivo de analisar as causas para essa exposição ao risco operacional, ou seja, para tentar encontrar variáveis/factores que possam explicar e, por sua vez, tentar controlar esse mesmo risco. Salientam-se, nomeadamente, as variáveis com poder explicativo sobre questões como a frequência ou a severidade dos riscos operacionais.

1.3. Estrutura do Trabalho

A presente dissertação de mestrado irá ter a seguinte estrutura:

Capítulo 2: Revisão da Literatura. Este capítulo vem espelhar toda a pesquisa bibliográfica efectuada em livros, artigos, ou outras teses de mestrado, para a realização do presente trabalho. Tem como objectivo conferir um melhor entendimento sobre os temas que irão ser abordados, nomeadamente a sua origem e desenvolvimento. Os referidos temas são o Risco Operacional, os modelos *KDD,* o Data Mining e a análise de *clusters*.

Capítulo 3: Metodologia de Investigação. Este capítulo ilustra a estratégia de investigação utilizada. Os aspectos mais relevantes que caracterizam a base de dados utilizada são mencionados,

assim como o método utilizado para a sua recolha. Posteriormente é feito um levantamento detalhado de todos os processos de tratamento de dados elaborados ao longo da tese.

Capítulo 4: Análise de Resultados. Este capítulo irá apresentar todas as análises realizadas aos dados de risco operacional. Começando pela análise exploratória da fonte de dados original, passando por uma análise mais detalhada da fonte de dados após esta ter sido trabalhada, terminando com a análise de *Clusters* realizada.

Capítulo 5: Conclusões. Este capítulo irá apresentar as conclusões obtidas após a realização desta dissertação.

Capítulo 6: Limitações e Investigação Futura. Este capítulo irá apresentar todas as limitações enfrentadas durante a realização deste trabalho. Por outro lado apresenta alguns pontos que podem ser explorados em futuras investigações, de forma a aprofundar o conhecimento sobre o fenómeno de estudo.

Capítulo 7: Bibliografia. Este capítulo irá apresentar todas as referências bibliográficas utlizadas na realização desta dissertação.

2. REVISÃO DA LITERATURA

2.1. Risco Operacional

O conceito de risco operacional tem gerado bastantes divergências entre académicos e profissionais. Se, por um lado, todos o consideram como heterogéneo, de difícil quantificação e previsão, a sua definição é ainda pouco consensual, assim como a sua classificação e abrangência. A ideia da heterogeneidade está presente em Moosa (2007) e em Gonçalves (2011) ao afirmarem que o risco operacional advém da materialização de uma variedade de eventos, incluindo fraude, roubo, usurpação de propriedade intelectual, perda de membros chave de uma equipa, processos judiciais, perda de informação, terrorismo, vandalismo e até mesmo de desastres naturais.

Crouchy (2001) considera o risco operacional como um conceito difuso, justificando que, para uma organização, é difícil distinguir de forma clara este tipo de risco de todas as habituais incertezas com que estas organizações são confrontadas no decorrer das suas actividades diárias. Por outro lado, Rao e Dev (2006) afirmam que há um par de anos não era de todo incomum considerar o risco operacional como algo residual. Na mesma linha de raciocínio, Power (2005) defende que este tipo de risco iniciou a sua vida na indústria financeira como uma categoria secundária, ou seja, todos os acontecimentos cujas metodologias de gestão de risco de mercado, ou de crédito, não conseguissem albergar, passariam a ser considerados como riscos operacionais. Lopez (2002) acompanha esta ideia e considera que a definição de risco de mercado assim como a de risco de crédito, são claras. Contudo, a definição de risco operacional tem vindo a evoluir com o passar dos anos, começando por ser definido como todo o tipo de risco incapaz de ser quantificado por parte dos bancos. Seguindo este raciocínio, Buchelt e Unteregger (2004) consideram que a diferença entre o risco operacional e os outros riscos, como o de mercado ou o de crédito, reside no facto de que, enquanto os últimos são riscos apenas associados à actividade bancária, o primeiro, assim como os perigos que dele decorrem, são problemáticos para qualquer actividade de negócio, incluindo o sector financeiro. Estes autores afirmam ainda que uma eventual perda operacional não se determina pelas suas consequências, mas sim pelas suas causas. Além disso, devido à sua natureza tão vasta e difícil de "agrupar", o risco operacional pode-se materializar directamente ou indirectamente através dos riscos de mercado e de crédito. Em suma, consideram estes autores que a única forma de definir risco operacional e de o diferenciar dos outros tipos de risco passa por encontrar as suas origens em cada contexto, e não apenas considerar risco operacional como todo o risco que não seja de mercado nem de crédito.

Jobst (2007) considera que qualquer risco que advenha de actividades de negócio, assumidas ou por negligência, processos ou sistemas inadequados, comportamento incorrecto de trabalhadores ou eventos externos, devem ser considerados como riscos operacionais. Considera também, assim como Power (2005), Cummins et al., (2006) e Dima e Orzea (2012), que ao longo dos anos os bancos sempre tiveram consciência dos problemas/incertezas provenientes do uso das tecnologias de informação (*IT*), assim como dos comportamentos humanos danosos/duvidosos (fraudes), das acções que a médio prazo poderiam vir a danificar o modo como estes executam as suas operações de negócio e até mesmo dos problemas legais associados. Contudo, foram necessários os escândalos financeiros ocorridos durante a década de 90 (Barings, Daiwa, Allied Irish Bank) e as suas consequentes perdas milionárias, quer a nível de valor de mercado quer a nível reputacional, para despertar/alarmar os média e as pessoas encarregues da regulação e das operações bancárias para a severidade/impacto destes riscos. O resultado foi uma aceitação geral por parte da comunidade financeira e de

regulamentação bancária, considerando o risco operacional como uma disciplina própria, por si só, complexa e que obriga a uma gestão activa de forma a garantir a sobrevivência de qualquer actividade de negócio, particularmente no sector da banca. Assim surge o risco operacional como "rótulo" para todos estes riscos difíceis de agrupar no antigo enquadramento teórico, dominado essencialmente pelo risco de mercado e o risco de crédito. Tornou-se uma categoria por si só, e neste momento, podemos afirmar que existe uma maior consciencialização da gestão do risco operacional, que se estende por toda a hierarquia de uma instituição financeira.

No início da década de noventa surgem as primeiras as primeiras definições para o conceito de risco operacional. No ano de 1993, o Grupo dos 30[1] considerou o risco operacional como a incerteza relacionada com perdas resultantes de sistemas ou controlos inadequados, erros humanos ou de gestão. O Banco da Austrália, em 1999, avançou com uma explicação mais abrangente de risco operacional, incluindo neste conceito todos os riscos que não sejam risco de mercado ou de crédito, que possam causar volatilidade nos proveitos, nas despesas e no valor de negócio dos bancos. No ano anterior o Banco da Reserva Federal de Nova Iorque, mais concretamente Shepheard-Walwyn e Litterman (1998) caracterizaram o risco operacional como um termo geral, passível de ser aplicado a todas as falhas que possam influenciar a volatilidade da estrutura de custos de uma empresa. De notar a diferença entre as últimas duas definições: se por um lado o Banco da Austrália admite que o risco operacional irá incidir sobre os dois lados do negócio, custos e lucros, a definição de Shepheard-Walwyn e Litterman incide apenas sobre os custos, exemplos claros da falta de consenso por parte dos académicos e profissionais na definição do tema. Por fim, nesse mesmo ano, Crouchy, Galai e Mark (1998) propuseram uma definição que identificava fontes externas e internas de risco operacional: o risco de eventos externos, deficiências em controlos internos, ou em sistemas de informação, que resultem em perdas, sejam elas antecipadas ou inesperadas.

Inicialmente o Comité de Supervisão Bancária de Basileia considerou o risco operacional como as perdas directas ou indirectas resultantes de processos internos inadequados, falhas, pessoas, sistemas, ou eventos externos. Contudo, a referência a perdas indirectas foi posteriormente eliminada devido ao pressuposto da quantificação de capital regulamentar, já que estas perdas apresentam obstáculos à sua medição. Assim, o Acordo de Basileia II definiu o risco operacional como sendo o risco de perdas resultantes de processos internos inadequados, falhas, pessoas, sistemas ou de eventos externos. Esta definição, que se baseia nas causas subjacentes (fontes) de risco operacional, inclui o risco legal, mas exclui os riscos de negócio e reputação. Contudo, esta definição não foi completamente aceite desde o início pelos académicos e profissionais da área. Três exemplos claros foram os de Herring (2002), quando criticou a definição pela sua total exclusão do risco de negócio; Turing (2003), quando curiosamente a criticou, não pela falta de conteúdo, mas por ser demasiado ampla e, como tal, pouco útil; e, por último, Thirlwell (2002), ao argumentar que esta definição apresenta uma visão mensurável de risco operacional para quem procura algo que seja quantificável, contudo considera-a inadequada para identificar as causas que levam os bancos a falharem. Apesar da falta de concordância, é possível afirmar que a definição apresentada no Acordo de Basileia II é a que apresenta uma utilização mais difundida na maioria dos bancos, seguradoras e entidades supervisoras.

[1] Grupo internacional, sem fins lucrativos, de carácter consultivo sobre assuntos económicos e financeiros, composto por elementos dos sectores públicos, privados e académicos

Utilizando esta definição, Jobst (2007) admite que o riso operacional pode ser dividido em duas categorias: riscos internos e riscos externos. Dentro dos riscos internos o autor identifica três causas para a exposição ao risco: (i) pessoas. O autor menciona as potenciais falhas que os seres humanos têm ao levarem a cabo qualquer tarefa, e no contexto empresarial apresenta como exemplos um treino incompleto, mecanismos insuficientes ao seu dispor, fraca supervisão ou a escassez de recursos; (ii) processos. O autor exemplifica com as falhas operacionais, processos mal desenhados, incapacidade em seguir as linhas orientadoras da empresa ou até mesmo a complexidade dos produtos e dos contractos; (iii) sistemas. Desta categoria fazem parte todos os erros relacionadas com os sistemas informáticos (*IT*) desde os dados às questões de segurança, de funcionalidade, de compatibilidade e de capacidade computacional. A categoria dos riscos externos remete para factores fora do controlo da firma, considerando que os competidores externos, as mudanças de paradigma da empresa, mudanças políticas, como diferentes partidos no poder ou aumento de impostos, desastres naturais, terrorismo e vandalismo, irão aumentar a exposição ao risco. A tabela 2, apresentada por Jednak e Jednak (2013), ilustra a categorização mencionada para um melhor entendimento.

Riscos Internos		
Pessoas	**Processos**	**Sistemas**
Fraude de trabalhadores	Erros de contabilidade	Qualidade dos dados
Erros de trabalhadores	Erros na quantificação de risco	Problemas de *software*
Delito	Riscos contractuais	Falhas de segurança
Lei do emprego	Incapacidade de cumprir prazos	Riscos estratégicos
Segurança e Saúde	Complexidade dos produtos/serviços	Incompatibilidades de sistema
Falta de conhecimento/*skills*	*Reporting* incompetente	Falhas de sistema
Perda de pessoas importantes	Erros transacionais	Performance insuficiente
Riscos Externos		
Questões legais	Incêndios	Catástrofes naturais
Lavagem de dinheiro	Riscos políticos	Segurança física
Regulamentação	Incumprimento dos fornecedores	Terrorismo
Impostos	Roubo	Usurpação de propriedade intelectual

Tabela 2 – Categorização do Risco Operacional segundo (Jednak & Jednak, 2013

2.1.1. Características distintas do Risco Operacional

No capítulo anterior foi feito um levantamento das diferentes definições de risco operacional que surgiram durante as ultimas décadas. Foi possível verificar como as diferentes definições/posições em relação ao tema se foram transformando e adaptando à medida que a consciencialização sobre este tipo de risco foi também crescendo no seio das organizações e das próprias entidades reguladoras. Neste capítulo serão apresentadas algumas características distintivas do risco operacional. Após variadas leituras foi possível identificar características nas quais os diferentes autores convergem e outras onde académicos e técnicos da área apresentam posições divergentes.

Existem quatro características que levam os autores a uma maior concordância: i) a diversidade do risco operacional parece ser a que maior consenso gera, não deixando de ser um problema para os analistas, visto que a esta diversidade está associado também um aumento do número de dimensões de análise deste fenómeno, tornando difícil limitá-la a um número restrito de dimensões (I. A. Moosa, 2007). Diferentes autores corroboram esta ideia. Buchelt e Unteregger (2004) admitem que o risco operacional nasce da interligação entre diferentes tipos de risco com as mais variadas origens; Milligan (2004) afirma que o risco operacional pode abranger desde uma queda, a uma bomba ou ao colapso de um banco; Hoffman (1998) considera que o risco operacional se encontra

presente em todos os processos de uma actividade de negócio, independentemente do nível hierárquico em que ocorram ou do seu maior ou menor poder de decisão. Este autor considera também que este tipo de risco, por se afastar dos moldes tradicionais relativamente aos quais as organizações se consideravam precavidas, ganha toda uma nova dimensão; ii) o problema da falta de indicadores financeiros e de dados robustos quando trabalhamos com risco operacional. De Koker (2006) argumenta que, enquanto a lógica de assumir riscos para conseguir posteriormente retirar maiores proveitos, que caracteriza as entidades bancárias, é passível de ser aplicado ao risco de mercado e de crédito, esta lógica torna-se muito mais difícil de ser aplicada ao risco operacional pela falta de indicadores financeiros. Por outras palavras, este tipo de risco não se encontra relacionado de forma próxima com nenhum indicador. Num exemplo simples, ninguém é capaz de acompanhar a evolução de algo tão aleatório como uma quebra num sistema informático. Por outro lado, Rempfler (2011) é da opinião de que, sem dados robustos (muitas vezes as instituições financeiras guardam este tipo de informação com demasiado secretismo), e tendo como segundo factor a dificuldade existente em controlar e modelar o comportamento humano (fraudes, incompetências, incapacidades, ignorância), torna-se muito difícil desenvolver uma abordagem quantitativa em relação ao risco operacional; iii) o facto de o risco operacional ser considerado também ele como uma questão cultural. De facto, devido à sua diversidade e pelo facto de se encontrar de tal forma difundido por toda a estrutura hierárquica de uma organização, a prevenção deste risco não pode apenas ficar a cargo dos gestores de topo, como acontece com os riscos de mercado ou de crédito. Buchelt e Unteregger (2004), Rao e Dev (2006) descrevem a gestão de risco operacional como uma actividade mais corporativa do que de gestão propriamente dita. Na medida em que, como todos os empregados e elementos da organização, podem ser considerados como elementos causadores de risco operacional, muito do seu impacto individual, ou falta dele, será influenciado por questões pessoais como as suas experiências, personalidade e cultura. A cultura de risco global de uma organização define-se então pela percepção que todas estas pequenas partes têm em relação ao risco e pela forma como tentam geri-lo e minimizá-lo em cada acção do seu dia-a-dia; iv) por último Moosa (2007) e Kaiser e Kohne (2006) afirmam que o risco operacional é considerado como um tipo de risco muito mais endógeno (passível de ser gerado pela própria actividade) do que os riscos de mercado e de crédito. Olhando por exemplo, para a definição de risco operacional do Comité de Supervisão Bancária de Basileia, as causas deste tipo de risco são de cariz muito mais interno do que externo e como tal as oportunidades para mitigar esse mesmo risco aumentam, dando maior controlo à organização.

As características acima mencionadas parecem consensuais, mas as que se irão seguir são alvo de debate e de algum desacordo por parte de diversos autores. Tipicamente, considera-se o risco operacional como unilateral (consideramos somente a parte das perdas da sua curva de distribuição probabilística) na medida em que surge como um "subproduto" indesejado de todas as actividades de negócio de uma organização, o que implica que, por exemplo, ao contrário do risco de mercado, a relação entre assumir riscos/recolher proveitos não tem equivalência no caso do risco operacional. Herring (2002) considera este risco como unilateral associando-o apenas à dimensão das perdas independentemente destas de facto ocorrem ou não, argumentando que é muito difícil imaginar um cenário onde o risco operacional possa originar lucros inesperados. Seguindo o mesmo raciocínio, Rempfler (2011) argumenta que as entidades bancárias, ao assumirem para si mesmas maior risco operacional, não o fazem com o objectivo de retirar daí maiores proveitos, afirmando que este tipo de risco na maioria dos casos é causador de grandes perdas que se alastram para as diferentes partes interessadas como accionistas, gestores de topo ou credores. É então possível concluir, segundo estes

autores, que ao assumir risco operacional um banco não consegue daí gerar lucros, ficando este tipo de risco apenas associado à dimensão dos custos. A dimensão dos lucros advém, por seu lado, da relação entre assumir riscos/recolher dividendos, somente para riscos de mercado ou de crédito. Admitem então estes autores que um banco não procura activamente expor-se ao risco operacional com a intenção subjacente do aumento dos seus lucros. Contudo, Moosa (2007) entra em desacordo com esta linha de pensamento, ou seja, não considera o risco operacional como unilateral, admitindo que a capacidade de gerar lucro é um argumento que não se deve descorar. Segundo este autor as organizações não se expõe a este tipo de risco, nas suas actividades diárias, por uma questão de gosto pessoal, mas sim pela possibilidade de monetizar tais actividades. O autor não exclui que este tipo de risco encontra-se presente nas diferentes vertentes da actividade bancária, desde transacções, investimentos ou gestão de activos, mas afirma que os bancos tendo noção de tudo isso, veem todas estas actividades como potenciais fontes de lucro, principalmente quando existe a possibilidade de tais riscos nem virem a ocorrer. Em suma, o autor considera que o assumir de alguns riscos com o objectivo de gerar proveitos futuros não só é a base que sustenta a actividade bancária como é uma relação que nunca deverá ser considerada como unilateral. Devemos olhar sempre para o lado positivo da curva de distribuição que representa os proveitos a colher, caso poucos ou nenhuns desses riscos se venham a materializar. De facto, a ideia de considerar o risco operacional como unilateral é mais fácil de aceitar, na medida em que parece difícil associar a riscos com tão grande potencial destrutivo, caso se venham a materializar, a capacidade de gerar lucros. Ainda assim, esta visão aparenta ser um pouco limitada/pessimista e parece ignorar a relação risco/lucro que baliza a actividade bancária.

A segunda questão que se levanta é a seguinte: deve ser o risco operacional considerado como idiossincrático, na medida em que, quando uma empresa é atingida por tal risco este não se dissemina por outras empresas. Lewis e Lantsman (2005) defendem esta ideia, afirmando que o risco operacional é de facto idiossincrático, uma vez que o risco de tais perdas e a sua manifestação não se encontram correlacionados com as forças gerais do mercado. Defende também o autor que esta característica só deve ser aplicada a tipos de risco com verdadeira capacidade de contágio, como são os casos do risco de mercado ou de crédito. A sua lógica aparenta ser sólida, ou seja, se uma economia se encontrar em recessão, por exemplo, os mercados financeiros irão oscilar num sentido descendente, aumentando por sua vez o número de incumprimento de créditos, o que levará a uma redução dos lucros e da própria liquidez, criando uma espiral negativa de qual todas as outras entidades financeiras irão também sofrer. Apesar de consistente, a sua proposição é contrariada por outros autores. Por exemplo, Moosa (2007) considera que qualquer um dos três tipos de risco já mencionados são passíveis de serem considerados como sistémicos dada a sua relação com o estado da economia. Por outro lado, faz uso do cenário económico acima referido para nele incluir o risco operacional, argumentando que, em momentos de recessão da economia, assistimos a um aumento de despedimentos ou de situações de possível fraude que mais facilmente surgem à superfície o que faz disparar o risco associado a custo legais. Segundo o autor, é nestas alturas que muitos cidadãos ficam sem capacidade para pagar os seus créditos directamente feitos aos bancos ou associados a cartões de crédito que foram utilizando de forma descontrolada em momentos de maior desafogo financeiro. Por último argumenta que a obrigatoriedade que o Comité de Basileia impõe sobre os bancos para que tenham na sua posse uma reserva de capital de forma a combater inesperados e potencialmente perigosos riscos operacionais é a maior prova de que este tipo de risco é de facto sistémico. Esta visão de idiossincrasia do risco operacional aparenta ser pouco sustentável. Assim, segundo Gonçalves (2011), sempre que um banco incorrer em perdas por força de flutuações no mercado (risco de

mercado) ou por incumprimento de créditos (risco de crédito), a sua capacidade de cumprir com as suas obrigações para com outros bancos será afectada. Até aqui tudo parece correcto. Mas se o problema tiver origem em acções fraudulentas e irresponsáveis (risco operacional), as consequências, a nível da incapacidade de cumprir obrigações para com terceiros não serão as mesmas? De recordar o exemplo da falência do banco inglês Barings em 1995 e das consequências sistémicas que daí resultaram. Parece de facto pouco crível não considerar o risco operacional como factor nocivo de contágio para outras entidades dentro do mesmo ramo.

Por último, surge a dúvida sobre se o risco operacional é capaz de se distinguir do risco de mercado ou de crédito, na medida em que, por vezes é difícil avaliar uma perda e saber a qual destas categorias se deve associar. Por exemplo, a última crise financeira nos Estado Unidos da América veio destacar a forte relação existente entre estes três tipos de risco. Por oposição surge esta ideia: um *trader* (pessoa ou entidade que compra e vende instrumentos financeiros como acções, títulos ou derivados, aproveitando-se das volatilidades dos mercados para gerar lucros) tem uma determinada acção, pois considera que a situação do mercado o irá favorecer, mas dada à volatilidade do mesmo, em vez de gerar lucro tem um grande prejuízo. A quem devemos imputar as culpas? Se por um lado o *trader* agiu de acordo com as orientações da sua organização, podemos estar perante um simples evento de risco de mercado. Se, por outro lado, o *trader* agiu de forma voluntária e contra essas orientações, estamos perante um evento de risco operacional. O mesmo critério deve ser aplicado em situações de risco de crédito. Segundo Kaiser e Kohne (2006) muitos eventos de risco de crédito só ocorrem porque as pessoas e os sistemas que deviam definir com critério a atribuição desses mesmo créditos falham. Apesar de cada situação de perda poder resultar de um, dois ou até vários diferentes tipos de riscos e considerando que a sua análise é feita de acordo com determinada perspectiva, as conclusões, quanto ao responsável, podem de facto variar. Sobre esta questão, diferentes autores têm a mesma opinião: Buchelt e Unteregger (2004) admitem que a única forma de definir risco operacional e de o diferenciar dos outros tipos de risco passa por encontrar as suas origens em cada contexto. Kaiser e Kohne (2006) têm a mesma opinião e argumentam que uma situação de perda deve ser atribuída ao tipo de risco que esteja na base da sua causa, exemplificando com o caso do banco *Barings,* onde, de facto, foi uma oscilação de mercado que gerou a perda, mas onde a verdadeira causa esteve directamente relacionada com risco operacional na forma de fraude e de supervisão inadequada. Em suma, certos eventos podem sempre desencadear uma corrente de acontecimentos nocivos, porém muitas das vezes esses eventos são, no fundo, secundários. O que queremos encontrar é o verdadeiro "culpado", isto é, quando procuramos classificar uma situação de perda como sendo, ou não, de risco operacional, devemos sempre procurar identificar as suas causas em vez de nos focarmos nas suas consequências.

2.2. Knowledge Discovery in Databases (KDD)

A progressiva "digitalização" da nossa sociedade deu aso a que qualquer empresa consiga aprender mais sobre os seus clientes. Hoje em dia, a mais pequena acção realizada por um individuo no seu dia-a-dia irá, com grande probabilidade, deixar um rasto digital, consequentemente traduzido em dados. Estes dados estão a ser gerados a uma velocidade sem precedente na história da humanidade e, como tal, a nossa capacidade de os analisar e compreender, enquanto humanos fica bastante aquém da nossa capacidade de os produzir e armazenar. "*We are drowning in information but starved for knowledge"* (Kuonen, 2004, p. 3). Neste sentido, Fayyad et al. (1996b) afirmam que, graças ao desenvolvimento tecnológico a nível de *hardware*, *software* e das técnicas de

armazenamento de dados, é possível, e pouco dispendioso, recolher e aceder a grandes quantidades de dados. Por outro lado, estes mesmos autores consideram que: (i) o processo pelo qual se cria conhecimento a partir de dados é utilizado em variadíssimas áreas desde a ciência, ao marketing, às finanças, ou à saúde; (ii) este processo é maioritariamente manual e depende da interpretação e do grau de familiaridade que o analista tiver com os dados. Através destes dois pontos, os autores concluem que, seja em que área for, uma análise manual de uma base de dados será sempre lenta, dispendiosa e bastante subjectiva, tendo tendência para piorar à medida que o volume de dados aumenta (Fayyad et al., 1996a). Neste contexto, torna-se imperativo desenvolver uma nova geração de técnicas e de ferramentas computacionais, potentes e versáteis, capazes de auxiliar os seres humanos na extracção de conhecimento desses mesmos dados. Esta necessidade está na base do desenvolvimento dos processos de *KDD* e consequentemente do Data Mining.

A literatura define *KDD* como sendo o processo não trivial capaz de identificar nos dados padrões válidos, novos, potencialmente úteis e compreensíveis. Fayyad, Shapiro e Smyth (1996b) entendem que o processo *KDD*: (i) deve ajustar modelos aos dados e encontrar estruturas robustas nos mesmos; (ii) por ser um processo é composto por várias etapas onde cada uma poderá vir a ter várias iterações; (iii) necessita de pesquisa prévia e só posteriormente se deve iniciar a análise, ou seja, não é algo linear como juntar dois valores e fazer a sua soma aritmética; (iv) os padrões descobertos devem ter capacidade de generalização para dados desconhecidos com algum grau de assertividade; (v) por último, os padrões descobertos devem ser novos aos olhos do sistema, assim como úteis para o utilizador que necessitará de os compreender. O *KDD* tem acompanhado a evolução tecnológica das últimas décadas e, mantendo o objectivo da busca pelo conhecimento, tem beneficiado de uma interligação com outras áreas como a estatística, reconhecimento de padrões, visualização de dados, inteligência artificial ou sistemas de bases de dados (Fayyad et al., 1996a).

Como tem sido referido ao longo deste trabalho, todo o processo de *KDD* assenta em dados e na qualidade dos mesmos, contudo, o inicio deste processo requer maior rigor e critério do que a simples produção e recolha destes recursos. Antes de mais, é indispensável compatibilizar todos os diferentes tipos de dados provenientes de diversas fontes, garantir a sua qualidade, organizá-los de forma consistente e garantir a sua fácil mobilização. Foi com o objectivo de resolver estes problemas que nasceu o conceito de *data warehouse*. De facto, este surge como consequência directa da natural evolução das tecnologias de informação, nomeadamente na área das bases de dados e da indústria de recolha e gestão de dados. Segundo Han et al. (2012), esta evolução começou a acentuar-se nos anos 60 com os processos, na altura ainda algo primitivos, de recolha de dados e com a criação das primeiras bases de dados computorizadas com capacidade de processamento de ficheiros. Avançando pelos anos 70 e 80, a problemática assentava na gestão dos dados, incluindo processos de recolha e armazenamento de dados, sistemas de base de dados relacionais, hierárquicos em rede e o desenvolvimento de sistemas assentes em *querys*[2]. Finalmente atingindo os anos 90, o foco passou a incidir nas análises avançadas, possíveis de serem realizadas sobre as bases de dados. Aqui podemos incluir a criação do conceito de *data warehouses* e dos processos *KDD.* Segundo Han et al. (2012), um *data warehouses* é um repositório central de todos, ou de pelo menos uma parte significativa dos dados recolhidos no âmbito de uma actividade de negócio que, regra geral, se encontram alojados num servidor da própria empresa. Estes são criados após os processos de limpeza, integração,

[2] Linguagem informática utilizada em bases de dados para extrair específicas subpartes dos dados em cada momento de acordo com as necessidades do utilizador

transformação, carregamento e actualização de dados serem concluídos, sendo que no caso da actualização esta tarefa é realizada sempre que necessário de acordo com as necessidades dos seus utilizadores. Considerado também como um reservatório que unifica múltiplas e heterógenas fontes de dados, permite uma melhor acessibilidade, selecção e organização destes mesmos dados para que posteriormente possam ser utilizados por aplicações analíticas, com o objectivo de melhorar o processo de tomada de decisão. Inclui também técnicas de limpeza e de integração de dados assim como *online analytical processing (OLAP),* ou seja, técnicas de análise com funcionalidades de agregação, sumarização ou consolidação dos dados, assim como a possibilidade de realizar análises com base em diferentes ângulos ou dimensões, análises multidimensionais. Segundo Kuonen (2004), é possível considerar um *data warehouses* como a memória de uma organização, pela sua capacidade de, em qualquer momento, este a relembrar sobre todas as informações recolhidas sobre qualquer cliente ou operação executada. Contudo, para que a memória tenha utilidade, é necessário traçar outro paralelismo, neste caso com a inteligência. Segundo o autor, a inteligência é conferida pelo processo *KDD* com destaque para o Data Mining. Sendo esta a inteligência que explora a memória em busca de regras, padrões e conhecimento.

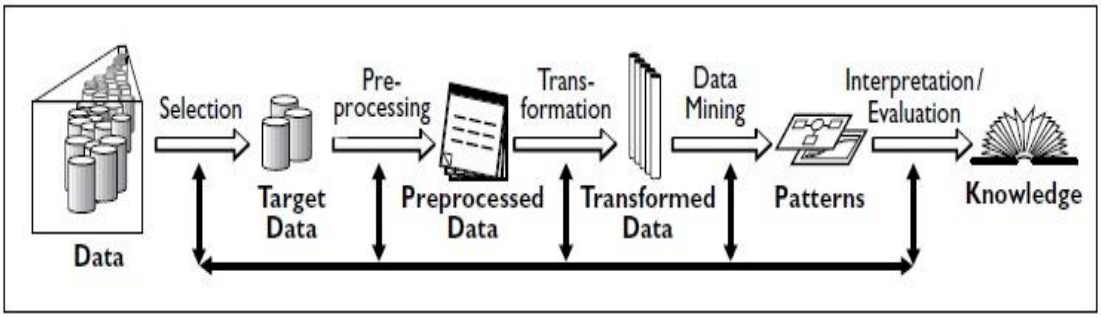


Imagem 1 - Processo *KDD* segundo (Fayyad, Piatetsky-Shapiro, & Smyth, 1996b)

O processo *KDD* é considerado como interactivo, iterativo e composto por diferentes passos, onde a componente humana é determinante em cada um deles (Brachman & Anand, 1994). A *imagem* 1 ilustra o processo *KDD* segundo, Fayyad et al. (1996b), assim como os seus principais passos. Não são contabilizados, por um lado, possíveis *loops* em caso de múltiplas iterações, nem por outro, pequenas acções entre cada passo que irão ficar a cargo da própria lógica e conhecimento adquiridos em etapas posteriores: (i) entender o contexto para o qual irá ser criado, quais os objectivos pretendidos e quais as melhores técnicas a utilizar, tendo em conta, por exemplo, experiências passadas ou o tipo de dados à disposição; (ii) selecionar caso exista, ou criar, um *dataset* de variáveis ou de exemplos para serem usados como alvo, estes dados devem ser recolhidos somente se forem úteis à específica tarefa do analista; (iii) proceder à limpeza dos dados e ao seu pré-processamento, ou seja, remover/tratar *outliers,* valores omissos e ruido, definir tipos de dados ou rever a melhor forma de gerir séries temporais; (iv) ponderar sobre qual a melhor abordagem a ter em relação aos dados, como por exemplo, reduzir a sua dimensionalidade ou proceder à sua transformação. A partir deste ponto o *data warehouses* encontra-se preparado para ser analisado; (v) decidir se o objectivo do processo irá passar por classificação, *clustering*, regressão ou por outra das funções possíveis do Data Mining; (vi) proceder à análise exploratória e definição das hipóteses do modelo: escolher o método apropriado para percorrer os dados em busca de padrões, decidir quais os modelos e parâmetros apropriados; (vii) percorrer os dados em busca de padrões de interesse utilizando uma única representação ou um conjunto de diferentes representações, como por exemplo, árvores de decisão, *clustering* ou

regressão; (viii) visualizar e interpretar os resultados obtidos na forma de dados ou de padrões. Sendo este um processo iterativo, é possível regressar ao primeiro passo e recomeçar até ao sétimo se for necessário; (ix) tomar decisões após a criação de conhecimento, por exemplo, aplicar esse conhecimento a um outro sistema, documentá-lo, reportá-lo ou agir em benefício de uma empresa/negócio.

2.3. Data Mining

Como foi anteriormente mencionado no capítulo 1 e ilustrado na *imagem 1*, o Data Mining constitui a mais importante, distintiva e complexa etapa de todo o processo *KDD.* Sendo este um processo muito complexo, composto por variadas etapas, em que, o objectivo final consiste em extrair padrões indiciadores de futuro conhecimento a partir de dados. Neste sentido o Data Mining, é considerado como a sua mais complexa e importante etapa (Han et al., 2012), sendo uma das suas marcas distintivas, segundo Bação e Painho (2003, p. 137), a "intensidade computacional dos algoritmos utilizados", que assim conseguem "substituir a elegância formal dos modelos matematicamente consubstanciados, pela busca baseada em potentes algoritmos de optimização".

O Data Mining tem provado ser uma ferramenta valiosa pela sua capacidade de analisar grandes quantidades de dados e daí conseguir extrair informação potencialmente valiosa, conferindo às organizações vantagens competitivas em relação aos seus concorrentes (Hormozi & Giles, 2004). A indústria bancária tem vindo a constatar a necessidade de usar este tipo de técnicas como ferramentas estratégicas para a criação de conhecimento e como meio impulsionador da sua própria competitividade, apontam Moin e Ahmed (2012). Já Bhambri (2011) vai mais longe e afirma que somente as organizações que consigam ter o *correcto* acesso à *correcta* informação no *correcto* e exacto período de tempo é que conseguirão dominar o mercado.

Apesar de a literatura usar diferentes definições, todas se aproximam da mesma linha de raciocínio. Bação e Painho (2003, p. 137) afirmam que "Data Mining é o processo não trivial de identificar nos dados padrões que sejam válidos, potencialmente úteis e compreensíveis, procurando traduzir dados em informação e informação em conhecimento, que por sua vez proporcionam a oportunidade de agir com propriedade e racionalidade". Hand et al. (2012), por uma questão de simplificação consideram-no como o processo que, através da descoberta de padrões interessantes em grandes quantidades de dados, consegue assim descobrir conhecimento. Da mesma forma Pivk et al. (2013) consideram-no como o processo de análise de dados, cujo resultado assenta na descoberta de informação implícita mas potencialmente útil, de padrões anteriormente desconhecidos e de relações que se encontram escondidos nos dados.

Diferentes autores apresentam interessantes e pertinentes justificações para a importância do Data Mining enquanto ferramenta capaz de elevar uma organização, conferindo-lhe vantagens competitivas consistentes sobre a concorrência. Moin e Ahmed (2012) afirmam que as ferramentas de Data Mining utilizam dados extraídos de diferentes áreas nomeadamente dados espaciais, multimédia, séries temporais, dados *web*, até aos mais tradicionais como por exemplo os dados referentes às transacções de uma empresa. Valverde (2015), vê o Data Mining como uma área que se distingue de todas as outras pela sua capacidade de procurar e encontrar padrões significativos dentro de um *dataset* e pela forma como é capaz de prever o futuro analisando o passado. Pivk et al. (2013) admitem que o Data Mining deve ser integrado nas várias etapas que constituem um processo de negócio, para

que desta forma seja possível uma maior partilha de informação e de conhecimentos adquiridos, quer ao nível das tomadas de decisão vindos do topo da organização, como também ao nível operacional. Hormozi e Giles (2004) apontam a importância do Data Mining para as organizações pela sua capacidade, não só de potenciar as tomadas de decisão, como também de ajudar a reduzir, em cada momento, a quantidade de dados a utilizar. Por outras palavras, em cada análise os responsáveis podem apenas utilizar a informação que melhor se adequa à resolução do problema em questão, não sendo necessário percorrer todo o reservatório de dados. Igualmente, Koyuncugil e Ozgulbas (2008b), afirmam que, utilizando ferramentas de Data Mining, as organizações serão capazes de prever tendências, comportamentos e, em alguns casos, descobrir informação preditiva inesperada, tornando-as mais proactivas, ao mesmo tempo que lhes confere um maior grau de certeza e confiança aquando das suas tomadas de decisão, visto que estas serão baseadas em conhecimento adquirido.

Contudo, é importante ter particular atenção às palavras de Edelstein (1998), pois na perspectiva do autor, o Data Mining representa um conjunto de ferramentas ao dispor de um analista e nada mais, isto é, não deve ser visto como um truque de magia que instantaneamente observa dados, produz resultados e transmite por correio electrónico as respostas correctas e capazes de potenciar um negócio. De forma alguma elimina a necessidade de um técnico conhecer a realidade do negócio para o qual trabalha, de entender os seus dados e os modelos analíticos a utilizar. Aos olhos do autor, é imperativo que o analista conheça bem os seus dados, ou seja, que saiba identificar e resolver problemas comuns (valores omissos, *outliers*, colunas de dados irrelevantes ou com idêntico poder explicativo) e que tenha uma clara noção de quão aplicáveis estes poderão ser para as tarefas que pretende. A mesma lógica é aplicada ao conhecimento sobre as técnicas ou sobre os algoritmos a utilizar. Nessa linha de raciocínio, Fayyad et al. (1996a) e Valverde (2015) concordam que o Data Mining consegue aproveitar as propriedades informáticas ao seu dispor para ajustar modelos aos dados ou encontrar padrões e relações desconhecidas nesses mesmos dados. Contudo, ambos os autores dão importância à componente humana do analista e à sua apreciação subjectiva no momento de decidir se os resultados provenientes da aplicação de uma determinada ferramenta/método podem ou não vir a ser considerados como acréscimo de conhecimento. Em suma, as técnicas de Data Mining não substituem a competência nem a especialização humana, são sim, potentes e promissoras ferramentas capazes de os auxiliar nas suas análises, ajudando-os a confirmarem por um lado, algumas observações/teorias empíricas previamente realizadas e, por outro, a descobrir novo conhecimento preditivo capaz de conferir valor à organização.

Por sua vez o Data Mining possui algumas características muito próprias que o distinguem da estatística clássica. Autores como Hand (1998), Fayyad et al. (1996a), Bação e Painho (2003), Friedman (1998) e Hand (1999) apontam algumas: (i) o Data Mining, fazendo uso da sua capacidade de processamento, consegue percorrer bases de dados, ou espaços de *input,* de muito maior dimensão (variáveis) e de maior profundidade (campos que descrevem cada registo); (ii) na estatística, o tamanho da amostra é uma preocupação, onde se deseja que esta seja o mais representativo possível da população. No Data Mining, esta questão é uma certeza, existirá certamente um volume de dados suficientemente grande para proceder à amostragem e, como se sabe, amostras de maior dimensão produzem erros de estimação e de variância mais pequenos, levando a inferências mais seguras, mesmo que se esteja a analisar um segmento reduzido da população; (iii) a proveniência dos dados é outro ponto de diferenciação. No caso da estatística os dados são recolhidos com uma finalidade estipulada, as questões de estudo foram levantadas e só depois foram recolhidos os dados para responder a essas mesmas questões (dados primários). Já no caso do Data Mining acontece o

contrário, os dados provêm de diferentes fontes, foram recolhidos com diferentes propósitos e, só depois da selecção, tratamento e alocação dos mesmos num *data warehouse* serão analisados com o intuito de encontrar relações insuspeitas que possam conferir valor ao negócio. Assim podemos afirmar que o Data Mining é um processo que vive de dados secundários; (iv) relacionado ao ponto anterior, podemos afirmar que os dados utilizados na estatística serão fixos, unicamente utilizados para o propósito da análise, o que difere dos dados utilizados no Data Mining que são dinâmicos. Estes últimos têm a vantagem de poderem ser alterados a qualquer momento, tornando o processo de análise mais interactivo e volátil, dando assim, a possibilidade de comparar a mesma realidade usando o mesmo modelo em diferentes períodos de tempo, por exemplo de forma diária ou semanal. Daqui provem grande parte do poder informativo, da capacidade de análise, monotorização e de previsão do Data Mining; (v) ainda na questão dos dados, merece ser referido que enquanto a estatística clássica apenas trabalha com dados numéricos, limpos e de preferência alocados numa única fonte, o Data Mining consegue, por um lado, utilizar dados provenientes de diferentes fontes, não se deixando limitar pelos seus tamanhos e, por outro, é capaz de analisar dados de diferentes tipos, como por exemplo, imagens, áudio, texto, ou dados geográficos; (vi) a estatística, tendo a matemática como base, assenta num maior rigor e conservadorismo, isto é, necessita provar concretamente que um determinado método funciona e só depois é que o irá utilizar, sendo em tese, uma abordagem correcta, por forma a minimizar erros de análise. O Data Mining vai mais além e, tendo herdado algum do experimentalismo das ciências computacionais e das técnicas de aprendizagem máquina consegue, sem descorar o rigor, ultrapassar as metodologias clássicas em prol de resultados favoráveis e por vezes inesperados; (vii) por fim, é possível afirmar que a estatística tem uma abordagem maioritariamente assente em dedução através de hipóteses, ao passo que o Data Mining é uma ferramenta muito mais indutiva.

De acordo com os seus objectivos podemos dividir o processo de descoberta de conhecimento em duas categorias. Segundo autores como Fayyad et al. (1996a), Fayyad et al. (1996b), Kuonen (2004), Valverde (2015) e Han et al. (2012), existe a **verificação**, cujo objectivo passa por utilizar os sistemas informáticos como ferramentas para validar as hipóteses desenvolvidas pelos utilizadores. A segunda categoria é apelidada de **descoberta**, no qual este trabalho se irá centrar, e onde as técnicas de Data Mining têm maior relevância. Aqui, o objectivo passa por utilizar os sistemas informáticos que, de forma autónoma, irão percorrer os dados na procura de padrões/estruturas relevantes, que o utilizador, após a sua análise, deverá traduzir em informação e posterior conhecimento. Dando maior foco ao processo de descoberta, é possível que este seja dividido em dois grandes grupos de acordo com os seus objectivos de partida, destacando de forma inequívoca as bases regentes de cada um. Por um lado, temos a **modelação descritiva**, cuja base assenta na produção de resumos sobre grandes quantidades de dados. Esta confere ao utilizador um maior entendimento sobre os resumos e facilita todo o processo de criação de informação útil e de posterior conhecimento. Por outras palavras, a ideia central passa por extrair padrões compactos que descrevam subconjuntos dos dados, produzir resumos que permitam avaliar rapidamente o potencial desses dados, produzir descrições sobre a sua distribuição e procurar estruturas de semelhança entre si. No fundo, isto implica utilizar ferramentas de natureza eminentemente exploratória, em muitos casos, ferramentas gráficas ou de imagem, que permitam construir interfases com as bases de dados, ou até mesmo técnicas de redução/compressão de dados. Desta forma, sem alienar o que de relevante se deseja reter, conseguimos aumentar o nosso grau de compreensão e de interpretabilidade. Para estes efeitos é possível utilizar técnicas como a análise de componentes principais ou o *Self Organizing Maps* que produzem projecções de baixa

dimensionalidade dos dados, assim como regras de associação e análises de *Clusters*. De referir que a modelação descritiva também é mencionada na literatura como aprendizagem não-supervisionada, em que o analista, por não especificar nenhum tipo de classe ou de variável *target*, não define um objectivo prévio para os seus algoritmos e desta forma, livres de restrições, estes conseguirão percorrer os dados na procura de relações, independentemente do conhecimento prévio que o utilizador possa ter.

No segundo grupo, **modelação preditiva,** o objectivo passa por aprender um critério ou uma regra de decisão, com base em padrões encontrados nos dados, que permita classificar novos exemplos desconhecidos de uma variável de interesse. Assim, a modelação preditiva é constituída por duas fases: (i) fase de aprendizagem, na qual o algoritmo irá extrair conhecimento a partir dos exemplos de treino de que dispõe (estão presentes as variáveis de input e de output), produzindo desta forma um classificador. Isto é, o algoritmo adquire a capacidade, após várias iterações, de atribuir um número ou uma classe a cada exemplo. (ii) fase de predição, na qual o classificador produzido irá ser utilizado para prever o valor/classe de futuros e desconhecidos exemplos (aqui só teremos as variáveis de input, tendo por objectivo a previsão da variável de output, ou dependente). A razão pela qual a literatura se refere a valores/classes diz respeito ao tipo de problemas com os quais o modelo se depara, se o novo exemplo a prever for categórico/discreto (género, sim/não, 0/1) estamos perante um problema de classificação, em que o modelo após analisar as características deste novo exemplo irá associá-lo a uma de várias das classes pré definidas. Este tipo de tarefas auxiliam, por exemplo, os bancos nas análises de *credit* scoring (cujo objectivo passa por classificar um novo pedido de crédito de acordo com uma lista de critérios que definem a escala de risco associada à concessão desse mesmo crédito), ou em processos fraudulentos de pedidos de indeminização a seguradoras (tentar encontrar as variáveis que caracterizam os pedidos fraudulentos através de uma análise histórica, para prever situações idênticas no futuro). Por outro lado, se o novo exemplo a prever for contínuo (peso, altura, idade, rendimento) estamos perante um problema de regressão. De referir que, por vezes, a regressão pode ser utilizada em problemas de classificação, quando em vez de atribuir uma classe queremos atribuir uma probabilidade de pertencer à classe (varia entre 0 e 1), passando a classificação a traduzir-se num limite. Um exemplo de um problema de regressão ocorre quando queremos, através do rendimento de uma pessoa, estimar o custo de uma casa que esta possa suportar (variável dependente). Na fase de aprendizagem iremos aplicar um algoritmo ao conjunto de dados por forma a produzir conhecimento. Este surge na forma de classificador, que neste caso se irá traduzir numa recta. Esta recta irá definir para cada valor de rendimento o preço da casa apropriado. Seguidamente, na fase de predição, um novo cliente irá ser introduzido e o modelo irá produzir uma estimação entre o seu rendimento e o valor da casa que poderá adquirir, baseado na recta de regressão produzida na fase anterior. Alguns algoritmos utilizados na modelação preditiva são as árvores de decisão, as redes neuronais e os classificadores *Naive Bayes*. De referir ainda que parte da literatura apelida a modelação preditiva de aprendizagem supervisionada, pelo facto de o processo ser regido pela tentativa de explicar o comportamento de uma variável *target* através de um conjunto de variáveis de *input.*

Por último e, segundo Bação e Painho (2003, p. 138), "a modelação requer um programa que proporciona instruções detalhadas sobre a forma como o processo se desenvolverá. Estas instruções são tipicamente equações matemáticas, que caracterizam a relação existente entre *inputs* e *outputs*. A formulação destas equações constitui o problema central da modelação." O objectivo será sempre entender a forma como os *outputs* são obtidos por via das variáveis de *input*. Os autores afirmam que

os modelos não paramétricos são normalmente os mais utilizados no Data Mining, são "modelos que dependem essencialmente da utilização de dados, por oposição à utilização de conhecimento específico do domínio do problema", muitas vezes também definidos como modelos *data-driven. "*A premissa essencial dos métodos não-paramétricos é a de que relações que ocorrem de forma consistente no conjunto de dados repetir-se-ão em observações futuras; uma perspectiva iminentemente indutiva"*(Bação & Painho, 2003, p. 139)*.

2.3.1. Análise de *Clusters*

Dada a importância que a análise de *clusters* irá ter numa fase posterior deste trabalho, neste ponto apresentamos uma visão geral deste tópico. Enquanto conceito a análise de *Clusters*, constitui uma actividade básica dos seres humanos. Segundo Kaufman and Rousseeuw (2005), o primeiro passo em qualquer processo de aprendizagem é o reconhecimento, quer isto dizer que quando um novo objecto é identificado, surge a necessidade de o classificar, procurando reconhecer nesse objecto semelhanças e diferenças relativamente ao universo que efectivamente conhecemos. Da mesma forma, uma criança cedo aprende a classificar os objectos do ambiente que a rodeia e a associar as classes daí resultantes a nomes da sua linguagem.

Análise de *Clusters* é a denominação genérica atribuída a uma variedade de metodologias que são utilizadas para classificar entidades. De acordo com Chen et al. (1996), a análise de *Clusters* é um processo que consiste em agrupar objectos físicos ou abstractos, com características semelhantes, em classes para assim construir partições significativas. Por sua vez estas partições irão, permitir uma simplificação dos processos de leitura referentes ao universo em análise. Mediante outro ponto de vista, Han et al. (2012), consideram a análise de *Clusters* como um processo de partição no qual, através de diferentes de algoritmos, se torna possível organizar um conjunto de objectos/entidades/observações multivariadas em subconjuntos. Cada subconjunto é denominado de *cluster*, sendo que os objectos dentro de cada subconjunto devem ser o mais similares possível entre si. Particularmente, no que às suas variáveis caracterizadoras diz respeito, criando assim *Clusters* o mais homogéneos possível, através de uma minimização das distâncias intra-cluster. Em sentido inverso, o objectivo passa também por apresentar *Clusters* o mais distintos possível entre si, *Clusters* com elevado grau de heterogeneidade, através de uma maximização das distâncias inter-cluster. Subjacente está a convicção de que objectos com características semelhantes em termos de variáveis utilizadas, irão ter comportamentos semelhantes. Estes autores defendem que este tipo de análise foi ao longo dos tempos utilizada nas mais variadas áreas, desde a biologia ao marketing, passando pelo reconhecimento de padrões através de imagens, ou à pesquisa web. Contudo, enquanto uma das funções do Data Mining, esta análise consegue ser uma ferramenta autónoma capaz de conferir ao utilizador, não só um maior conhecimento sobre a distribuição dos seus dados, como também permite que este se foque num *cluster* em específico, constituído por observações com determinadas características, na eventualidade de uma futura análise. Sor outra perspectiva, é interessante mencionar que Jain, Murty, and Flynn (1999) consideravam que na análise de *Clusters* exactamente, pela sua natureza essencialmente exploratória, a ideia passa por extrair padrões compactos que descrevam subconjuntos nos dados, enquadrando-se dentro daquilo que a literatura apelida de

aprendizagem não-supervisionada[3]. Este processo pode servir como uma fase de pré-processamento da informação para outros algoritmos/análise, como por exemplo no caso da classificação, em que o foco iria recair apenas sobre um determinado *cluster* de acordo com as suas características distintivas. De facto, segundo estes autores, existe uma grande diferença entre *clustering* e classificação que nem sempre é mencionada: o processo de classificação é iniciado com um conjunto de dados pré-classificados, dados compostos por variáveis de i*nput* que irão ser utilizadas na classificação, assim como a verdadeira classe a que cada registo pertence (variáveis de *output*). Desta forma é possível construir um modelo capaz de prever a verdadeira classe de um novo registo. Um dos maiores exemplos deste tipo de processos é o *credit scoring*, muito utilizado pelos bancos e pelas seguradoras. Por oposição, na análise de *Clusters* não existem dados pré-classificados. O objectivo consiste em utilizar um grande conjunto de objectos, caracterizados por diversas variáveis e tentar substitui-lo por um conjunto manifestamente menor de grupos (*Clusters)* de objectos que sejam semelhante entre si, mas o mais diferentes possível uns dos outros.

A literatura apresenta uma grande variedade de técnicas de *clustering*. Diferentes aspectos são tidos em consideração na hora de escolher o procedimento desejável, nomeadamente qual o critério que irá ser utilizado para minimizar a variância intra-cluster, ou qual o critério a utilizar no momento de maximizar a distância entre *Clusters*. Contudo, e por uma questão de simplificação, iremos utilizar uma distinção mais prática, diferenciando desta forma os **métodos hierárquicos** dos **métodos de optimização**, com destaque para o *k-means*.

Segundo Jain et al. (1999) e Mooi and Sarstedt (2011), os métodos hierárquicos, criam uma decomposição hierárquica de partições $P_1, P_2, \dots, P_h$ do conjunto total dos *n* elementos em *1,2,…h* grupos. A denominação dos métodos hierárquicos advém do facto de que, para cada par de partições P_j e P_{j+1}, cada grupo da partição P_{j+1} estará sempre incluído num grupo da partição P_j. Estes métodos podem ser aglomerativos ou divisivos, de acordo com a forma como a decomposição é feita. *Clustering* aglomerativo, também conhecido como *bottom-up*, inicia-se com cada objecto a representar um *cluster* individualmente. O algoritmo procede em cada iteração e em sentido ascendente à fusão dos *Clusters* com maior grau de similaridade - aqueles cuja distância entre si é mais pequena - até chegar a um único grupo, ou *cluster* que contém todos os objectos, ou até que algum critério de paragem seja atingido. Por oposição a isto, *o Clustering* divisivo, também conhecido como *top-down*, reúne todos os objectos num só *cluster* e, em cada iteração, agora em sentido descendente, vai decompondo esse grupo em *Clusters* de menor dimensão, até que seja possível obter um *cluster* constituído por um objecto apenas, ou até que um critério de paragem seja novamente atingido. Este segundo tipo de *clustering* tem pouca utilização em Data Mining e, como tal, não o iremos mencionar novamente. A principal desvantagem dos métodos hierárquicos, para além da dificuldade em lidar com grandes volumes de dados e da complexidade das suas matrizes, está relacionado com o facto de que, uma vez executada uma qualquer iteração, seja ela aglomerativa ou separativa, não se pode voltar atrás e associar um objecto a um diferente *cluster*. Esta rigidez é uma importante distinção entre estes métodos e os métodos de partição, nomeadamente o *k-means*, que iremos analisar mais à frente.

[3] O analista por não especificar nenhum tipo de variável *target*, não define um objectivo prévio para os seus algoritmos e desta forma, livres de restrições, estes conseguirão percorrer os dados na procura de relações, independentemente do conhecimento prévio que o utilizador possa ter.

Por fim, é fulcral que para se proceder à agregação sucessiva dos objectos, seja necessário estabelecer um critério que defina a proximidade entre grupos. De facto, quando estamos ao nível dos objectos é fácil calcular a proximidade entre si, dado que cada um pode ser representado como um ponto num espaço multidimensional. Contudo, após a primeira agregação, este tipo de proximidade deixa de existir, conduzindo a maiores distâncias entre alguns elementos de cada grupo. De facto, segundo, Mooi and Sarstedt (2011), existem vários procedimentos de aglomeração que se distinguem pela forma como conseguem definir as distâncias entre *Clusters* recentemente formados e os restantes objectos ainda por agrupar. De referir que diferentes definições destas distâncias poderão resultar em diferentes soluções finais. Os critérios de agregação mais utilizados são os seguintes: (i) *single linkage* (critério do vizinho mais próximo). Define como semelhança entre dois *Clusters* a menor das distâncias entre os seus elementos mais próximos; (ii) *complete linkage* (critério do vizinho mais afastado). Define como semelhança entre dois *Clusters* a maior das distâncias entre os seus elementos mais afastados ou menos semelhantes; (iii) *unweighted pair-group* (critério da média dos grupos). A distância entre dois *Clusters* é definida como a média das distâncias entre todos os pares de objectos desses dois *Clusters*; (iv) *unweighted pair-group centroid* (critério do centróide). A distância entre dois *Clusters* é definida pela distância entre os centroides de ambos os *Clusters*, sendo os centroides os pontos definidos pelas médias das variáveis caracterizadoras dos objectos de cada *cluster*; (v) método de Ward. Este método, em vez de ir juntando sucessivamente os objectos como maior grau de similaridade, procede à junção dos objectos que garantam a variância intra-cluster mais baixa possível. Este método é útil, quando o objectivo passa pela criação de *Clusters* proporcionais em termos de tamanho e quando não existem *outliers* nos nossos dados.

Os métodos de partição são o segundo grupo de procedimentos de *clustering* em destaque. Segundo Han et al. (2012) e Jain et al. (1999), tendo uma base de dados *D* com *n* objectos e *k* número de *Clusters* para criar, um algoritmo de partição consegue organizar os objectos em *k* partições, sendo $k \leq n$, onde cada partição irá representar um *cluster*. Estes são desenvolvidos tendo em consideração a optimização de um certo critério, uma função de semelhança, como por exemplo uma distância. Garantindo desta forma um grau de similaridade elevado entre os objectos agrupados num determinado *cluster*, ao mesmo tempo que os objectos classificados em diferentes *Clusters* irão ser tão diversificados quanto possível. Neste capítulo iremos apenas descrever o procedimento de partição *k-means*, sendo este o processo que iremos utilizar numa fase posterior deste trabalho, nomeadamente quando for desenvolvida a análise de *Clusters* à base de dados de risco operacional. Segundo Kaufman and Rousseeuw (2005), este procedimento de *clustering* é um dos mais antigos, dos mais simples de implementar, dos que produz melhores resultados e, como tal, o mais utilizado também. Segundo Jain et al. (1999) o algoritmo *k-means* recebe como *input* um parâmetro *k* e procede à partição de *n* objectos em *k* grupos, de forma a garantir uma elevada semelhança intra-cluster e uma elevada dissemelhança inter-cluster. A semelhança/dissemelhança é medida em função dos valores médios dos objectos que constituem cada *cluster,* não havendo a necessidade de escolher uma medida de distância para que seja possível iniciar a análise (distância Euclideana ou a distância de Manhattan por exemplo), como ocorre nos métodos hierárquicos. Han et al. (2012) afirmam também que o *K-means* se baseia na variância intra-cluster e inter-cluster como critério para a criação de *Clusters* homogéneos no seu interior e o mais heterogéneos entre si. Ao mesmo tempo estes autores afirmam que as medidas de homogeneidade/heterogeneidade são calculadas pela aproximação de cada individuo à média do seu *cluster*/afastamento da média do *cluster* à média global dos objectos. No caso da aproximação, esta é calculada através da soma dos quadrados das diferenças de cada objecto

à média do *cluster*, de certa maneira analisando de que forma os objectos se concentram em redor do "centro" do seu *cluster*. No caso do afastamento da média do *cluster* à média global dos restantes *Clusters*, esta é obtida pela soma dos quadrados das diferenças da média de cada *cluster* à média global dos vários *Clusters*. Podemos considerar o *k-means* como um processo iterativo que permite uma combinação dos objectos de um espaço multidimensional em *k* grupos homogéneos no seu interior, sendo estes o mais heterogéneos possível entre si. A combinação resultante deste processo interactivo é considerada a repartição óptima

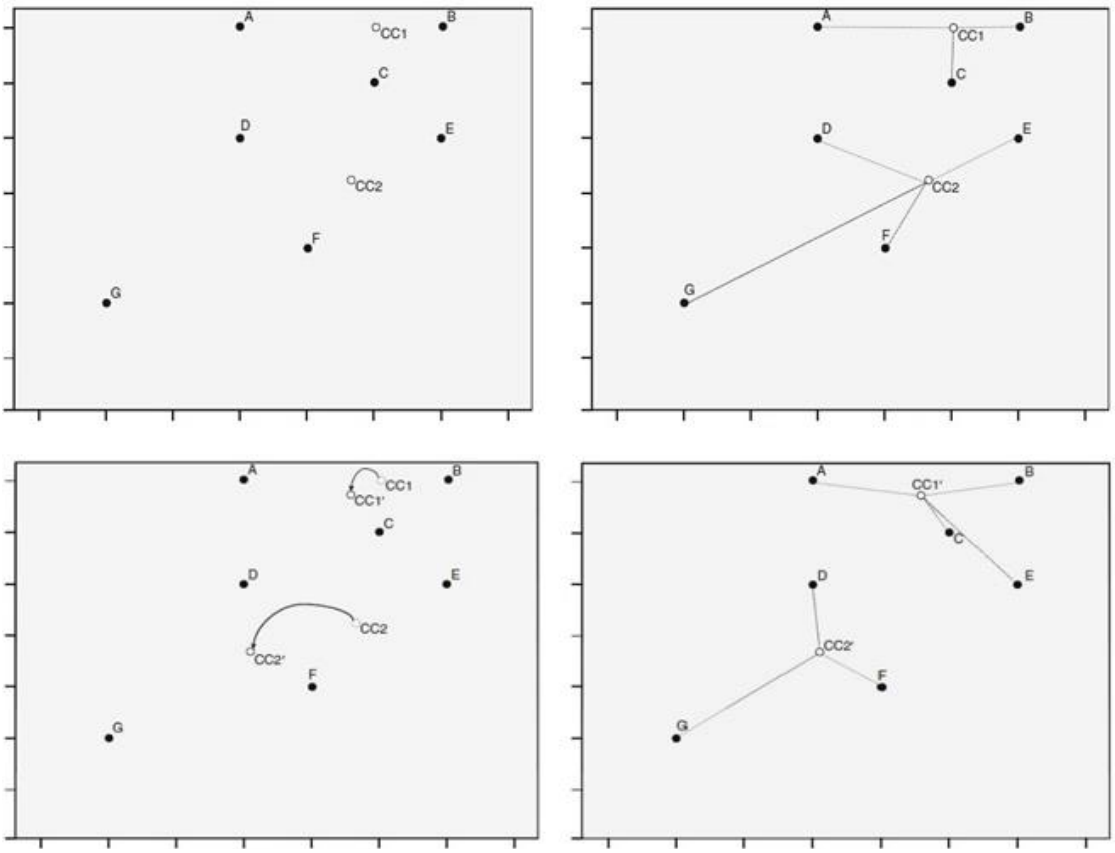


Imagem 2 – Processo k-Means segundo (Mooi & Sarstedt, 2011) : topo esquerdo 1ª fase (i), topo direito 2º fase (ii), fundo direito 3º fase (iii), fundo esquerdo 4º fase (iv)

Segundo Mooi and Sarstedt (2011), e como ilustra a *imagem 2*, podemos decompor o processo da seguinte forma: (i) após estar definido o número de *k Clusters* a utilizar, o algoritmo de forma aleatória irá selecionar *k* objectos, ou *seeds*, que irão representar o centroide, o centro de cada *cluster*; (ii) as distâncias Euclidianas entre os restantes objectos e os centroides dos *Clusters* previamente definidos são calculadas. Cada objecto é alocado ao *cluster* cuja distância para o seu centroide for menor. A partição inicial dos objectos está concluída; (iii) estando a fase de atribuição completa, os centroides dos *Clusters* são recalculados. Desta forma é necessário voltar a actualizar a média dos diferentes valores que constituem cada observação de cada *cluster*, o que se irá traduzir numa actualização da média do próprio cluster. Este cálculo irá fazer com que os centroides previamente definidos se movimentem de posição no espectro multidimensional; (iv) nesta fase as distâncias entre os objectos e os centroides têm de ser também elas recalculadas, visto que o centro do *cluster* já não se encontra no mesmo sítio. Como consequência disto, os objectos são novamente recolocados no *cluste*r com menor distância entre si e o centroide do respetivo grupo. Provavelmente irá ocorrer uma mudança de *cluste*r em bastantes observações o que levará a uma alteração da própria estrutura dos *cluster*, assim como, a uma alteração das suas próprias características. O algoritmo irá repetir os passos

(iii) e (iv) até que o número de iterações previamente definido seja atingido, ou até que não seja possível recentrar os centroides de nenhum *cluster*.

Este algoritmo apresenta resultados superiores relativamente aos métodos hierárquicos, principalmente quando a análise recai sobre um maior conjunto de dados multivariados, na medida em que requer menos poder computacional para ser executado, ou também, quando os *Clusters* são constituídos por nuvens compactas e separadas de forma clara umas das outras (Han et al., 2012). Contudo, o algoritmo *k-means* apresenta algumas deficiências: este algoritmo é bastante sensível à partição inicial, ou seja, ao número de *k Clusters* que o utilizador seleciona no inicio do processo, assim como às posições iniciais das próprias *seeds.* Por outro lado, este algoritmo tem algumas limitações para encontrar *Clusters* com formas complexas, ou seja, *Clusters*, com formas não-convexas. Finalmente este método é muito sensível à presença de "ruído" e de *outliers*. Isto ocorre pelas constantes actualizações das médias de cada *cluster,* sendo as médias altamente enviesadas por este tipo de dados (Mooi & Sarstedt, 2011).

3. METODOLOGIA DE INVESTIGAÇÃO

Este capítulo será dividido em duas partes: **em primeiro** lugar, irá ser realizada uma apresentação e uma análise global da base de dados de risco operacional providenciada por uma instituição financeira, a fonte de todos os dados utilizados ao longo deste trabalho. Iremos enunciar em linhas gerais a constituição da fonte de dados, assim como algumas das variáveis que ajudam a caracterizar os variados eventos de risco. Algumas considerações irão ser feitas, sobre a existência de valores omissos. **Em segundo lugar,** irá ser feita uma descrição de todo o processo de tratamento de dados que foi executado: (i) iremos descrever o modo como trabalhámos os dados de forma a permitir a sua análise exploratória. Serão elaborados gráficos que auxiliam na compreensão da *frequência* e da *severidade* dos eventos de risco ocorridos, tendo estes sido repartidos por diferentes intervalos temporais, assim como pelas suas *categorias* e *processos* de risco. Aqui iremos aprofundar a forma como estes gráficos foram construídos, a lógica e os critérios por detrás de tal construção, tendo sempre em conta a matéria-prima (dados) à disposição; (ii) iremos descrever a forma como foi criada uma segunda tabela de dados. Iremos descrever as colunas que a constituem e, a forma como em alguns casos essas mesmas colunas tiveram de ser criadas de raiz. Iremos desta forma explicar a lógica por detrás de tais criações; (iii) por último iremos mencionar o diagrama produzido pelo *software SAS ENTREPRISE MINER* e explicar alguns dos seus componentes. Estes passos foram essenciais durante a análise de *Clusters* posteriormente produzida.

3.1. Dados

Irá ser utilizada uma base de dados de risco operacional proveniente de uma instituição do sector financeiro Português. Estes dados foram recolhidos pela própria instituição financeira através de um interface informático que foi disponibilizado a todas as direções e balcões de forma a possibilitar a recolha e o armazenamento de eventos de risco operacional. Esta fonte de dados consiste num conjunto de 3594 eventos de risco. Destes, 1597 não se materializaram em perdas financeiras para a instituição, ao contrário das restantes 1997 que efectivamente geraram. Todos os eventos ocorreram entre 1 de Janeiro de 1993 e 14 de Julho de 2014 e, a cada evento correspondem 28 atributos que o caracterizam. Contudo, existem casos em que muitos destes atributos não se encontram preenchidos, como iremos ilustrar mais adiante. Alguns dos atributos utilizados dizem respeito a datas que determinam as diferentes etapas de um evento de risco, nomeadamente a data em que este se inicia, quando é descoberto e quando é finalizado. Por outro lado, foram também utilizadas variáveis quantitativas correspondentes a perdas efectivas que cada evento teve para a organização. Por último, é importante referir que a base de dados continha outros dois grandes grupos de atributos: (i) atributos referentes às categorias de risco, onde se inserem as categorias de Basileia com os seus respectivos quatro níveis de atributos, assim como as categorias de risco referentes à actividade da instituição bancária em concreto e, os seus respectivos quatro níveis; (ii) atributos referentes aos processos de risco que caracterizam os eventos. Este último apresenta quatro níveis distintos de atributos.

A base de dados encontra-se com bastantes valores omissos em diferentes colunas: por exemplo, nas colunas que dizem respeito às datas existem 733 eventos de risco sem data de início, 1225 sem data de descoberta e 1204 sem data de termo. Existem também 677 registos que apenas possuem data de descoberta, sem início nem fim, contudo 57 destes apresentam perdas somadas de quase meio milhão de euros. Existem também 56 eventos sem nenhuma datas, sendo que 33 destes efectivamente causaram perdas para a instituição num total a rondar os 60.000€. Como consequência

destas inconsistências as análises elaboradas ao longo deste trabalho irão incidir sempre sobre uma parte parcial dos nossos dados de acordo com o seu propósito, mantendo o máximo de representatividade possível.

3.2. Tratamento de Dados

O processo de tratamento de dados, foi realizado com dois objectivos em mente. Num **primeiro momento** o objectivo passou pela realização de uma análise exploratória da base de dados. Esta análise foi dividida, por sua vez, em duas fases, ou seja, foi elaborada uma análise das variáveis *frequência* e *severidade* à luz de intervalos temporais previamente criados, mas também, utilizando as *categorias* e os *processos* de risco com maior significância e representatividade. Num **segundo momento** o objectivo passou pela criação de uma nova tabela com as seguintes colunas retiradas da base de dados inicial: identificador do evento, tipo de processo de risco e tipo de categoria de risco. Às colunas mencionadas foram criadas e acrescentadas mais três: cada qual com uma legenda e uma codificação assentes numa lógica que mais à frente irá ser explicada. A tabela criada foi a fonte de dados final e, aquela sobre a qual recaiu a análise de *Clusters*.

Com o objectivo de realizar a análise exploratória foi, numa primeira fase, necessário criar gráficos que ilustrassem duas variáveis/realidades diferentes: a *frequência* e *severidade* dos eventos. Desta forma, foram criados dois gráficos para cada cenário de acordo com as suas datas de início, descoberta e fim. Por exemplo, o primeiro gráfico que irá ser apresentado mais adiante analisa somente os eventos de risco que contenham obrigatoriamente uma data de início e uma data de descoberta e, dentro desse conjunto iremos analisar a sua *frequência*, ou seja, iremos quantificar o número de eventos que ocorreram em diferentes períodos temporais e, desses totais iremos identificar quantos foram os eventos que se traduziram em perdas para a organização. Da mesma forma, iremos analisar a sua *severidade*, ou seja, em que intervalo temporal é que ocorreu a maior percentagem de perdas, em volume, para a organização. Para que esta lógica não fosse comprometida pelo tipo de dados de que dispúnhamos, foi necessário proceder ao tratamento, transformação e criação e algumas colunas da base de dados: (i) utilizando a ferramenta excel, procedemos ao cálculo das diferenças entre as três datas, obtendo assim valores inteiros e quantificáveis. De referir que, segundo a fórmula, 1 ano corresponde a 360 dias. Desta forma foi possível observar quantos dias ocorreram entre cada diferente momento de um evento de risco e, criar intervalos temporais para a realização de gráficos. Desta forma, foram criadas três novas colunas na base de dados; (ii) de seguida resolvemos o problema dos valores omissos nas colunas temporais descoberta e fim. Foi criada uma data limite de 1/1/2016 para todas essas células, de forma a harmonizar os dados, caso contrário ao executar o passo anterior, os valores atingidos nestas células seriam de desmedida/irreal proporção. A partir daí recomeçamos o passo anterior. Contudo, apesar desta tentativa e, como irá ser ilustrado mais à frente, alguns gráficos acabaram por ser criados com base em intervalos exclusivamente referentes a anos e não a diferenças de dias. Isto ocorre principalmente em casos onde as datas de início são muito antigas, conferindo uma diferença de dias muito discrepante em relação à data de termo estipulada. De notar que todos os 677 eventos que não possuem data de início foram retirados de qualquer análise por uma questão de lógica; (iii) por último, foi criada uma coluna binária para ajudar a compreender e a quantificar o número de eventos que tinham efectivamente originado perdas (1) e os casos onde tal não ocorreu (0).

A segunda fase da análise exploratória teve como objectivo analisar as variáveis *frequência* e *severidade* mas, ao contrário do que foi feito anteriormente, onde a componente temporal era o mais importante, aqui o foco recaiu essencialmente sobre os diferentes tipos de processos de risco e de categorias de risco que ajudam a explicar a natureza dos eventos ocorridos na instituição financeira. No que aos processos de risco diz respeito, é importante referir que foi utilizada apenas a coluna referente aos processos de segundo tipo, apesar de existirem ainda os de primeiro, terceiro e quarto. Os critérios de escolha dizem respeito, por um lado, ao facto de os processos de primeiro tipo serem demasiadamente agregadores, ou seja, apresentam poucos valores distintivos e, por outro lado, o facto de os processos de terceiro e quarto tipo serem demasiadamente específicos e apresentarem um elevado número de valores omissos na base de dados. Desta forma optámos pela coluna que garantia um equilíbrio entre a representatividade e uma capacidade de caracterização de um evento equilibrada.

De referir que, por um lado, a base de dados de risco operacional possui 54 diferentes tipos de processos de risco de segundo tipo, contudo por uma questão de representatividade e de capacidade de leitura foram apenas selecionados os 21 processos cuja *frequência* era superior a 10 eventos. Por outro lado, realçar que o tipo "sem nome" e "outros" foram incluídos pelo seu alto volume de ocorrências, apesar de não se entender quais as suas origens. Referir também que, no gráfico referente à severidade, o processo "sem nome" possuía um *outlier* (valor discrepante em relação à distribuição média da variável). Neste caso um evento, sem nome de processo, cujas perdas para a instituição eram superiores a 14 milhões de euros, quando o total de perdas de todos os processos e respectivos eventos somados não chega a 28 milhões. Foi por isso retirado para não enviesar a análise do gráfico de severidade. Por último, referir que dos 54 tipos de processos inicias foram apenas selecionados os 13 cuja severidade foi mais significativa a nível de volume de perdas para a instituição. Foram selecionados os tipos de processos cujas perdas acumuladas foram iguais ou superiores a 100.000€.

No que às categorias de risco diz respeito, é importante referir que foi utilizada apenas a coluna referente às categorias de segundo tipo, apesar de existirem ainda as de primeiro, terceiro e quarto tipo, assim como as categorias de Basileia. De facto, o critério para a selecção da coluna referente ao segundo tipo em detrimento das restantes foi semelhante ao critério utilizado para os processos de risco: representatividade e capacidade de caracterizar de forma equilibrada um evento de risco. A grande diferença foi a razão pela qual se optou por uma categoria de risco de segundo tipo em vez do seu equivalente segundo Basileia. Nesta decisão entrou em consideração o facto de que, enquanto as categorias de Basileia são linhas gerais muito padronizadas, capazes de serem aplicadas a todos os eventos de todas as instituições bancárias, as restantes categorias dizem efectivamente respeito às actividades da instituição financeira em causa e, como tal, faz mais sentido optar pelas segundas.
De referir também que, a base de dados contem 20 diferentes categorias de risco de segundo tipo sendo que, uma destas categorias alberga os eventos com valores omissos aos quais apelidámos de "sem nome". Após futuras análises e, por uma questão de representatividade, optou-se por apenas utilizar as 13 categorias cuja *frequência* era superior a 10 eventos ocorridos, independentemente destes se terem traduzido em perdas para a instituição ou não. Na realização da análise de *severidade* foi decidido que, por uma questão de lógica e de representatividade, seria melhor ignorar todas as categorias cujas perdas acumuladas não fossem iguais ou superiores a 100.000€.

O segundo grande objectivo encontra-se relacionado com a criação de uma nova tabela de dados. Esta tabela foi a nova fonte de informação aquando da utilização das técnicas de Data Mining (análise de *Clusters*) numa fase posterior do trabalho e é composta por 2519 registos (70% da fonte de

dados original). Inicialmente do total de 3594 registos, foram retirados apenas os 733 eventos que não continham data de início. Contudo, conclui-se que seria mais lógico, por um lado, retirar os 81 eventos que, apesar de conterem data de início, não continham simultaneamente data de descoberta nem data de termo e, de seguida, retirar todos os eventos que apresentassem valores omissos nas colunas de processos e de categorias. Desta forma, foram retirados 155 eventos sem valor na coluna de processos, 106 sem valor na coluna de categoria. Referir também que a nova tabela é composta por seis colunas: três foram retiradas da base de dados original, como é o caso das colunas: *identificador* de evento, tipo de *processo* e tipo de *categoria de risco*. As três colunas criadas de raiz são: *montantes perdidos*, *duração* de um evento de risco e *detecção* de um evento de risco.

A primeira coluna criada diz respeito aos montantes que um determinado evento teve de prejuízo financeiro para a instituição. Após analisar todos os valores referentes a estas perdas, foram criados intervalos por forma a agrupar os diferentes prejuízos facilitando assim a própria leitura. A *Tabela 3* ajuda a ilustrar os intervalos criados, assim como o número de eventos inseridos em cada um.

Intervalo	Legenda	Nº Eventos
S/ Valor	L0	21
[0 - 100€]	L1	1152
[101 - 500€]	L2	536
[501 - 1500€]	L3	374
[1501 - 10.000€]	L4	280
[10.001 - 50.000€]	L5	93
[50.001 - 300.000€]	L6	54
[300.001 - 1.000.000€]	L7	7
> 1.000.000€	L8	2

Tabela 3 – Variável criada referente aos montantes perdidos

A segunda coluna criada diz respeito à duração de cada evento de risco, ou seja, a diferença de dias que decorreram entre a sua data de início e a sua data de termo. De referir que 403 eventos não possuíam data de fim, contudo foram incluídos na análise visto que preenchiam os critérios acordados para os tipos de processos e de categorias. Todos estes eventos foram assim colocados numa classe intitulada "*S/ Data de Termo*". A *Tabela 4* ajuda a ilustrar os intervalos criados, assim como o número de eventos inseridos em cada um.

Intervalo	Legenda	Nº Eventos
S/ Data de Termo	D0	322
24 Horas	D1	287
1º Mês	D2	733
Primeiros 3 Meses	D3	436
Entre 3 Meses e 1 Ano	D4	435
Entre 1 e 3 Anos	D5	198
Entre 3 e 10 Anos	D6	96
Mais de 10 Anos	D7	12

Tabela 4 – Variável criada referente à duração de um evento

A última coluna foi criada para observar principalmente o número de dias que decorrem desde que um evento é iniciado até à sua descoberta e, em segundo plano, observar se a sua descoberta é feita após este ter sido finalizado ou se aconteceu antes. Um evento que seja descoberto antes da sua data de termo tem a sua legenda registada na coluna A, se o contrário ocorrer, a sua legenda irá aparecer na coluna *B*. De referir que existem 1044

Intervalo	Nº Eventos	Legenda (A)	Legenda (B)
24 horas	561	A1 = 561	B1 = 0
1º Mês	365	A2 = 303	B2 = 62
1º Ano	369	A3 = 225	B3 = 144
[1 - 3 Anos]	97	A4 = 88	B4 = 9
[3 - 7 Anos]	68	A5 = 54	B5 = 14
[7 - 10 Anos]	4	A6 = 3	B6 = 1
> 10 Anos	11	A7 = 4	B7 = 7

Tabela 5 - Detecção de um evento antes e após a sua data de termo

eventos (30% do total) sem data de descoberta, mas com data de termo. Este número foi considerado demasiadamente elevado para se ignorar e, como tal, foi definida novamente a data 1/1/2016 como a data de descoberta de todos estes eventos. Sendo esta data superior a qualquer data de fim existente na base de anos, estes eventos foram automaticamente colocados na coluna *B* (eventos que são descobertos depois da sua data de termo). Para este cenário foi atribuída a legenda *B0*. Esta lógica não foi seguida para os 322 eventos que apresentam data de descoberta mas que não apresentam valores para a data de termo. Visto que é possível calcular as reais diferenças de dias entre as suas datas de início e de descoberta, estes eventos podem ser agrupados em diferentes intervalos de acordo com o número de dias em questão, não sendo necessário criar uma classe à parte (*A0*), para que a sua inclusão na análise seja possível. Por outro lado parece importante referir que todos os eventos que foram iniciados e descobertos no mesmo dia, mesmo aqueles cuja data de termo corresponde a esse mesmo dia, foram aglomerados na classe *A1*. Esta é a razão pela qual não existem valores para *B1*, dado que estamos apenas a falar de um período de 24 horas. A *Tabela 5* ajuda a ilustrar os intervalos criados, assim como o número de eventos inseridos em cada um.

Por último iremos mencionar alguns dos passos que foram tomados na construção do diagrama que permitiu a elaboração da análise de *Clusters*. Para realizar esta análise foi utilizado o *SAS Enterprise Miner*, um *software* apropriado à realização de diversas técnicas de Data Mining e estatística, onde através de diagramas de processos interactivos e dinâmicos, se torna possível fazer diversas análises aos dados, com um considerável grau de interpretabilidade dos mesmos. O *Anexo 1* ilustra o diagrama referente a todo este processo. Em linhas gerais, o processo inicia-se com uma selecção da fonte de dados, seguindo-se uma análise exploratória dos mesmos, através do nó *StatExplore*. De seguida, e após se escolher quais as variáveis da fonte de dados que irão efectivamente entrar no processo, inicia-se a análise de *Clusters* no respectivo nó e posteriormente será feita uma análise aos perfis de cada grupo criado. Numa última nota, o nó *SASCode* foi utilizado para a elaboração do gráfico do cotovelo, como irá ser ilustrado e explicado no ponto 4.3 (análise de *Clusters*).

4. ANÁLISE DE RESULTADOS

No capítulo anterior foi descrito todo o processo de tratamento de dados que permitiu a execução, por um lado, da primeira análise exploratória (elaborada sobre toda a base de dados) e, por outro, a criação de novas tabelas de dados. Neste capítulo iremos apresentar a análise exploratória efectuada. Esta análise assenta em gráficos, elaborados com o objectivo de explicar o comportamento das variáveis *frequência* e *severidade,* quer numa primeira fase, onde foram criados intervalos temporais, quer numa segunda, onde estas variáveis foram analisadas à luz dos seus processos e categorias de risco. Por outro lado, iremos apresentar também uma segunda análise exploratória, cujo foco irá incidir sobre a segunda tabela de dados, mencionada no ponto anterior. Esta tabela, composta por colunas previamente existentes e por outras que foram posteriormente criadas, irá ser a fonte de informação a analisar aquando da utilização das técnicas de Data Mining.

4.1. Análise Exploratória

A análise exploratória de dados corresponde à parte da modelação descritiva do Data Mining, onde o objectivo passa por produzir resumos sobre grandes quantidades de dados, de forma a conferir ao utilizador um maior entendimento sobre os mesmos, assim como, facilitar todo o processo de criação de informação útil e de posterior conhecimento. Sendo um dos pioneiros desta abordagem Tukey (1977), defende que todos os estatísticos deviam explorar os dados, por forma a maximizar o seu conhecimento sobre os mesmos, encontrar estruturas robustas e desconhecidas, proceder ao tratamento de algumas inconsistências como valores omissos, e finalmente formular hipóteses que conduzissem a novo conhecimento e novas descobertas. Esta abordagem assenta maioritariamente em técnicas visuais e gráficas para que a exploração seja o mais livre e "*open-minded*" possível. De seguida irão ser apresentados os gráficos realizados para um melhor entendimento das variáveis *frequência* e *severidade* dos eventos de risco observados.

4.1.1. Eventos Iniciados e Descobertos

Em sintonia com o que foi mencionado previamente e, como é possível observar pelos gráficos em baixo apresentados, sentiu-se a necessidade de criar intervalos temporais de acordo com as indicações providenciadas pelos dados. Desta forma criaram-se os seguintes intervalos: eventos iniciados e descobertos nas primeiras **24 horas**, **no primeiro mês** (30 dias), **nos primeiros três meses** (31 a 90 dias), **entre três meses e um ano** (91 a 360 dias), **entre 1 e 2 anos** (361 a 720 dias), **entre 2 e 3 anos** (721 a 1080 dias), **entre 3 e 5 anos** (1081 a 1800 dias), **entre 5 a 15 anos** (1801 a 5400 dias) e eventos **superiores a 15 anos**, ou seja, eventos que desde a sua data de início, demoraram mais de 5401 dias até serem descobertos. De referir que para a realização destes dois gráficos apenas foram utilizados 47,1% dos registos à disposição o que perfaz um total de 1692 eventos. Como foi referido no exemplo do capítulo anterior (3.2), apenas foram contabilizados os eventos que apresentavam valores para ambas as datas, sendo que todos os outros foram excluídos, desta forma ignoraram-se 1169 registos (32.5%) sem data de descoberta e 733 (20.4%) sem data de início. Por último referir que dos 1692 registos utilizados 801 (47.3%) não apresentam perdas, ao passo que 891 (53.7%) apresentam.

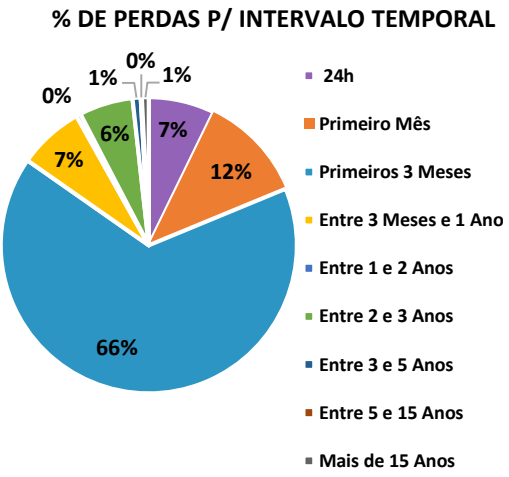


Gráfico 2 - Análise da variável severidade

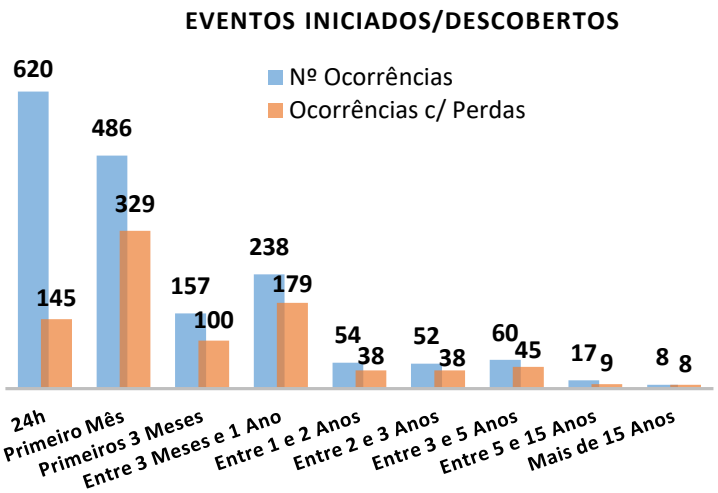


Gráfico 1 - Análise da variável Frequência

Fazendo uma leitura de ambos os gráficos apresentados em cima é possível concluir: (i) praticamente 89% dos 1692 eventos analisados foram iniciados e descobertos entre as primeiras 24 horas e o primeiro ano. Este período concentra 753 eventos com perdas para a instituição, ou seja, 85% do total de eventos com efectivas perdas. Por outro lado, esses eventos são responsáveis por 92% do total de prejuízos obtidos. No entanto, 66% deste último valor tem origem em apenas 179 eventos, iniciados e descobertos durante os primeiros 3 meses; (ii) curioso constatar que os dois primeiros intervalos, apesar de juntos, concentrarem 53% dos eventos com efectivas perdas, estas apenas se materializam em 19% do total dos prejuízos; (iii) não existe nenhuma relação entre um maior número de dias ocorridos entre o início e a descoberta de um evento e um possível aumento das perdas para a instituição. Aliás, é possível observar que eventos com diferenças, por exemplo, entre os dois e os quinze anos, apesar de mais de metade serem efectivamente eventos com perdas para a intuição, estes têm um impacto praticamente marginal não atingindo sequer os 10% do total de perdas.

4.1.2. Eventos Iniciados não Descobertos

Estes gráficos foram elaborados, utilizando para o efeito, intervalos com base em anos e não em diferenças de dias. Por outras palavras foram contabilizados, por exemplo, quantos eventos foram iniciados entre 1997 e 2004 e que ainda não foram descobertos. Analisámos também quantos é que efectivamente registaram perdas para a instituição. Foram utilizados 1169 registos (32.5%), sendo que 1016 (86.9%) apresentaram prejuízo, ao passo que 153 (13.1%) não apresentaram.

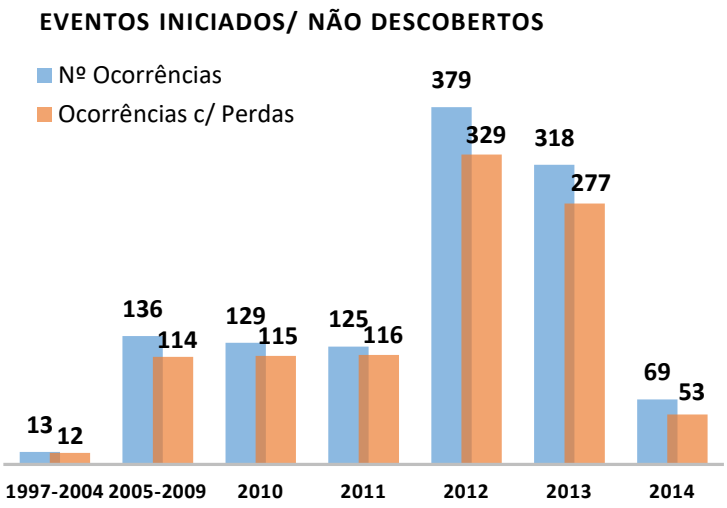


Gráfico 3 - Análise de variável Frequência

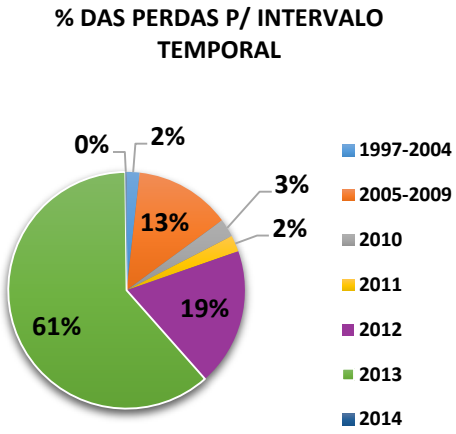


Gráfico 4 - Análise da variável Severidade

Fazendo uma leitura de ambos os gráficos é possível concluir: (i) em todos os intervalos criados existe, para cada total de eventos ocorridos, um valor de pelo menos 75% que corresponde a eventos com perdas efectivas; (ii) do total de eventos com perdas, 60% ocorreram entre 2012 e 2013, sendo que entre 2005 e 2013 ocorreram 93%; (iii) apesar de nos três intervalos, 2005-2009, 2010 e 2011 os totais de eventos com perdas serem praticamente os mesmos - 114, 115 e 116 - os seus prejuízos totais são bem diferentes com o primeiro o obter um total de 15% de perdas somadas, ao passo que os restantes apenas têm 3% e 2% respectivamente. (iv) os dois anos onde ocorreram maior número de eventos com perdas são também os que apresentam maior preponderância em termos de severidade: 2012 é o ano com um maior número de eventos e corresponde a 19% do total de perdas, mas é em 2013 que verificamos maior percentagem de danos financeiros atingindo os 61%, sendo este o segundo ano com maior número de eventos de risco.

4.1.3. Eventos Iniciados e Terminados

Os gráficos apresentados em baixo seguiram a mesma lógica dos gráficos 1 e 2 aquando da sua realização. Foram retirados 733 (20.4%) registos que apresentavam valores omissos em ambas as colunas e, foram também retirados 471 registos (13.1%) que não apresentavam valores apenas na coluna referente ao termo de um evento. No total foram utilizados 2390 registos (66.5%), sendo que 1841 (77%) apresentaram prejuízos para a instituição, ao passo que 549 (23%) não apresentaram.

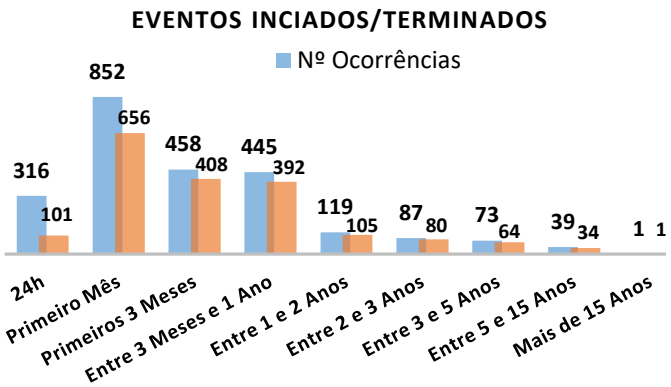


Gráfico 6 - Análise da variável Frequência

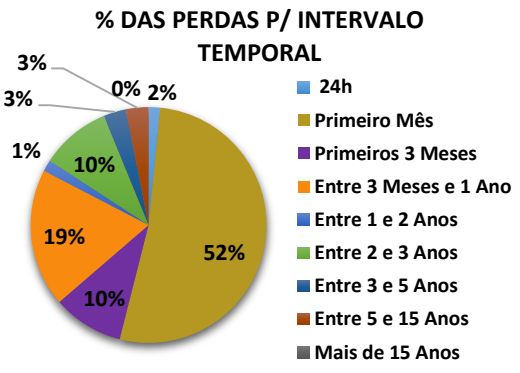


Gráfico 5 - Análise da variável Severidade

Fazendo uma leitura de ambos os gráficos é possível concluir: (i) após o primeiro mês, mais de 80% dos eventos ocorridos são efectivamente eventos com perdas em todos os intervalos criados; 73% dos eventos ocorrem entre o primeiro mês e o primeiro ano, sendo que neste mesmo período ocorrem cerca de 80% dos eventos com perdas. Este valor de eventos com perdas corresponde, a nível de severidade, a 81% do total de perdas verificadas entre eventos que foram iniciados e terminados; (ii) os eventos iniciados e descobertos no primeiro mês correspondem a 51% do total de perdas obtidas, sendo este o intervalo com maior número de eventos prejudiciais à instituição (656). Contudo o intervalo dos 2 aos 3 anos apesar de apenas ter 80 eventos deste tipo contribui com 10% para o total de perdas.

4.1.4. Eventos Iniciados não Terminados

Os gráficos apresentados em baixo seguem a mesma lógica dos gráficos 3 e 4 anteriormente apresentados. Foram utilizados os 471 (13.1%) eventos com data de início e sem data de termo, sendo que desses, 405 (86%) não apresentaram perdas, ao passo que 66 (14%) apresentaram.

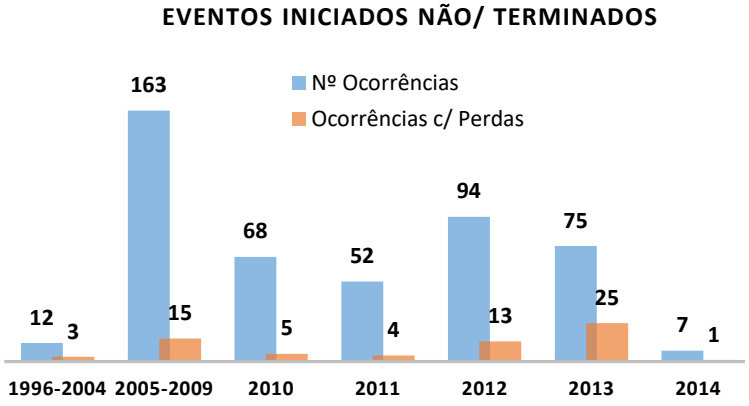


Gráfico 7 - Análise da variável Frequência

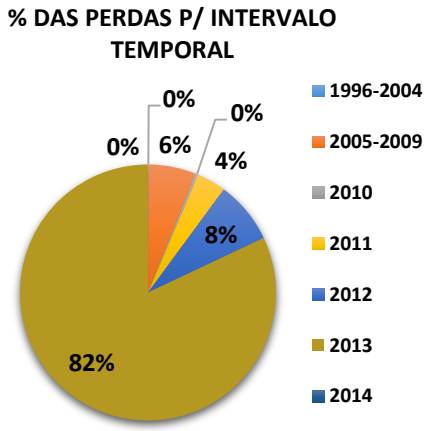


Gráfico 8 - Análise da variável Severidade

Fazendo uma leitura de ambos os gráficos é possível concluir: (i) o número de eventos de risco com prejuízos é muito baixo em todos os intervalos temporais, contudo observamos que 57% destes ocorrem entre 2012 e 2013 e que entre 2005 e 2009 ocorreram 22% destes eventos. A percentagem de eventos ocorridos com perdas, em relação ao total de eventos ocorridos em cada intervalo temporal, só uma vez foi superior aos 30%. (ii) quanto à severidade dos eventos, o ano de 2013 destaca-se, pois foi aqui que 82% do total de perdas ocorreram, seguido por 2012 com 8%. Sendo estes os anos onde ocorreram um maior número de eventos com perdas, 25 e 13 respectivamente, a diferença a nível de volume de perdas é por demais evidente entre estes dois anos.

4.1.5. Eventos Descobertos e Terminados

Os gráficos apresentados seguem a mesma lógica de outros apresentados anteriormente. A única diferença diz respeito aos intervalos temporais, onde, por falta de informação, acabou por se retirar os últimos dois intervalos temporais, aglomerando os eventos com maior diferença entre datas na classe "mais de 3 anos". Inicialmente iriam ser utilizados 1310 eventos, visto que eram os únicos que possuíam ambas as datas necessárias, mas após uma segunda análise, conclui-se que seria interessante analisar à parte, os eventos que só foram descobertos após terem sido terminados. Assim sendo, estes gráficos foram realizados utilizando 1056 registos (30%), sendo que destes, 635 (60%) apresentaram perdas para a instituição, ao passo que 421 (40%) não apresentaram.

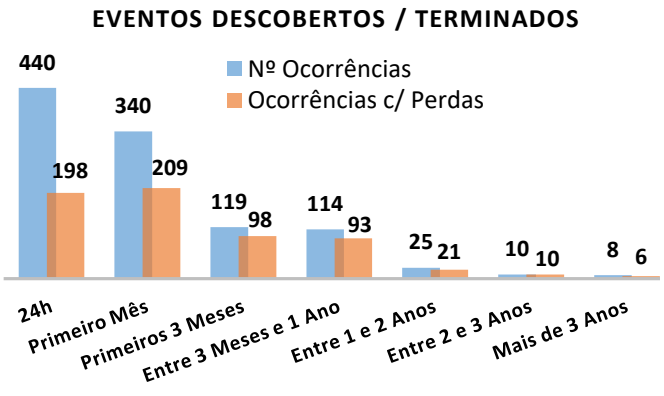


Gráfico 9 - Análise da variável Frequência

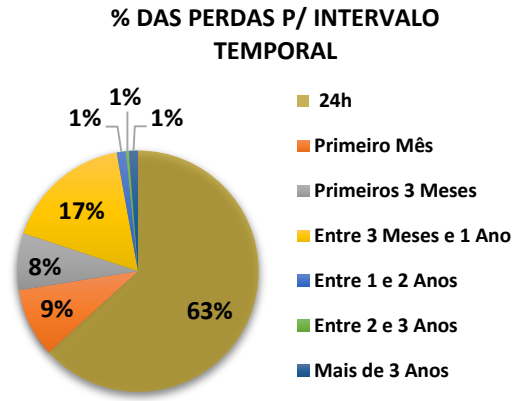


Gráfico 10 - Análise da variável Severidade

Fazendo uma leitura de ambos os gráficos é possível concluir: (i) 74% dos eventos ocorrem entre as primeiras 24 horas e o primeiro mês e, a esse valor correspondem 64% dos eventos com efectivas perdas para a instituição; (ii) os restantes intervalos apresentam valores superiores a 80% no que à relação evento ocorrido/evento com perda diz respeito, apesar de os seus números assumirem uma tendência descendente; (iii) em relação à *severidade* dos eventos, é possível observar que nas primeiras 24 horas ocorre um maior volume de perdas, 63%. Curiosamente o intervalo com maior número de eventos de perdas apenas possui 9% do volume total das mesmas, ao passo que o intervalo dos 3 meses a um ano com apenas 14% dos eventos prejudiciais para a instituição consegue atingir os 17%.

4.1.6. Eventos Descobertos após terem sido Terminados

Após uma segunda análise, considerou-se interessante analisar o comportamento dos eventos que somente são descobertos após terem sido terminados. Apenas 254 valores (7%) preenchem este critério, pelo que se optou por criar apenas um gráfico que ajude a observar a variação da variável *frequência*. Da mesma forma, a variável *severidade* fica fora desta análise, dado que com tão poucos registos as suas perdas acumuladas não ultrapassam os 400.000€, tornando-se assim pouco significativo. De referir que deste número de registos, 206 (81%) apresentam perdas para a instituição, ao passo que os restantes 48 (19%) nada apresentam.

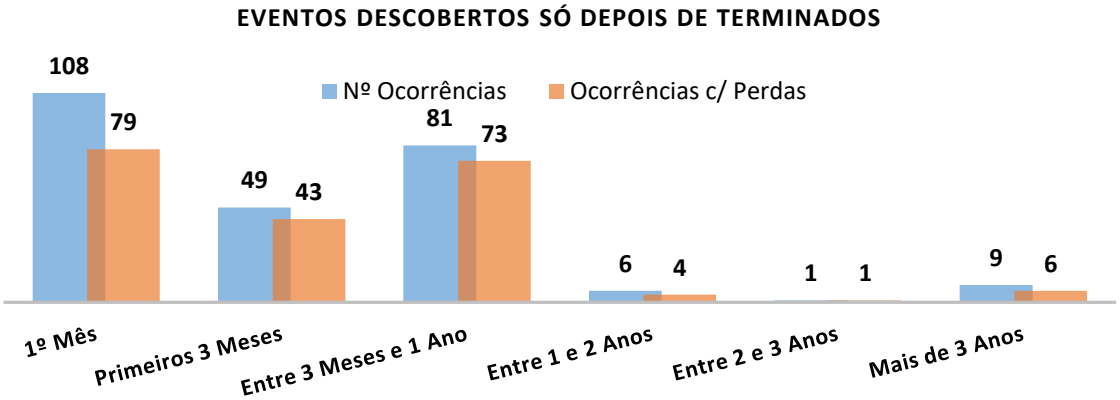


Gráfico 11 - Análise da variável Frequência

Fazendo uma leitura do gráfico em cima apresentado é possível concluir: (i) mais uma vez uma grande percentagem de eventos ocorrem entre o primeiro mês e o primeiro ano, neste caso 93%. Neste espaço temporal estão inseridos 94% dos eventos com efectivas perdas para a instituição. (ii) a mais interessante conclusão a reter é que existem eventos de risco, neste caso 16, que só foram descobertos após um ano, dois ou até mais de três anos após terem sido dados como terminados

4.1.7. Eventos Descobertos não Terminados

Os gráficos apresentados em baixo seguem a mesma lógica de outros apresentados anteriormente. Os intervalos temporais foram ajustados em função dos respectivos dados. Foram utilizados 1059 registos (30%) que preenchem os critérios exigidos, sendo que destes, apenas 106 registos (10%) correspondem a eventos com perdas para a instituição, ao passo os restantes 953 (90%) não apresentam.

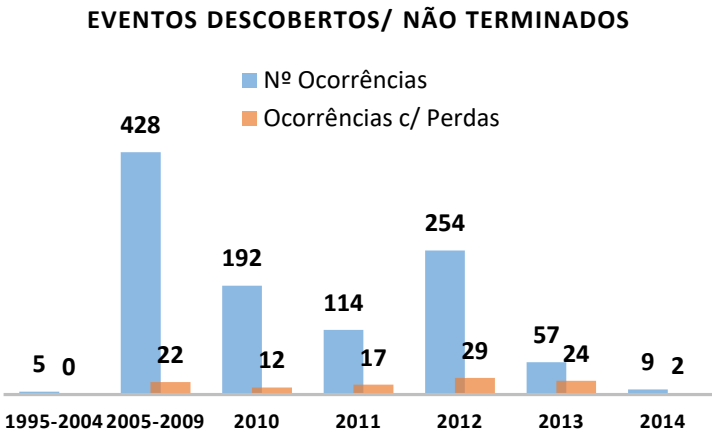


Gráfico 12 - Análise da variável Frequência

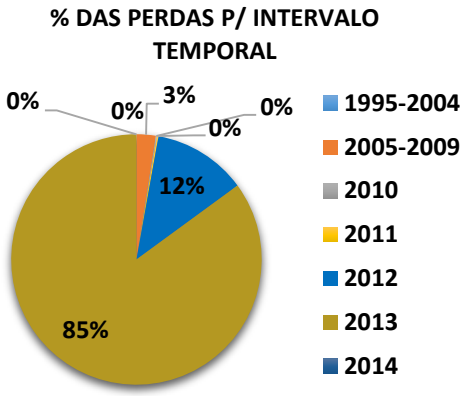


Gráfico 13 - Análise da variável Severidade

Fazendo uma leitura dos gráficos é possível concluir: (i) excluindo o ano de 2013 onde perto de metade dos eventos ocorridos correspondem a eventos com efectivas perdas, em todos os outros intervalos a proporção de eventos com perdas é pouco mais do que residual. Contudo o ano que apresenta maior número de eventos com prejuízo é 2012 e, se o somarmos com os valores de 2013 atingimos os 50% de eventos onde o prejuízo financeiro foi uma realidade; (ii) numa tendência que se tem repetido, os intervalos referidos, assim como o 2005-2009, são os intervalos com maior *frequência* e esta tendência também tem reflexo na *severidade* dos eventos ocorridos. Aliás nesse gráfico, apenas estas três categorias apresentam valores, compondo assim os 100% das efectivas perdas para a instituição. Destes, o destaque vai para 2013 que mesmo não sendo o ano com maior número de eventos, alberga 85% do total de perdas, ficando 12% no ano de 2012 e apenas 3 % entre 2005 e 2009.

4.1.8. Processos de Risco

A segunda fase da análise exploratória passou por analisar as variáveis *frequência* e *severidade* mas, ao contrário do que foi feito anteriormente, onde a componente temporal era o mais importante, nos gráficos que se seguem o foco recai essencialmente sobre os diferentes tipos de processos de risco e de categorias de risco que ajudam a explicar a natureza dos eventos ocorridos na instituição financeira. Na realização destes gráficos foram utilizados inicialmente todos os 3594 eventos de risco, ou seja, a totalidade da base de dados, embora em análises posteriores alguns ajustes tenham sido feitos como será mencionado de seguida.

Como foi referido previamente, dos 54 diferentes tipos de processos de risco apenas foram utilizados os 21 cuja frequência era superior a 10 eventos. Por fim, indicar que na realização do seguinte gráfico, mesmo utilizando apenas 21 processos, foi possível analisar 3453 eventos (96%), sendo que desses, 1986 (57%) apresentaram perdas para a instituição, ao passo que 1467 (43%), não apresentaram.

Como foi também mencionado previamente o gráfico referente à variável severidade, foi criado utilizando os 13 tipos de processos, do total de 54, cujas perdas acumuladas foram iguais ou superiores a 100.000€. Uma última nota para referir que, utilizando este critério, foram analisados 1861 eventos, o que corresponde a 93% do total de eventos com efectivas perdas.

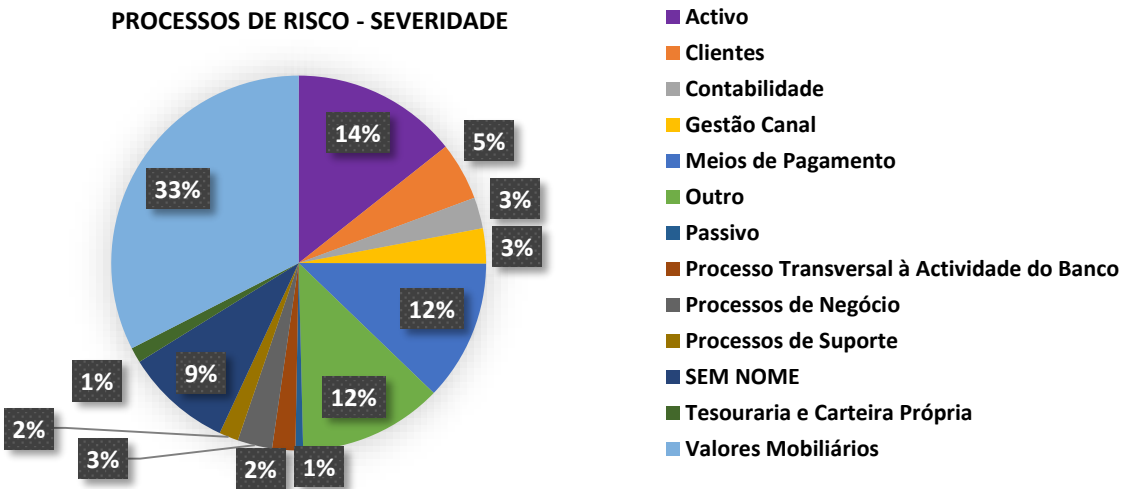


Gráfico 14 - Análise da variável *Severidade*

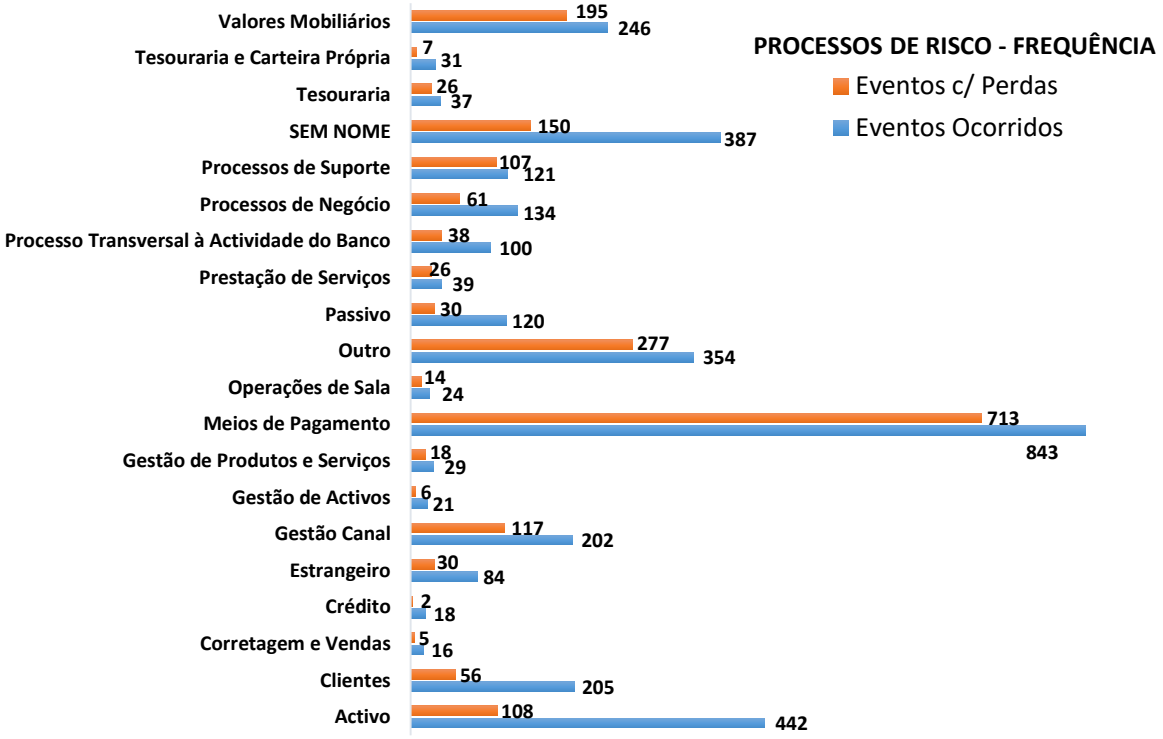


Gráfico 15 - Análise da variável *Frequência*

Fazendo uma leitura dos gráficos apresentados em baixo é possível concluir: (i) existem 7 tipos de processos de risco que juntos englobam 75% do total de eventos ocorridos. Sendo que estes eventos correspondem a 84% dos eventos com efectivas perdas (*frequência*). Por sua vez este último valor, a nível de *severidade*, corresponde a 85% das perdas para a instituição. Os processos em questão são: *activos, gestão canal, meios de pagamento, outro, processos de suporte, SEM NOME, valores mobiliários*. O processo *Meios de Pagamento* destaca-se claramente das restantes albergando quase 36% dos eventos com perdas (frequência); (ii) dos processos com maior volume de perdas para a instituição, 12 em 13 encontram-se também no gráfico de *frequência* o que ajuda a relacionar as duas variáveis. Contudo esta relação não pode ser vista como algo linear pois, se é verdade que de forma

geral, a processos com menor número de eventos de risco, com ou sem perdas, irão corresponder geralmente, menores percentagens de volume perdido, os gráficos apresentados ilustram alguns exemplos que ajudam a contrair esta lógica: ao observar os processos *Meios de Pagamento* e *Outro* concluímos que são os processos com maior número de eventos com efectivas perdas, 36% e 14% respectivamente. Contudo estes altos valores não têm qualquer tipo de correlação na variável *severidade*, onde estes valores correspondem apenas a 12% em cada um dos casos, sendo o exemplo do processo *Meios de Pagamento*, por demais evidente desta contradição. No sentido inverso, os processos *Activos* e *Valores Mobiliários* apresentam um número consideravelmente mais baixo de eventos com perdas, 5% e 10% respectivamente, contudo estes valores têm uma maior preponderância no total de perdas obtidas, atingindo no primeiro caso os 14% e no segundo uns surpreendentes 33%, sendo estes os processos que mais contribuíram para o total de perdas que a instituição obteve.

4.1.9. Categorias de Risco

Na continuação da segunda fase na análise exploratória, fomos analisar a variável *frequência* e *severidade* à luz das categorias de risco previamente escolhidas. Como foi mencionado previamente, na realização do gráfico de frequência, foram utilizadas apenas as 13 categorias que obrigatoriamente apresentavam mais de 10 eventos ocorridos, das 20 possíveis, independentemente destes se terem traduzido em prejuízos para a instituição ou não. Após esta restrição, o gráfico foi criado utilizando um total de 3540 eventos (98%), sendo que destes, 1978 (56%) correspondem a eventos com perdas para a instituição. Como foi também mencionado previamente, na realização do gráfico de severidade, foram utilizados apenas as 9 categorias cujas perdas acumuladas fossem iguais ou superiores a 100.000€. Desta forma a realização deste gráfico foi feita utilizando 1861 eventos o que corresponde a 94% do total de eventos com efectivas perdas.

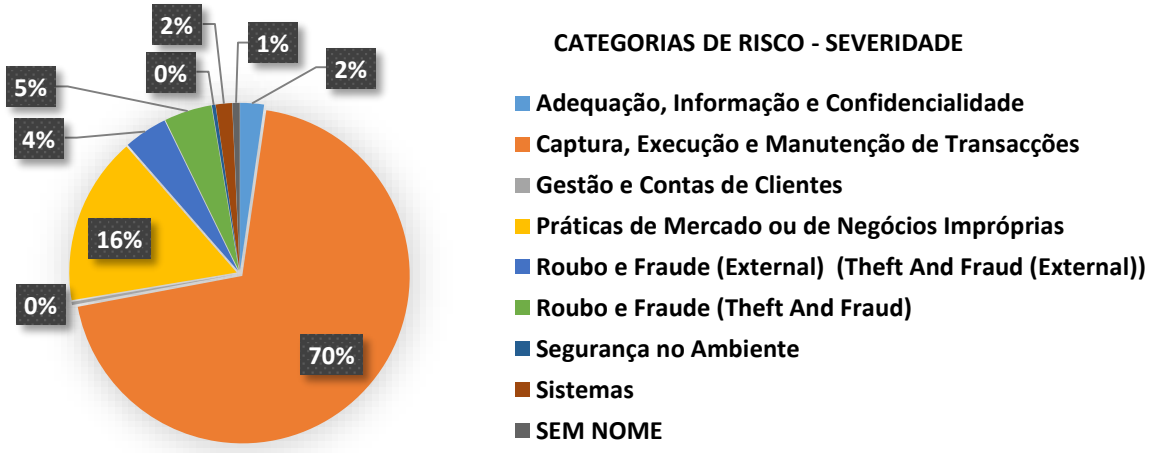


Gráfico 16 - Análise da variável Severidade

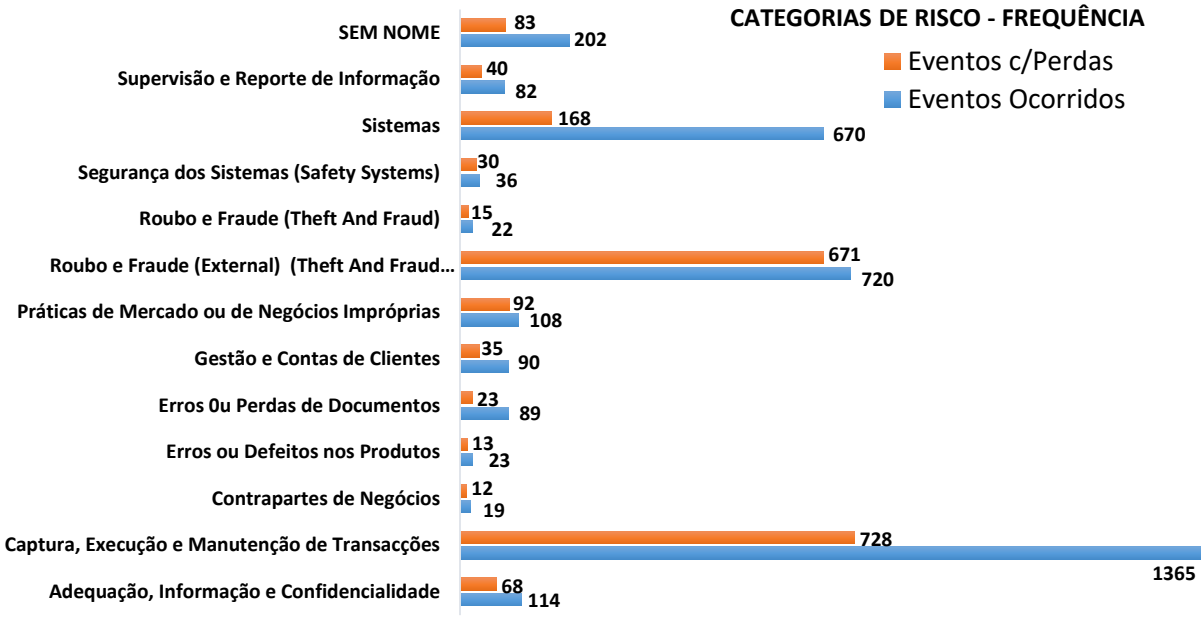


Gráfico 17 - Análise da variável Frequência

Fazendo uma leitura dos gráficos apresentados em baixo é possível concluir: (i) no que à frequência diz respeito, as categorias *Captura Execução e Manutenção de Transacções, Roubo e Fraude Externas e Sistemas* contêm 38%, 20% e 19% dos eventos de risco respectivamente, o que perfaz um total de 77% de eventos. Por outro lado estas são as categorias com um maior valor de eventos com efectivas perdas: 37%, 34% e 8% respectivamente, sendo que as restantes categorias apresentam valores inferiores. (ii) é possível observar que 8 das 9 categorias que entram na análise da *severidade* também se encontram na análise de *frequência*, numa realidade semelhante à que foi descrita nos resultados referentes aos processos de risco. (iii) na análise de *severidade* são rapidamente observáveis dois fenómenos: a categoria com maior número de eventos de risco com perdas, *Captura Execução e Manutenção de Transacções,* é a que se destaca também em volume percentual, fazendo os seus 37% de eventos corresponder a 70% do volume perdido. Contudo, a categoria que a precede em termos de número de eventos com perdas, *Roubo e Fraude Externas,* apenas faz corresponder os seus 34% de eventos a pouco mais de 4% em volume perdido. Seguindo a mesma lógica, a terceira categoria previamente mencionada com maior percentagem de eventos com perdas, apresenta um volume praticamente residual 2%. (iv) curioso observar que a categoria *Práticas de Mercado ou de Negócios Impróprias*, ganha destaque pela forma como atinge os 16% em volume percentual perdido com apenas 4% do total de eventos que tal originaram.

4.2. Análise da Informação Trabalhada

Como foi mencionado no capítulo 3.2 o segundo grande objectivo do processo de tratamento de dados foi o de criar uma nova tabela de dados. Esta tabela foi a nova fonte de informação aquando da utilização das técnicas de Data Mining e é composta por 2519 registos (70% da fonte de dados original). Por outro lado é composta por seis colunas, sendo que três transitam da base de dados inicial: identificador do evento, processos de risco e categorias de risco. As restantes três colunas foram criadas de raiz. A primeira, divide as perdas financeiras de cada evento em nove classes de acordo com os montantes envolvidos, sendo que, as perdas mais baixas correspondem à classe *L1*, ao passo que, perdas superiores a 1 milhão de euros correspondem à classe mais alta *L8.* A segunda coluna utiliza

novamente classes, neste caso oito, para identificar o número de dias que decorreram desde que um evento é iniciado até que este seja finalizado. A classe *D0* corresponde a eventos que foram iniciados e finalizados no mesmo dia, ao passo que a classe máxima *D7* engloba eventos cuja diferença de dias ultrapassa os 10 anos, ou seja, os 3600 dias. A última coluna foi criada com um duplo objectivo: em primeiro plano observar o número de dias que decorrem desde que um evento é iniciado até que este seja descoberto e, em segundo plano, observar se a sua descoberta é feita antes ou após a sua data de termo. Um evento que seja descoberto antes da sua data de termo irá pertencer a uma das classes do tipo *A*, se o contrário ocorrer, irá pertencer a uma das classes do tipo *B*.

4.2.1. Análise Descritiva da Informação Trabalhada

A análise descritiva apresentada foi desenvolvida através da criação de gráficos e de tabelas que desta forma: (i) irão contribuir para um melhor entendimento sobre a forma como se distribuem os eventos de risco em função das três colunas previamente mencionadas; (ii) irão permitir uma comparação entre as três variáveis criadas, ou seja, observa, por exemplo, para cada classe da variável referente aos montantes perdidos, quais são as classes das outras duas variáveis criadas que ocorrem com maior frequência, sendo desta forma possível caracterizar um determinado evento com base nestes três prismas; (iii) irão permitir a realização de uma análise que relacione os processos e as categorias de riscos com maior número de observações e as três colunas criadas. Ou seja, é possível observar, por exemplo, para um determinado tipo de processo de risco, qual o montante perdido na maior parte dos casos, ou por outro lado observar, na maioria dos casos, quantos dias demora um evento com determinada categoria de risco, desde que é iniciado até à sua data de termo.

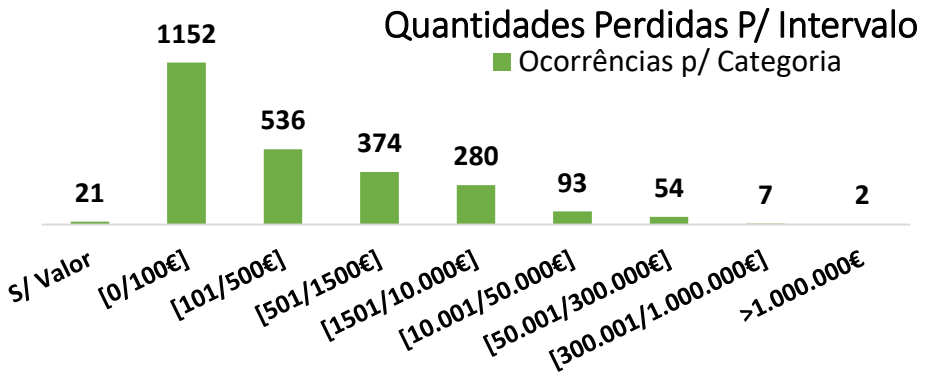


Gráfico 18 – Análise das quantidades perdidas por intervalos

O primeiro gráfico remete para a distribuição dos eventos de risco em função da coluna referente às perdas que cada evento teve para a instituição financeira. **Como é possível observar 93% das perdas estão compreendidas entre os 0 e os 10.000€**, sendo que, **desse valor percentual, 46% corresponde a valores de perdas compreendidos entre os 0 e os 100€**. Apenas existem dois eventos cujas perdas ultrapassam 1 milhão de euros.

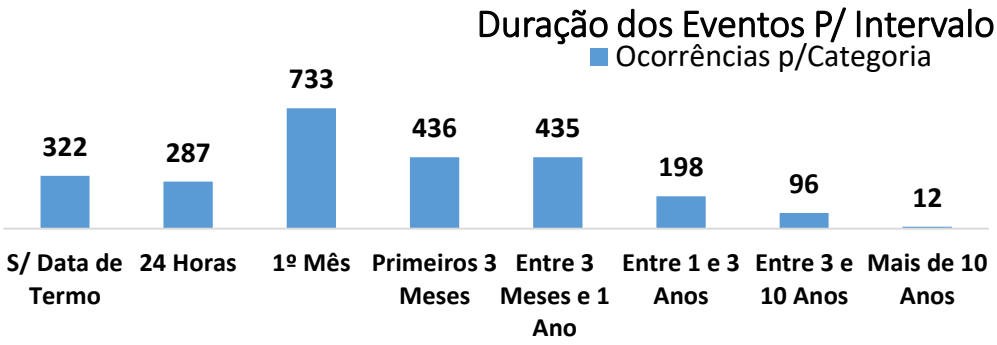


Gráfico 19 – Análise da duração dos eventos por intervalo

O segundo gráfico remete para a distribuição dos eventos de risco em função da coluna referente à diferença de dias que decorreram entre as suas datas de início e as suas datas de termo. Como é possível observar **75% dos eventos, desde que são iniciados, demoram entre 24 horas até um ano para serem finalizados**, sendo que, deste valor percentual, **praticamente 30% corresponde a eventos iniciados e finalizados no decorrer dos primeiros 30 dias**. Interessante observar a existência de 12 eventos que só foram terminados após mais de 10 anos de vida.

A tabela que se segue remete para a distribuição dos eventos em função do número de dias que passaram desde que o evento foi iniciado até que este foi descoberto. Se esse evento foi descoberto antes da sua data de termo irá ter um dos valores de *A*, caso contrário irá ter um dos valores de *B*. Uma nota para mencionar que a classe *B0* foi retirada desta análise e de todas as restantes análises até ao final desde capítulo.

Como foi referido anteriormente, houve a necessidade de criar a classe *B0,* com o propósito de agrupar os 1044 eventos (30% da base de dados original) sem data de descoberta, mas com data de termo. A data de descoberta atribuída a todos foi a de 1/1/2016, uma data superior a qualquer data de termo e como tal estes eventos foram automaticamente colocados na coluna *B*. Esta lógica não foi seguida para os eventos com data de descoberta mas sem data de termo, visto ser possível calcular as reais diferenças de dias entre as suas datas de inicio e de descoberta e assim agrupá-los em diferentes classes de acordo com o número de dias em questão, sendo este o principal objectivo desta nova coluna criada. Assim, não foi necessário criar uma classe *A0*. Por último referir que todos os eventos que foram iniciados e descobertos no mesmo dia, mesmo aqueles cuja data de termo corresponde a esse mesmo dia, foram colocados na classe *A1*. Esta é a razão pela qual não existem valores para *B1*, dado que estamos na presença de um período de apenas 24 horas.

Intervalos Temporais	Legenda	Observações
24 horas	A1	561
	B1	0
1º Mês	A2	303
	B2	62
1º Ano	A3	225
	B3	144
[1 - 3 Anos]	A4	88
	B4	9
[3 - 7 Anos]	A5	54
	B5	14
[7 - 10 Anos]	A6	3
	B6	1
> 10 Anos	A7	4
	B7	7

Tabela 6 – Análise da Detecção de um evento antes e após a sua data de termo

Como é possível observar **88% dos eventos são iniciados e descobertos durante o primeiro ano de existência, sendo distribuídos pelas três primeiras classes com os valores de 38%, 25% e 25% respectivamente.** Por outro lado é possível observar que a **maioria dos eventos são descobertos antes da sua data de termo, mais concretamente 84%, contra 16% de casos onde o contrário ocorre**. Por último mencionar que existe apenas uma classe cujo número de eventos descobertos após a sua data de termo é superior ao número de eventos descobertos antes dessa data, referimo-nos à classe cujo intervalo temporal é superior a dez anos.

Classe	Legenda	Classe	Legenda	Obs	Classe	Legenda	Obs
L0	S/ Valor	**D1**	24 Horas	**7**	**A1**	24 Horas	**13**
L1	[0/100€]	**D2**	1º Mês	**348**	**A1**	24 Horas	**436**
L2	[101/500€]	**D2**	1º Mês	**163**	**A3**	1º Ano	**74**
L3	[501/1500€]	**D4**	Entre 3 Meses e 1 Ano	**102**	**A2**	1º Mês	**47**
L4	[1501/10.000€]	**D3**	Primeiros 3 Meses	**75**	**A2**	1º Mês	**41**
L5	[10.001/50.000€]	**D2**	1º Mês	**35**	**A1**	24 Horas	**9**
L6	[50.001/300.000€]	**D2**	1º Mês	**32**	**A2**	1º Mês	**5**
L7	[300.001/1.000.000€]	**D2**	1º Mês	**3**	**A3**	1º Ano	**2**
L8	>1.000.000€	**D5**	Entre 1 e 3 Anos	**1**	**A4**	[1-3 Anos]	**1**

Tabela 7 – Comparação da coluna "Perdas" com as restantes colunas criadas

A *tabela 7* ilustra a relação entre a variável (azul) referente às *perdas* que cada evento teve para a instituição financeira e as outras duas variáveis criadas, nomeadamente a coluna referente à *duração* dos eventos (verde) e a coluna referente à *detecção* dos eventos de risco (laranja). Para cada classe da coluna azul foram observadas todas as classes das restantes duas variáveis com o objectivo de encontrar as classes a que correspondiam um maior número de observações. Podemos observar que 5 das 9 classes de *perdas* têm correspondência com a classe *D2* (1º Mês), ou seja, **independentemente dos montantes em *perda* estes ocorrem maioritariamente em eventos com a *duração* de um mês**. Referir que o montante mais elevado de *perdas* tem correspondência com a classe *D5* (entre 1 e 3 anos), ou seja, **as *perdas* superiores a 1 milhão ocorrem em eventos com uma *duração* compreendida entre 1 e 3 anos**. Por outro lado podemos observar que **8 das 9 classes de *perdas* têm correspondência com eventos que demoram entre 24 horas (*A1*) até 360 dias (*A3*) desde que são iniciados até que são *descobertos***. Referir também que nenhuma classe de *montantes perdidos* tem correspondência com qualquer evento da classe *B*, ou seja, *eventos descobertos* após a sua data de termo. Referir por último que mais uma vez **a classe de *montantes perdidos* mais elevada (*L8*) corresponde a uma classe de *detecção* alta *A4* (entre 1 e 3 anos).**

Classe	Legenda	Classe	Legenda	Obs	Classe	Legenda	Obs
D0	S/ Data de Termo	**L1**	[0/100€]	**293**	**A1**	24 Horas	**185**
D1	24 Horas	**L1**	[0/100€]	**207**	**A1**	24 Horas	**190**
D2	1º Mês	**L1**	[0/100€]	**348**	**A2**	1º Mês	**171**
D3	Primeiros 3 Meses	**L1**	[0/100€]	**127**	**A2**	1º Mês	**46**
D4	Entre 3 Meses e 1 Ano	**L2**	[101/500€]	**142**	**A3**	1º Ano	**108**
D5	Entre 1 Ano e 3 Anos	**L2**	[101/500€]	**61**	**A4**	[1-3 Anos]	**66**
D6	Entre 3 e 10 Anos	**L2**	[101/500€]	**30**	**A5**	[3-7 Anos]	**38**
D7	Mais de 10 Anos	**L4**	[1501/10.000€]	**6**	**B7**	> 10 Anos	**6**

Tabela 8 – Comparação da coluna "Duração" com as restantes colunas criadas

A *tabela 8* ilustra a relação entre a variável (verde) referente à *duração* de um evento de risco (diferença de dias entre a sua data de inicio e de termo) e as outras duas variáveis criadas, nomeadamente a coluna referente aos *montantes perdidos* por cada evento (azul) e a coluna referente à *detecção* dos eventos de risco (laranja). Para cada classe da coluna verde foram observadas todas as classes das restantes duas variáveis com o objectivo de encontrar as classes a que correspondiam um maior número de observações. Podemos observar que, às primeiras quatro classes de *duração* corresponde a classe mais baixa de *perdas*, *L1* (0/100€), assim como é possível observar que às seguintes três classes de *duração* corresponde a seguinte classe de perdas *L2* (101/500€). Apenas a classe mais elevada de *duração* tem correspondência com a classe *L4* (1501/10.000€). **Existe portanto, uma tendência que relaciona eventos com reduzidas *durações* a *perdas* igualmente pequenas e vise versa.** Por outro lado **é possível observar uma relação entre as classes de *duração* e as classes de *detecção*, ou seja, os intervalos temporais das classes de *detecção* coincidem com os das classes de *duração***, o que faz sentido, dado que a probabilidade de um evento ser iniciado e descoberto, dentro do período em que este é iniciado e terminado é maior do que o contrário. A maior surpresa surge na classe de duração *D7* (mais de 10 anos) à qual corresponde a classe *B7* (> 10 anos), ou seja, **aos eventos que têm um maior período de duração correspondem, com maior número de observações, os eventos que têm um maior período de descoberta, mas são todos descobertos após terem sido terminados** o que não deixa de ser assinalável.

Classe	Legenda	Classe	Legenda	Obs	Classe	Legenda	Obs
A1	24 Horas	D1	24 Horas	190	L1	[0/100€]	436
A2	1º Mês	D2	1º Mês	171	L1	[0/100€]	152
A3	1º Ano	D4	Entre 3 Meses e 1 Ano	108	L1	[0/100€]	108
A4	[1-3 Anos]	D5	Entre 1 Ano e 3 Anos	66	L1	[0/100€]	36
A5	[3-7 Anos]	D6	Entre 3 e 10 Anos	38	L2	[101/500€]	18
A6	[7-10 Anos]	D6	Entre 3 e 10 Anos	2	L2	[101/500€]	2
A7	> 10 Anos	D7	Mais de 10 Anos	2	L1	[0/100€]	2
B2	1º Mês	D2	1º Mês	31	L1	[0/100€]	31
B3	1º Ano	D2	1º Mês	64	L1	[0/100€]	73
B4	[1-3 Anos]	D4	Entre 3 Meses e 1 Ano	4	L1	[0/100€]	4
B5	[3-7 Anos]	D6	Entre 3 e 10 Anos	12	L4	[1501/10.000€]	5
B6	[7-10 Anos]	D6	Entre 3 e 10 Anos	1	L1	[0/100€]	1
B7	> 10 Anos	D7	Mais de 10 Anos	7	L4	[1501/10.000€]	4

Tabela 9 – Comparação da coluna “Detecção” com as restantes colunas criadas

A *tabela 9* ilustra a relação entre a variável (laranja) referente à *detecção* de um evento de risco (diferença entre a sua data de inicio e de descoberta) e, as outras duas variáveis criadas, nomeadamente a coluna referente à *duração* dos eventos de risco (verde) e a coluna referente aos *montantes perdidos* por cada evento (azul). Para cada classe da coluna laranja foram observadas todas as classes das restantes duas variáveis com o objectivo de encontrar as classes a que correspondiam um maior número de observações. Como foi referido anteriormente as colunas *B0* e *B1* foram retiradas da análise. Podemos observar que praticamente a qualquer classe de *detecção* corresponde uma classe de *duração* com igual intervalo temporal, ou seja, um qualquer evento iniciado e posteriormente descoberto, antes da sua data de termo, tem como classe correspondente com maior número de

observações, uma classe de duração com igual período temporal. **Podemos inferir uma relação de ordem entre um evento que primeiramente é iniciado, posteriormente descoberto e por fim finalizado, tudo dentro do mesmo período de tempo**. Observamos que as **classes *de detecção* têm correspondência com as classes mais baixas da variável referente aos *montantes perdidos*: cinco são do tipo *L1* (0/100€) e duas do tipo *L2* (101/500€)**, sendo que as últimas se encontram relacionadas com classes *A* com um período temporal mais alargado.

Relativamente aos eventos da classe *B*, ou seja, eventos *descobertos* após a sua data de termo, observamos novamente, uma relação de grande semelhança entre os seus intervalos temporais e os que caracterizam as classes da variável *duração*. As duas excepções ocorrem nas classes *B3* e *B4.* Relativamente aos *montantes perdidos*, a tendência para montantes baixos mantem-se, com as primeiras três categorias *B* a obterem um maior número de observações na classe *L1* (0/100€). Curioso observar que à classe B5 (3-7 anos) e B7 (>10 anos) correspondem classes mais elevados, nomeadamente *L4* (1501/10.000€).

A tabela 10 ilustra a relação entre, os processos e as categorias de risco com um maior número de observações e as três variáveis criadas. Desta forma é possível compreender o tipo de eventos que caracterizam maioritariamente cada processo ou categoria à luz das três variáveis criadas. Foi acrescentado o processo "Valores Mobiliários" e a categoria "Práticas de Mercado ou de Negócios Impróprias" pela forma como a sua relação com as variáveis criadas contrasta com as restantes, apesar de terem um número de observações relativamente menor. **Analisando os *processos de risco* podemos observar que os três processos com maior número de observações são constituídos maioritariamente por eventos responsáveis por *perdas* de reduzidos montantes, *L1* e *L2*, sendo que a estas correspondem classes de *detecção* compreendias entre *A1* e *A3*, também elas baixas. Por sua vez, as classes de *duração* variam entre *D0*, *D2* e *D4*.** Curioso observar que **o processo "Meios de Pagamento" é constituído maioritariamente por eventos *descobertos* durante o primeiro mês** (*A2*), contudo, estes **chegam a demorar um ano até que sejam *terminados*** (*D4*). **Contrariamente a esta tendência, o processo com menos observações, "Valores Mobiliários", é constituído na sua maioria por eventos responsáveis por *perdas* elevadas, classe *L5*.** Por outro lado é relevante identificar que este tipo de processo de risco **é composto por**

	Nome	Obs	Classe	Legenda	Valor %
Processos	Activo	241 Obs	**L1**	[0 - 100€]	**60%**
			D0	S/Data de Termo	**33%**
			A1	24 Horas	**45%**
	Meios de Pagamento	739 Obs	**L1**	[0 - 100€]	**35%**
			D4	Entre 3 Meses e 1 Ano	**29%**
			A2	1º Mês	**30%**
	Outro	292 Obs	**L2**	[101 - 500€]	**50%**
			D2	1º Mês	**36%**
			A3	1º Ano	**34%**
	Valores Mobiliários	218 Obs	**L5**	[10.001 - 50.000€]	**24%**
			D2	1º Mês	**39%**
			A5	[3-7 Anos]	**30%**
Categorias	Captura, execução e manutenção de transações	943 Obs	**L1**	[0 - 100€]	**47%**
			D2	1º Mês	**35%**
			A1	24 Horas	**26%**
	Roubo e Fraude (Externa)	688 Obs	**L3**	[501 - 1500€]	**29%**
			D4	Entre 3 Meses e 1 Ano	**31%**
			A2	1º Mês	**28%**
	Sistemas	444 Obs	**L1**	[0 - 100€]	**78%**
			D1	24 Horas	**27%**
			A1	24 Horas	**52%**
	Práticas de Mercado ou de Negócios Impróprias	99 Obs	**L5**	[10.001 - 50.000€]	**37%**
			D2	1º Mês	**67%**
			A1	24 Horas	**60%**

Tabela 10 - Comparação de Processos e Categorias com as restantes colunas criadas

eventos que, apesar de serem *terminados* dentro do primeiro mês, classe *D2*, **chegam a demorar sete anos até serem *descobertos*.**

Relativamente às ***categorias* de risco,** podemos observar que **duas das três categorias com maior número de observações são maioritariamente constituídas por eventos que causaram pequenas *perdas* à instituição**, classe *L1*, **assim como, por eventos com baixas classes de *duração***, *D2* e *D1*, e também **baixas classes de *descoberta*** (*A1*). Contudo, existem duas categorias que apresentam uma **tendência diferente**: a classe "Roubo e Fraude (Externa)" é constituída maioritariamente por eventos, cujas *perdas* remetem para a classe *L3*, assim como, por eventos que apesar de demorarem um mês até serem *descobertos*, na sua maioria chegam a demorar até um ano para serem *finalizados* (*D4*). **A *categoria* com menos observações, mais uma vez, atinge a classe de *perdas* mais elevada *L5*, sendo constituída maioritariamente por eventos com baixos períodos temporais de *descoberta* e de *duração*, *A1* e *D2* respectivamente.**

4.3. Análise de *Clusters*

Este capítulo irá abordar em detalhe a análise de *Clusters* efectuada aos dados de risco operacional, anteriormente trabalhados e tratados. A fonte de dados utilizada é composta por 1475 registos, cada um caracterizado por seis variáveis distintas, amplamente descritas no capítulo anterior: *identificador*, tipo de *processo*, tipo de *categoria, detecção, duração* e montantes *perdidos*. A *Imagem 3* confirma que todas as seis variáveis foram descarregadas para o diagrama, sendo que todas são consideradas variáveis de *input* com a excepção da variável responsável por identificar cada evento de risco, classificada como *ID*. As restantes são variáveis de classe, do tipo nominal, na medida em que são variáveis não numéricas. Como é possível, observar todas as variáveis foram utilizadas para a realização da análise de *Clusters* com excepção do identificador dos eventos.

Name	Use	Report	Role	Level
Categoria_risk	**Yes**	**No**	Input	Nominal
Deteccao	**Yes**	**No**	Input	Nominal
Duracao	**Yes**	**No**	Input	Nominal
ID	**No**	**No**	ID	Nominal
Loss_AMT	**Yes**	**No**	Input	Nominal
Processos_Risk	**Yes**	**No**	Input	Nominal

Imagem 3 – Variáveis utilizadas na análise de *clusters* pelo *software Miner*

NumClusters	AvgErro
1	2898.013
2	2782.991
3	2712.126
4	**2662.334**
5	2519.159
6	2521.51
7	2423.227
8	2462.44
9	2353.297
10	2360.305

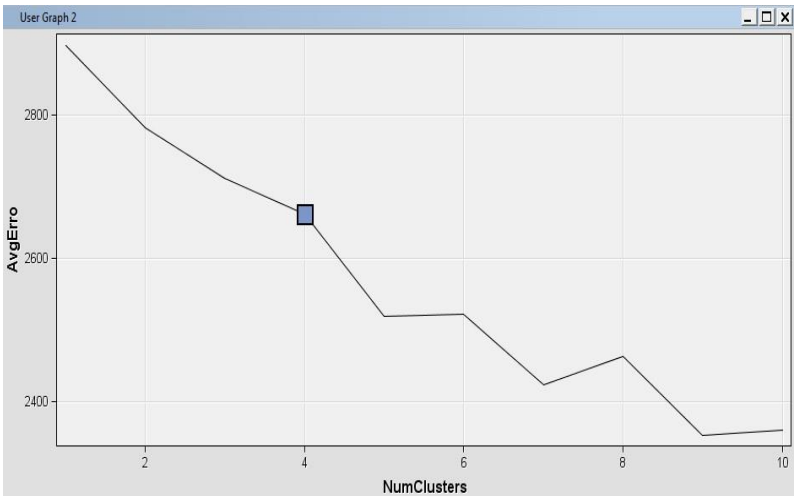


Imagem 4 - Escolha do número de *Clusters (Software SAS Miner)*

O *software SAS Miner* utiliza o processo de partição *K-means* para a realização do *clustering*. Este processo é rápido e funciona de forma eficiente com quantidades de dados relativamente altas. Ao mesmo tempo é bastante utilizado para gerar *Clusters* com formas simples, principalmente quando o número de iterações a realizar é significativamente reduzido. Para se conseguir identificar o número de *Clusters* ideal, utilizou-se, como referido anteriormente, o nó *SASCode* cujo código se encontra no *Anexo 2*, para criar o *Gráfico do Cotovelo*. Este gráfico (*Imagem 4*) tem por objectivo ilustrar a evolução da variabilidade intra-cluster, isto é, a soma das distâncias de todos os objectos ao centróide do seu *cluster*. **Sabendo que, por um lado, uma menor variabilidade intra-cluster é sinónimo de uma maior coesão do *cluster*, assim como que, à medida que o número de *k Clusters* aumenta essa mesma variabilidade tende a decrescer, o objectivo a que nos propormos passa por encontrar no gráfico o ponto onde o declive da curva seja menos acentuado**. Por outras palavras, **encontrar o ponto onde o ganho de coesão dos *Clusters* deixe de justificar a criação de um *cluster* adicional**. Dito isto, executou-se o processo *k-means* para valores de *k*, tal que, *k=1,…,10,* construindo o gráfico apresentado em cima. **Tendo em conta o objectivo anteriormente mencionado e, observando quer o gráfico do cotovelo, quer a tabela com os valores exactos da variabilidade intra-cluster, concluiu-se que o número ideal de *Clusters* seria 4**. Esta solução irá garantir que cada *cluster* será o mais coeso possível (mantendo a maior variabilidade intra-cluster possível), não construindo conjuntos em demasia, o que iria prejudicar a sua própria coesão interna

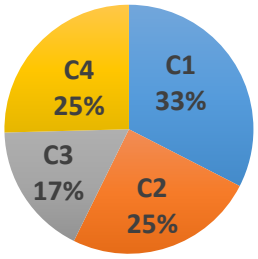


Gráfico 20 - Distribuição dos eventos por *cluster*

De seguida iremos apresentar os resultados da análise de *Clusters*. Como foi referido anteriormente foram utilizados quatro *Clusters*. O *Gráfico 20* ilustra a distribuição percentual dos registos pelos diferentes grupos criados. De referir que o primeiro *cluster* albergou 480 registos, o segundo 365, o terceiro 255 e o quatro 375. A solução parece bastante aceitável também a nível da distribuição dos dados de *input.*

Após observar a distribuição dos registos pelos diferentes *Clusters* e, antes de começar a decompor cada grupo de acordo com as suas variáveis dominantes, fomos observar uma aplicação particular do S*oftware Miner* que resulta da aplicação de uma técnica específica à matriz de distâncias entre *Clusters*, o *Multidimensional Scalling.* Esta técnica, ilustrada na *Imagem 5*, tem como principal objectivo apresentar as distâncias e as proximidades entre cada *cluster* num espaço bidimensional. Sendo uma técnica de redução de dimensionalidade, tenta através de um processos iterativo reduzir um elevado número de variáveis apenas a duas dimensões, ao mesmo tempo que preserva a hierarquia das distâncias entre as médias de cada *cluster.* Como é possível observar, o segundo *cluster* apresenta um claro distanciamento relativamente aos restantes três. Por outras palavras, existe uma variável que domina de forma inequívoca o segundo *cluster* fazendo com que este se distancie dos restantes. Mais à frente veio-se a comprovar que essa era a variável referente à *duração* dos eventos de risco.

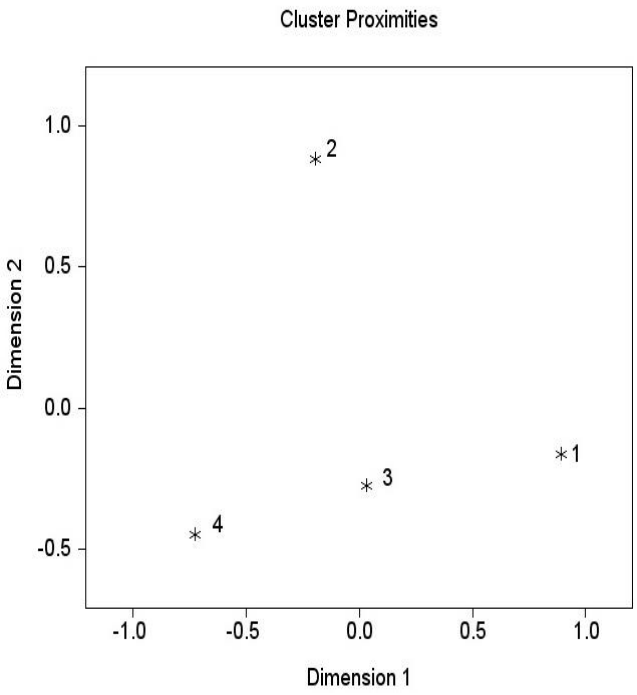


Imagem 5 - Visualização das distâncias entre os *clusters* *(Software SAS Miner)*

Para reforçar esta conclusão fomos observar dois *outputs* do S*oftware Miner* que se encontram relacionados: (i) fomos analisar a representação dos *Clusters,* sob a forma de uma estrutura hierárquica, em árvore, onde é possível identificar as percentagens e os números de observações atribuídos a cada *cluster*, assim como os valores limite de cada variável de *input*. Desta forma é possível identificar qual das variáveis de *input* teve um maior peso, aquando do agrupamento das observações pelos diferentes *Clusters*. Esta representação encontra-se completa no *Anexo 3*, contudo, decidimos apresentar na *Imagem 6*, **a ramificação que ilustra o referido afastamento entre o *cluster* 2** e os restantes; (ii) de seguida fomos observar de seguida a tabela que, não só apresenta as variáveis responsáveis pela elaboração de cada *cluster*, como ainda a importância relativa, numa escala de 0 a 1, que cada uma teve nessa mesma construção. Esta encontra-se na *Imagem 7*. Analisando ambas as imagens é possível concluir que a variável referente à duração de um evento de risco teve uma enorme importância na construção dos *Clusters*, sendo esta a responsável pelo distanciamento entre o *cluster* 2 e os restantes.

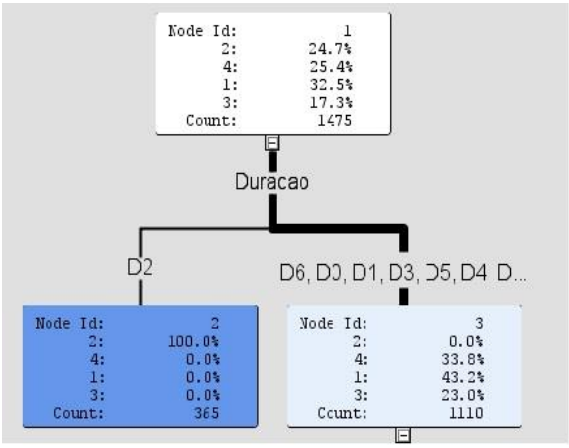


Imagem 6 - Perfil *dos clusters* em árvore *(Software SAS Miner)*

Variable Name	Label	Number of Splitting Rules	Importance
Duracao	Duracao	6.0	1.0
Deteccao	Deteccao	3.0	0.9348
Processos_Risk	Processos_Risk	0.0	0.7396
Loss_AMT	Loss_AMT	1.0	0.5781
Categoria_risk	Categoria_risk	2.0	0.5031

Tabela 11 - Importância das variáveis para a construção dos *clusters*

De seguida iremos apresentar uma descrição de cada um dos quatro *Clusters.* O *Anexo 4* apresenta um perfil de cada grupo, onde é possível observar a preponderância que cada classe de uma determinada variável teve na elaboração de um certo *cluster*. De forma a separar a análise, foram criadas tabelas para cada grupo de modo a permitir um melhor entendimento da sua análise.

A *Tabela 12* apresenta a composição do primeiro *cluster* em função das suas variáveis caracterizadoras. De referir que apenas os valores percentuais mais altos foram incluídos na tabela. Como é possível observar, este *cluster* juntou quase na sua totalidade eventos com as mais baixas *perdas* financeiras possíveis para a instituição, eventos esses que também tiveram uma *duração* bastante baixa. Curiosa a inclusão dos eventos sem data de *duração*[4]. A tendência mantem-se a nível da *detecção* dos eventos, em que mais de 70% dos registos reunidos foram *descobertos* nas primeiras 24 horas. No que aos *processos* de risco diz respeito, estes encontram-se muito repartidos no *cluster* em análise, o que é possível observar pelos valores percentuais apresentados, podendo-se apenas destacar: (i) *activo*. Este *processo* diz respeito à forma como a instituição financeira gere todos os seus activos, desde os seus imóveis aos seus investimentos; (ii) *processos de negócio*. Este *processo* diz respeito a todas as actividades normais de um banco, desde a abertura de contas à concessão de créditos; (iii) *clientes*. Este *processo* diz respeito à relação do banco com os dados dos seus clientes, desde a forma como organiza os seus contactos à forma como utiliza os dados referentes a futuras reclamações, ou ao futuro interesse, ou não, numa nova campanha. Relativamente às *categorias* de risco, praticamente metade dos registos que o *cluster* agregou dizem respeito aos *sistemas*: esta *categoria* corresponde a tudo o que tem a ver com a tecnologia que a instituição financeira utiliza, desde os seus serviços informáticos e bases de dados até ao *homebaking.*

Cluster 1			
Variáveis	**Classes**	**Valor Estatistico**	**%**
Loss_AMT	**L1** [0/100€]	0.9375	**94%**
Duração	**D0** (S/ Data de Termo)	0.6562	**66%**
	D1 (24 Horas)	0.3252	33%
Detecção	**A1** (24 Horas)	0.7292	**73%**
	A3 (1º Ano)	0.1208	12%
	A2 (1º Mês)	0.0854	9%
Processos_Risk	Activo	0.1896	**19%**
	Processos De Negócio	0.1146	**11.4%**
	Clientes	0.0958	**10%**
	Outro	0.0749	8%
	Meios De Pagamento	0.0708	7%
Categoria_risk	Sistemas	0.4667	**47%**
	Captura, execução e manutenção de transacções	0.3104	31%

Tabela 12 - Caracterização do primeiro *cluster*

[4] Eventos sem data de fim, mas incluídos à mesma na análise, por respeitarem de forma correcta o preenchimento de todas as outras variáveis.

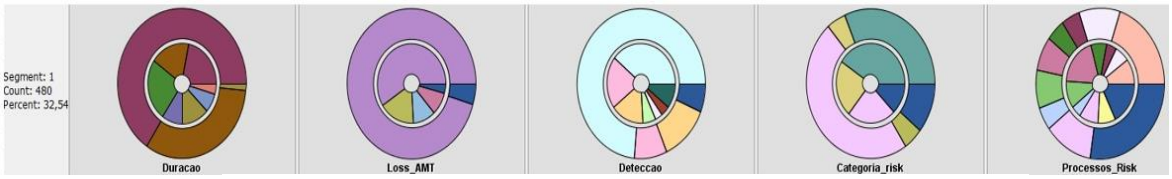


Imagem 7 – *Segment Profil*e do primeiro *cluster (Software SAS Miner)*

A *imagem 7* apresenta um *output* que ajuda a complementar a análise anterior: (i) todas as cinco variáveis foram utilizadas na construção do primeiro *cluster*; (ii) o círculo interno apresenta a distribuição normal de cada variável em função das suas classes, ao passo que o círculo externo apresenta a importância que cada classe teve para a construção do *cluster* em análise; (iii) é possível observar a preponderância que as classes de *duração D0* e *D1* tiveram, assim como a classe de *montantes perdidos L1* e a classe de *duração A1*, todas bastante acima da sua distribuição normal; (iv) no mesmo sentido é possível observar que as variáveis referentes às *categorias* e aos *processos* de risco se encontram bastante fragmentados, ao ponto de só aparecerem no círculo interno as classes com maior frequência. Contudo, podemos validar a preponderância da *categoria sistemas* e do *processo activo*. Curioso ainda observar na variável *processos* o segmento azul do círculo externo. Este identifica várias classes de *processos*, que sozinhas têm muito pouca expressão, mas que juntas, têm uma preponderância considerável na formação do *cluster,* numa *categoria* à qual o *software* apelidou de "*others*". Por último, a *imagem 8* ilustra a distribuição das observações mais relevantes aquando da formação do primeiro *cluster*.

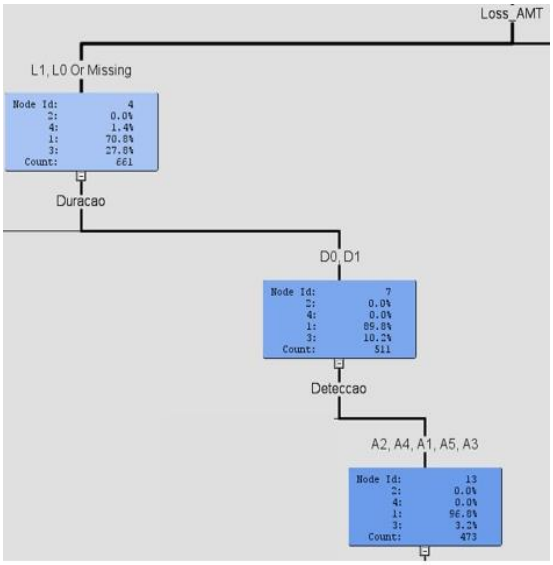


Imagem 8 - Perfil *dos clusters* em árvore *(Software SAS Miner)*

Em suma, o primeiro *cluster* é dominado pelas classes de *perdas*, de *duração* e de *detecção* mais baixas, às quais de juntam os *processos* de risco (*Activo*, *Processos de Negócio e Clientes*), assim como a *categoria* de risco (*Sistemas*).

A *Tabela 13* apresenta a composição do segundo *cluster*. Como é possível observar, existem algumas diferenças em relação ao *cluster* anterior: apesar de as *perdas* de classe mais baixa continuarem a dominar, o seu valor percentual foi reduzido em 33% e, como consequência, novas classes de *perdas* mais elevadas surgem agregadas a este *cluster.* Uma segunda nota que diz respeito ao **poder da classe *D2* que domina totalmente o *cluster*, como aliás, já tinha sido demonstrado anteriormente**. Contudo não deixa de representar um aumento temporal relativamente à mesma variável no *cluster* anterior. O mesmo aumento temporal acontece na variável referente à *detecção* dos eventos. Curiosa a aparição de eventos com classe de *duração B3*[5]. Respectivamente aos *processos* de risco, existe novamente uma grande fragmentação das classes, sendo possível destacar três: (i) *processos de suporte*. Este *processo* diz respeito a todas as componentes capazes de auxiliar as

[5] Eventos que só foram dados como descobertos após terem sido finalizados.

instituições financeiras nas áreas da informática, logística ou recursos humanos; (ii) *meios de pagamento*. Este *processo* diz respeito aos riscos das operações comerciais e dos pagamentos diferidos desde cheques sem cobertura até créditos mal parados; (iii) *estrangeiro*. Este *processo* diz respeito aos riscos associados com todas as transacções efectuadas com bancos estrangeiros. Relativamente às *categorias* de risco praticamente metade dos registos que o *cluster* agregou dizem respeito à *captura, execução e manutenção de transacções*. Esta *categoria* agrega os diferentes problemas que uma transacção pode sofrer durante estas três etapas.

Cluster 2			
Variáveis	**Classes**	**Valor Estatistico**	**%**
Loss_AMT	**L1** [0/100€]	0.6109	**61%**
	L2 [101/500€]	0.1699	17%
	L3 [501/1500€]	0.1041	10%
Duração	**D2** (1º Mês)	1	**100%**
Detecção	**A2** (1º Mês)	0.4684	**47%**
	A1 (24 Horas)	0.2712	**27%**
	B3 (1º Ano, após data de termo)	0.1753	17.5%
Processos_Risk	Processos de Suporte	0.1781	**18%**
	Meios de Pagamento	0.1260	**13%**
	Estrangeiro	0.1096	**11%**
	Processos de Negócio	0.0932	9%
Categoria_risk	Captura, execução e manutenção de transacções	0.4658	**47%**
	Roubo E Fraude (Externa) (Theft And Fraud (External)	0.2192	22%
	Sistemas	0.1479	15%

Tabela 13 – Caracterização do segundo *cluster*

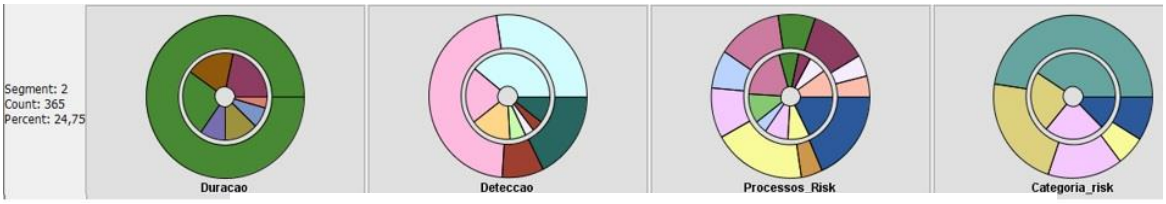


Imagem 9 – *Segment Profil*e do segundo *cluster (Software SAS Miner)*

Através da *Imagem 9* podemos observar as características que definem o segundo *cluster:* (i) observamos que apenas quatro variáveis tiveram real preponderância na construção do cluster, ou seja, foram agregados registos com base em todas as variáveis menos a que diz respeito aos montantes em *perda*. (ii) **validamos o argumento de que a variável *duração* e a sua classe *D2* foi determinante na elaboração do grupo e no agregar dos seus registos**; (iii) observamos um aumento da classe de *detecção A2* em detrimento da *A1*, assim como o aparecimento da classe *B3*; (iv) continuamos a observar uma fragmentação das variáveis *processos* e *categorias*, contudo validamos a preponderância da *categoria captura, execução e manutenção de transacções*, assim como das classes de *processo Processo de Suporte* e *Meios de Pagamento*. Novamente, as classes de *processos* de risco mais baixas, classes com pouca significância individualmente, foram aglomeradas pelo *software* na classe "*Others*" ganhando desta forma uma maior expressão (secção a azul).

Em suma, o segundo *cluster* diferencia-se do primeiro por incorporar classes de *perdas*, de *duração* e de *detecção* um pouco superiores. Apesar de no caso dos montantes *perdidos* o domínio da classe *L1*, ainda ser bastante prenunciado. Diferencia-se também pelo claro domínio da variável *duração*, nomeadamente a classe *D2* (1º Mês), pela importância que teve aquando da agregação dos registos que constituem este *cluster.* A estas classes associam-se os *processos* de risco (*Processos de Suporte*, *Meios de Pagamento* e *Estrangeiro*), assim como a *categoria* de risco (*Captura, Execução e Manutenção de Transacções*)

A *Tabela 14* apresenta a composição do terceiro *cluster*. Como é possível verificar este *cluster* continua a agregar bastantes registos com baixas classes de perda, tendo o seu valor percentual aumentado 11% em relação ao grupo anterior. Contudo, as outras variáveis apresentam algumas alterações significativas: no que à *duração* diz respeito é possível observar que as classes *D3* e *D4* juntas ultrapassam os 40% dos registos agregados, ao passo que os registos *descobertos* após a sua data de termo, representados pela classe de *detecção B3* ultrapassam os 31%. Relativamente à variável *processos de risco,* mantem-se a tendência para uma grande fragmentação, podendo apenas destacar-se as classes já descritas *Processos de Negócio* e *Meios de Pagamento*. Relativamente à variável *categorias de risco* podemos observar que a classe *Roubo e Fraude (Externa)* teve uma subida de 15%, ao passo que a classe *Captura, Execução e Manutenção de Transacções* teve uma descida de 16%, sendo estas as duas classes mais relevantes desta variável.

Cluster 3			
Variáveis	**Classes**	**Valor Estatistico**	**%**
Loss_AMT	**L1** [0/100€]	0.7216	**72%**
	L3 [501/1500€]	0.1137	11%
	L2 [101/500€]	0.0863	9%
Duração	**D1** (24 Horas)	0.44314	**44%**
	D3 (Primeiros 3 Meses)	0.22745	23%
	D4 (Entre 3 Meses e 1 Ano)	0.18824	19%
Detecção	**B3** (1º Ano, após data de termo)	0.30196	**31%**
	A1 (24 Horas)	0.20784	21%
	A3 (1º Ano)	0.20392	20%
Processos_Risk	Meios Pagamento	0.22745	**23%**
	Processos de Negócio	0.14118	**14%**
	Gestão Canal	0.10980	11%
	Processos de Suporte	0.10980	11%
Categoria_risk	Roubo E Fraude (Externa) (Theft And Fraud (External)	0.36863	**37%**
	Captura, execução e manutenção de transacções	0.28628	**29%**
	Sistemas	0.12549	13%

Tabela 14 – Caracterização do terceiro *cluster*

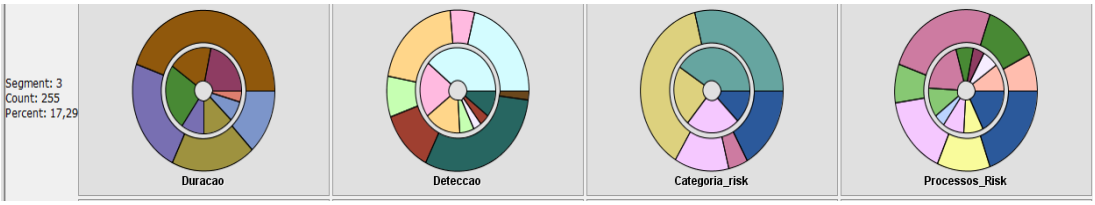


Imagem 10 – *Segment Profile* do terceiro *cluster (Software SAS Miner*)

Através da *Imagem 10* podemos validar a caracterização do *cluster:* (i) observamos que o grupo aglomerou objectos com base em apenas quatro das cinco variáveis, tendo voltado a ficar de fora a variável referente aos montantes *perdidos*, como já tinha ocorrido no segundo *cluster*; (ii) validamos o crescimento de novas classes de *duração* (*D3* e *D4*) e de *detecção* (*B3* e *A3*), contrariando a normal distribuição das variáveis (circulo interno); novamente contrariando a tendência da variável referente às *categorias de risco*, a variável *Roubo e Fraude (Externa)* ganha uma grande preponderância neste *cluster* ao passo que na variável referente aos *Processos de Risco* é a classe *Processos de Negócio* a que de forma mais notória contraria a média da variável. Por último, a *imagem 11* ilustra a distribuição dos registos da classe *B3* para o terceiro *cluster.*

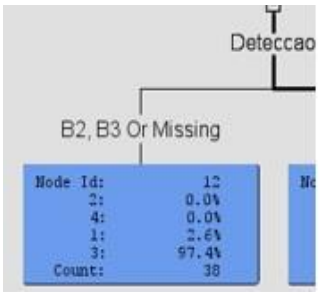


Imagem 11 - Perfil dos *cluster*s em árvore (*Software SAS Miner*)

Em suma, o terceiro é caracterizado por eventos com classes de perdas baixas. Continua o domínio da classe *L1*. Contudo, estes eventos apresentam as classes de *duração* mais altas até ao momento, podendo chegar a demorar um ano até que sejam *descobertos*. A classe de *detecção* mais relevante é a *B3*, logo estamos na presença de eventos que só são *descobertos* após terem sido *terminados*. Não deixa de ser interessante analisar as classes de *detecção* e de *duração* em conjunto. Quanto aos processos, destacam-se os *Meios de Pagamento* e os *Processos de Negócio*. Quando às *categorias,* destacam-se o *Roubo e Fraude (Externa)* e a *Captura, Execução e Manutenção de Transacções*.

A *Tabela 15* apresenta a composição do quarto *cluster*. Este último grupo apresenta bastantes diferenças relativamente aos anteriores: (i) constatamos que a classe de perdas *L1* deixou de ter significância para a construção do grupo e, no seu lugar surgem classes de *perdas* mais elevadas como *L3* e o *L4*, que juntos perfazem praticamente 50% dos registos aglomerados; (ii) estes registos apresentam os períodos temporais mais elevados de qualquer *cluster* podendo demorar até 10 anos para serem *descobertos*; (iii) os períodos temporais de *detecção* também apresentam um aumento, nomeadamente a classe *A3* que ultrapassa os 30% de registos aglomerados. Por outro lado surge pela primeira vez a classe *A4*; (iv) relativamente aos *processos de risco*, as duas classes que se destacam são a classe *Outros* e a classe *Meios de Pagamento*; relativamente às *categorias de* risco, observamos que a classe predominante é a *Captura, Execução e Manutenção de Transacções* com 47% dos registos aglomerados pelo *cluster.*

Cluster 4			
Variáveis	**Classes**	**Valor Estatistico**	**%**
Loss_AMT	**L2** [101/500€]	0.4159	**42%**
	L3 [501/1500€]	0.2427	24%
	L4 [1.501/10.000€]	0.2267	23%
Duração	**D4** (Entre 3 Meses e 1 Ano)	0.3707	**37%**
	D3 (Primeiros 3 Meses)	0.2213	**22%**
	D5 (Entre 1 e 3 Meses)	0.2	**20%**
	D6 (Entre 3 e 10 Anos)	0.1653	17%
Detecção	**A3** (1º Ano)	0.3067	**31%**
	A2 (1º Mês)	0.2053	**21%**
	A1 (24 Horas)	0.1573	16%
	A4 [1-3 Anos]	0.1413	14%
Processos_Risk	Meios de Pagamento	0.3359	**34%**
	Outro	0.2347	**23%**
	Gestão Canal	0.0773	8%
Categoria_risk	Captura, execução e manutenção de transacções	0.4747	**47%**
	Roubo E Fraude (Externa) (Theft And Fraud (External)	0.3387	**34%**

Tabela 15 – Caracterizacão do quarto *cluster*

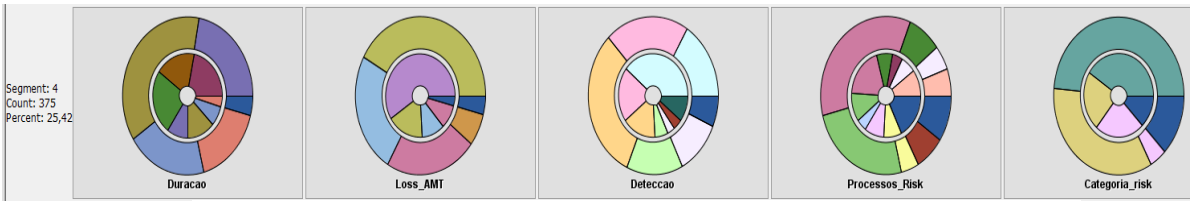


Imagem 12 – *Segment Profile* do quarto *cluster (Software SAS Miner*)

Através da *Imagem 12* podemos validar a caracterização do *cluster:* (i) observamos que o grupo aglomerou objectos com base em todas as variáveis de *input;* (ii) podemos observar que, para as variáveis de *duração* e de *perdas*, existe uma total inversão relativamente à tendência distributiva dos seus registos, ou seja, o *cluster* aglomerou algumas classes destas duas variáveis, que de facto, têm pouca expressão quantitativa. Daí a grande diferença de cores entre os círculos internos e externos; (iii) observamos também uma inversão na variável de *detecção*, principalmente nas classes *A3* e *A5*, classes com um considerável intervalo temporal mas com pouca expressão quantitativa na própria

variável; (iv) observamos que a variável referente às *categorias de risco* apresenta um aumento referente às classes *Roubo e Fraude (Externa)* e *Captura, Execução e Manutenção de Transacções* relativamente à tendência geral da variável; (v) observamos por último que as classes *Meios de Pagamento* e *Outros*, são as classes da variável de *processos de risco* que de forma mais notória rompem com a distribuição normal da sua variável.

A *Imagem 13* ilustra a distribuição dos registos refentes às classes das variáveis de *duração*, de *detecção* e de *perdas financeiras* na criação do quarto *cluster*.

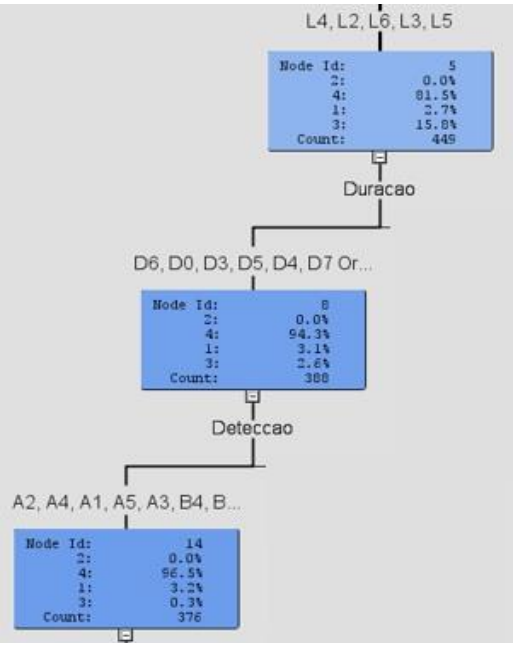


Imagem 13 - Perfil dos *cluster*s em árvore *(Software SAS Miner)*

Em suma, dos quatro *clusters* criados, o quarto é aquele que aglomerou as maiores classes de perda, de duração e de detecção: classes de *perda L1*, *L3* e *L4*, classes de *duração D3, D4, D5*, D6 assim como classes de *detecção A2*, *A3* e *A4*. Por outras palavas, este é o *cluster* que aglomerou: eventos com altos prejuízos para a instituição financeira, podendo chegar aos 10.000€, eventos que demoraram até 10 anos para serem *descobertos* e, por fim, eventos que demoraram até 3 anos para serem d*etectados*. Quanto aos *processos* de risco, destacam-se os *Meios de Pagamento,* assim como a classe *Outro*. As *categorias* de risco em destaque são novamente a *Captura, Execução e Manutenção de Transacções* e a classe *Roubo e Fraude (Externa)*.

Para terminar, podemos afirmar que o **primeiro *cluster*** foi o **principal agregador de eventos, com reduzidas *perdas* (entre 0 e 100€),** e com **reduzidos períodos temporais de *duraç*ão e de *detecção* (24 horas)**. A **classe de *processos* de risco** em destaque é o ***Activo,*** o que não deixa de ser curioso dado que esta classe é a segunda que mais impacto tem a nível de *severidade*, pese a sua *frequência* ser reduzida. A **classe das *categorias* de risco** em maior destaque é a dos ***Sistemas***, uma **classe com uma *frequência* mediana mas com uma *severidade* muito baixa**. O **segundo *cluster*** apresenta uma **tendência evolutiva no que aos montantes *perdidos* (0 a 1.500€) e aos intervalos temporais de *duração* (1º mês)** e de ***detecção* (1º mês)** dizem respeito. A **classe de *duração* D2 (1º mês) foi determinante na forma como agregou os registos que vieram a constituir este *cluster*.** As **classes de *processos* de risco** em destaque são, por um lado, os ***Processos de Suporte*** cujos eventos **têm pouca expressão quer ao nível da *frequência* quer ao nível da *severidade*** e, a classe ***Meios de Pagamento* cujos eventos são muito significativos a nível de *frequência* mas com pouca expressão** também **para a variável *severidade*. A classe com maior expressão das *categorias* de risco é a *Captura, execução e manutenção de transacções*** uma classe **muito relevante pela sua alta *frequência*, o que acaba por se traduzir também numa alta *severidade***, apesar dos seus eventos apresentarem individualmente perdas reduzidas. O **terceiro *cluster* apresenta algumas nuances divergentes:** por um lado as ***perdas* financeiras que o caracterizam são baixas (0 a 1.500€).** Contudo o **intervalo temporal de *duração* parece aumentar**, com o aparecimento de **períodos temporais que**

vão desde as primeiras 24 horas até ao 1º ano. Por outro lado podemos afirmar que este ***cluster*** **agregou eventos que só são *descobertos* após de um ano de estes terem sido terminados,** sendo este o único *cluster* onde tal acontece. Relativamente aos ***processos*** **de risco,** a **classe *Meios de Pagamento*** é a que se destaca. Relativamente às ***categorias de risco***, as **classes *Roubo e Fraude (externa)* e *Captura, execução e manutenção de transacções* voltam a assumir o protagonismo** sendo ambas **muito relevantes em termos de *frequência,*** apesar de apenas a segunda ser, e muito, relevante em termos de *severidade.* O **quarto *cluster*** é de todos, o que **aglomerou os eventos com os maiores montantes perdidos (101 a 10.000€)** e com os **maiores períodos temporais de duração (1 mês a 10 anos) e de detecção (1 mês a 3 anos)**. Quanto aos **processos de risco, destacam-se os *Meios de Pagamento,*** assim como a **classe *Outro*. As *categorias* de risco em destaque são novamente a *Captura, Execução e Manutenção de Transacções* e a classe *Roubo e Fraude (Externa)*.**

5. CONCLUSÕES

Neste capítulo iremos apresentar as várias conclusões adquiridas durante a realização desta dissertação. O objectivo principal consistia em **identificar factores que podem ter impacto na *frequência* e na *severidade* do risco operacional dentro de uma instituição financeira.** Para que tal fosse atingido, foi necessário desenvolver etapas de trabalho. Foram então criados três objectivos específicos relacionados entre si. Não só a conclusão do anterior iria ser essencial para a realização do seguinte, como também cada um à sua maneira proporcionou um diferente incremento de conhecimento sobre o fenómeno em estudo. A concretização destes três objectivos específicos forneceu a base para responder ao objectivo principal deste trabalho.

A **análise exploratória** da base de dados de risco operacional foi o primeiro objectivo específico proposto. Estes dados foram analisados sempre em função de duas variáveis: *frequência* e *severidade*. Foram anteriormente definidos quais os tipos de *processos* e de *categorias* de risco que iriam ser utilizadas e qual o melhor critério para identificar o momento temporal em que um evento de risco era iniciado, descoberto e terminado. Relembrar que cada análise foi executada sobre um subconjunto de dados diferente, pois apenas se podiam utilizar os registos que preenchessem os critérios exigidos, como por exemplo as suas datas ou os seus valores de perda.

Tendo feito estas análises, estamos em condições de concluir que: as análises efectuadas entre eventos *iniciados/descobertos*, *iniciados/terminados* e *descobertos/terminados* indicam que **o período compreendido entre o 1º mês e o 1º ano é aquele onde ocorre um maior número de eventos de risco,** oscilando entre os 73% e os 89%; por outro lado, **é neste período onde ocorre o maior número de eventos de risco com efectivas perdas para a instituição**, oscilando entre os 64% e os 85%. Importante referir também a **existência de uma relação entre a variável *frequência* e a variável *severidade* nestes intervalos temporais**. Podemos observar que a maioria do volume de perdas em cada situação ocorre neste período de maior *frequência*. A *severidade* oscila entre os 72% e os 92%.

Remetendo para uma análise mais especifica, podemos afirmar que: o período onde a ***severidade* é mais elevada na análise *eventos iniciados/descobertos*** é durante o segundo e o terceiro mês, albergando 66% do volume perdido; o período onde a ***severidade* é mais elevada na análise eventos *iniciados/terminados*** é durante o primeiro mês, albergando 52% do volume perdido; o período onde *a* ***severidade* é mais elevada na análise eventos *descobertos/terminados*** é durante as primeiras 24 horas, albergando 63% do volume de perdas; a análise *eventos descobertos após terem sido terminados* não foi incluída no parágrafo anterior pela opção de não se analisar o seu volume de perdas, contudo, a nível de *frequência* o período mantem-se, indo, do 1º mês ao 1º ano. Neste período ocorreram 93% dos eventos de risco, sendo que 82% foram eventos com efectivas perdas.

Observando agora as análises cujos intervalos temporais foram faseados de 1995 a 2014, estamos em condições de concluir que: as análises efectuadas entre *eventos iniciados/não descobertos, iniciados não terminados* e *descobertos não terminados* indicam que os anos de **2012** e **2013**, apesar de nem sempre serem os momentos em que um maior número de eventos de risco ocorre, **são,** por sua vez**, os anos onde ocorrem um maior número de eventos com efectivas perdas,** oscilando entre os 50% e os 60%. **Existe uma relação entre a variável *frequência* e a variável *severidade* nestes intervalos temporais,** visto que é neste par de anos que ocorrem os eventos que mais contribuem para o volume total perdido pela instituição, oscilando entre os 70% e os 97%. De referir por último que, a **nível de *severidade,* 2013 é o ano mais prejudicial de todos**. Nos três cenários

mencionados no início deste parágrafo, 2013 aglomerou 61% de 80%, 82% de 90% e, por último 85% de 97% do volume perdido entre 2012 e 2013, sendo que, apenas num caso este ano obteve um maior número de eventos com perdas (*frequência*) do que 2012. Contudo foi uma diferença marginal.

Remetendo para uma análise mais especifica, podemos afirmar que: a análise de *eventos iniciados não descobertos* é a única das três previamente mencionadas em que, efectivamente ocorreram mais eventos de risco em 2012 e 2013 do que em qualquer outro ano, atingindo os 60**%. Quer isto dizer que este tipo de eventos ocorrem maioritariamente nestes dois anos, verificando-se aqui o valor mais elevado da *frequência* de eventos com perdas financeiras. São estes os eventos que mais contribuem para o volume perdido (*severidade*).** Ainda na análise de *eventos iniciados não descobertos*, esta é a única das três onde em todos os intervalos temporais mais de 75% dos eventos ocorridos são de facto eventos com *perdas* financeiras. Contrariamente, nas restantes duas, a proporção de eventos com efectivas perdas em relação ao número total de eventos ocorridos é sempre baixa, mesmo nos anos críticos de 2012 e 2013.

Observando agora as análises aos *processos de risco,* estamos em condições de concluir que: de todas as classes de *processos* de risco observadas existem sete (***activos, gestão canal, meios de pagamento, outro, processos de suporte, SEM NOME* e *valores mobiliários)* que juntas contêm 75% do total de eventos de risco**, sendo as restantes residuais neste aspecto. **As classes com maior número de eventos são aquelas onde a *frequência* é mais elevada**, atingindo os 85**%; existe uma relação entre a variável *frequência* e *severidade***, dado que o referido valor de eventos com perdas, é responsável por 84% do volume perdido pelos eventos dos sete processos em destaque, **sendo possível aceitar a relação entre *frequência* e *severidade* em praticamente todas as classes. Tal não ocorre em quatro das cinco classes com maior frequência,** quando observadas individualmente**.**

Remetendo para uma análise mais especifica, podemos afirmar que: **as classes de *processos* de risco *Meios de Pagamento* e *Outro*, são as classes com o maior número de eventos de risco com efectivas perdas (*frequência*)** com 36% e 14% respectivamente. **Contudo, a nível de volume de perdas (*severidade*) apenas atingem os 12% cada**. As **classes de *processos* de risco *Activos* e *Valores Mobiliários*, são classes que apenas agregam 5% e 10% respectivamente da variável *frequência.*** Ainda assim, **têm uma muito maior relevância para a variável *severidade* atingindo os 14% e os 33%, sendo estes os processos que mais contribuem para esta variável.**

Observando agora as análises às *categorias de risco,* estamos em condições de concluir que: de todas as classes de *categorias de risco* observadas existem três **(*Captura Execução e Manutenção de Transacções, Roubo e Fraude Externas e Sistemas)* que juntas contêm 77% (38%-20%-19%) dos eventos de risco ocorridos**, sendo as restantes residuais neste aspecto. **Estas classes são também aquelas nas quais a *frequência* é mais elevada**, atingindo os **79% (37%-34%-8%)** do total de eventos com efectivas perdas. **Existe uma relação entre as variáveis *frequência* e *severidade***, dado que o valor de eventos com perdas anterior é responsável por 76% do volume perdido pelas três categorias de risco em análise; **a categoria de risco que merece maior destaque é a *Captura Execução e Manutenção de Transacções***, **dado que os seus 37% da variável *frequência* são responsáveis por 70% da variável *severidade***. Sendo possível aceitar a relação entre *frequência* e *severidade* em praticamente todas as classes**, tal não ocorre em três categorias,** se as observarmos individualmente**.**

Remetendo para uma análise mais especifica, podemos afirmar que: as **classes de *categorias de risco Roubo e Fraude Externa* e *Sistemas*** são duas das três classes com **maior número de eventos**

de risco com efectivas perdas (*frequência*) com 34% e 8% respectivamente, contudo a **nível de volume de perdas (*severidade*) apenas atingem os 5% e os 2% respectivamente**; a classe ***Práticas de Mercado* ou de *Negócios Impróprios* tem para a variável *frequência* um valor residual de 4%, contudo, esta classe contém eventos de risco responsáveis por 16% do total de volume perdido (*severidade*),** sendo assim a segunda categoria que mais contribui para esta variável.

A organização da informação de forma a construir um conjunto de análises que permitam um maior conhecimento sobre os eventos de Risco Operacional foi o segundo objectivo específico proposto. Após a conclusão da análise exploratória foi necessária a criação de uma nova tabela de dados que foi posteriormente utilizada na análise de *Clusters*. Esta tabela contem 2519 registos (70% da fonte de dados original) e é composta por seis variáveis, sendo que três transitam da base de dados inicial*:* o *identificador do evento*, os *processos de risco* e as *categorias de risco*. As restantes três colunas foram criadas de raiz, nomeadamente, a coluna referente aos *montantes perdidos*, à *detecção* de um evento de risco (tempo que demora desde que um evento é iniciado até que seja posteriormente descoberto) e à *duração* de um evento de risco (tempo que demora desde que um evento é iniciado e posteriormente finalizado).

As análises realizadas recaíram essencialmente sobre as três colunas criadas de raiz, tendo estas sido analisadas individualmente e em conjunto. Começando pelas **análises individuais**, podemos concluir que; relativamente aos ***montantes perdidos***, os eventos de risco em análise são **maioritariamente eventos com *perdas* médias/baixas**; 94% das perdas são compreendidas entre os **0 e os 10.000€**, sendo que 46% desse valor corresponde a perdas compreendidas entre os **0 e os 100€.** Relativamente à ***duração* dos eventos de risco**, podemos concluir que **75% dos eventos demoram entre 24 horas a um ano para serem finalizados**, sendo que desse valor 30% corresponde a eventos que **iniciam e *terminam* dentro dos primeiros 30 dias**. Relativamente **à *detecção* de um evento de risco** podemos concluir que **88% dos eventos são *descobertos* durante o 1º ano de existência**, sendo que praticamente **63% são *descobertos* entre as primeiras 24 horas e o 1º mês.** Por outro lado concluímos também que **84% dos eventos são descobertos antes de serem terminados.**

Observando agora as análises conjuntas das três variáveis anteriores, podemos concluir que: a **maioria dos eventos de risco**, independentemente das quantias *perdidas,* **caracterizam-se não só por um período de *duração* compreendido entre o 1º mês e o 1º ano, como por um período de *detecção* compreendido entre as primeiras 24 horas e o 1º ano**. A grande excepção ocorre em **eventos com *perdas* iguais ou superiores a 1.000.000€, dado que estes eventos são caracterizados por períodos de *duração* e de *detecção* que podem chegar aos três anos**. **Regra geral, existe uma tendência que relaciona eventos com médias/baixas perdas a intervalo temporais de duração e de detecção igualmente médios/baixos e vice-versa. É possível também afirmar que os intervalos temporais de *duração* e de *detecção* quase sempre coincidem**, o que faz sentido, dado que é maior a probabilidade de um evento ser iniciado e *descoberto* dentro do período em que este é iniciado e *terminado* do que o seu contrário. Curioso observar uma excepção que ocorre nos eventos que demoram mais tempo até serem *terminados* (mais de 10 anos). Estes são eventos que, não só se caracterizam por *perdas* médias (até 10.000€), como também se caracterizam por apenas serem *descobertos* após terem sido dados como *terminados*. Sendo esta uma excepção, mais uma vez os seus intervalos temporais são idênticos. **Regra geral existe uma relação que faz coincidir os intervalos de duração e de detecção de um evento de risco.**

Iremos observar agora as análises que relacionam as três variáveis anteriores juntamente com os *processos de risco* e com as *categorias de risco* selecionadas. **Para esta análise foram utilizadas as classes de processos *Activo*, *Meios de Pagamento*, *Outro* e *Valores Mobiliários***, **assim como as classes de categorias de risco *Sistemas, Captura, execução e manutenção de transacções*, *Roubo e Fraude (externa)* e *Práticas de Mercado ou de Negócios Impróprias***. Desta forma estamos em condições de concluir que: relativamente aos ***processos de risco*, as três classes que registam um maior número de observações** são facilmente **caracterizadas por eventos com *perdas* reduzidas**, não ultrapassando os **500€** e por **intervalos de *duração* e de *detecção*** também eles **compreendidos entre as primeiras 24 horas e o 1º ano.** Curioso observar que a classe com o maior número de observações (*Meios de Pagamento*) é composta por eventos *descobertos* durante o primeiro mês, mas que chegam a demorar até um ano para serem *terminados.* Pode-se então **concluir que estes processos são mais danosos para a instituição financeiras em função da variável *frequência* do que da variável *severidade***. O seu excessivo número de ocorrências tornam-nos conjuntamente mais prejudiciais do que a sua *severidade* individual. Contudo, esta **tendência é invertida na classe de processos *Valores Imobiliários***. De facto esta classe já tinha merecido destaque durante a análise exploratória pela forma como conjugava um **número de eventos com perdas quase residuais com o maior volume perdido de entre os vários *processos de risco***. Nesta análise a tendência mantem-se, esta **classe é caracterizada por eventos com *prejuízos* de 10.001 a 50.000€, assim como por eventos que apesar de demorarem um mês até serem *terminados***, **podem demorar até cinco anos para que sejam *descobertos***. Relativamente às ***categorias de risco***, duas das três classes com um maior número de observações **seguem a tendência das reduzidas *perdas***, não ultrapassando os **100€** e **dos reduzidos intervalos de *duração* e de *detecção***, compreendidos entre as **primeiras 24 horas e o 1º mês**. **A excepção ocorre na classe *Roubo e Fraude (externa)* caracterizada por eventos com *perdas* médias**, **de 501 a 1500€,** assim como **por eventos que apesar de demorarem um mês até serem *descobertos*, chegam a demorar um ano até serem *terminados***. Apesar desta excepção, podemos **concluir novamente que estas categorias são mais danosas para a instituição financeira em função da variável *frequência* do que da variável *severidade***. Contudo, mais uma vez foi analisada **uma excepção a esta tendência**, a classe ***Práticas de Mercado ou de Negócios Impróprias***. De facto, esta classe já tinha merecido destaque durante a análise exploratória pela forma como conjugava um número de eventos com *perdas* **quase residuais com o segundo maior valor percentual em termos de volume perdido** de entre as várias *categorias de risco*. Esta classe caracteriza-se por eventos com ***perdas* médias** entre os **10.001€ e os 50.000€,** tendo mais uma vez um **período temporal de *duração* e de *descoberta*** dentro do normal, **entre as primeiras 24 horas e o 1º mês**.

O desenvolvimento de segmentos capazes de identificar factores que possam vir a ser indicadores da evolução das variáveis *frequência* e *severidade* foi o terceiro e último objectivo específico proposto. Para responder a este objectivo procedemos a uma análise de *clusters* com o objectivo de encontrar novas e mais robustas relações entre os nossos dados de forma a conferir um maior grau de conhecimento sobre o fenómeno em estudo.

Através da análise de *Clusters* conseguimos retirar as seguintes conclusões: as variáveis mais relevantes para a construção dos quatro *Clusters*, isto é, **as variáveis que maior impacto tiveram na agregação de registos foram a variável *duração* e *detecção***. **Os *Clusters* criados** apesar de terem uma distribuição a nível de registos, bastante equilibrada **são bastante influenciados pelo número de**

observações de cada classe e, consequentemente, de cada variável. Esta realidade explica porque é que em **três grupos existem *perdas* baixas ou médias/baixas, assim como intervalos de *duração* e de *detecção* com iguais características**, visto que os dados ao nosso dispor apresentam estas mesmas particularidades. **Observamos novamente uma tendência que relaciona *perdas* mais elevadas a intervalos temporárias de *duração* e de *detecção* igualmente mais elevados e vice-versa**. Uma tendência perfeitamente **ilustrada no primeiro e no quarto *cluster*.** Relativamente às classes de ***processos de risco,*** podemos concluir **que apenas a classe *Meios de Pagamento* e *Outro* foram relevantes no agregar de eventos com maiores prejuízos financeiros**. Contudo, esta **relevância ocorreu muito mais pelo lado da variável *frequência*, na qual estas apresentam elevados valores**, **do que pelo lado da variável *severidade,*** na qual estas são relativamente modestas. **Todas as restantes classes foram agregadas por influência das variáveis *duração* e *detecção***. Relativamente às classes de ***categorias de risco,*** podemos concluir que tanto a classe ***Roubo e Fraude (externa)***, como a classe ***Captura, execução e manutenção de transações* são as mais relevantes para explicar a agregação de valores com maiores prejuízos financeiros.** Destaca-se em particular **a segunda,** porque não só **domina a sua variável em termos de *frequência* e de *severidade***, como é também a classe que **agrega praticamente metade dos eventos do *cluster.* Sendo estes eventos caracterizados por *perdas* médias/altas (101 a 10.000€).** Observando os ***Clusters* 2 e 3** concluímos que **sempre que as classes de montantes *perdidos,* nomeadamente a *L2* (101/500€) e a *L3* (501/1500€), tendem a crescer**, a classe de *categoria* predominante é a referida. **Podemos então concluir que, dentro das classes observadas, existe uma maior importância das *categorias de risco* relativamente aos *processos de risco* no que ao aglomerar de eventos diz respeito, eventos esses caracterizados pela sua propensão de trazerem maiores prejuízos financeiros a longo prazo para a instituição.**

Com base nos dados de que dispomos e nas respectivas análises efectuadas, podemos apresentar algumas conclusões que irão ajudar a responder ao objectivo final:

- As diferentes análises indicam que **os eventos com períodos temporais compreendidos entre o 1º mês e o 1º ano são aqueles onde ocorrem perdas efectivas para a instituição, sendo assim estes eventos os que mais contribuem para o agravamento do volume perdido pela instituição.**

- Conclui-se que **os anos 2012 e 2013**, sempre que tal análise foi possível, **são os piores anos quer a nível da *frequência* de eventos, quer a nível da variável *severidade***, sendo que, dos dois, o pior é o ano 2013.

- Concluímos também que o **tempo que um evento demora até ser *descoberto*, assim como o tempo que demora até que este seja *terminado* são dois factores que ajudam a ditar os *prejuízos* que um certo evento irá ter, ou não, para a instituição.** Esta relação encontra-se espelhada em praticamente todas as análises efectuadas. Se estes períodos aumentam, passando de meses para anos, os seus eventos irão acompanhar esta tendência agravando as suas perdas financeiras.

- Regra geral, **existe a tendência para que os intervalos temporais de *duração* e de *detecção* coincidam.**

- Relativamente aos ***processos de risco***, podemos concluir que a **classe *Valores Imobiliários* deve preocupar as instituições, não pela quantidade** de eventos de risco que produz, **mas pela *severidade* das suas perdas** quando algum evento ocorre, principalmente **se considerarmos que são eventos *detectados* a longo prazo (entre 3 a 7 anos).**

- Sendo este uma excepção, **os restantes processos mais problemáticos** como os ***Meios de Pagamento*** ou o ***Activo*** devem, por sua vez, **captar a atenção das instituições financeiras**, mas por um motivo diferente. Neste caso, **o impacto nefasto destes eventos é justificado pela elevada *frequência* com que os seus eventos ocorrem, muito mais do que pela sua *severidade* individual.**

- Apenas **um dos três exemplos** apresentados **deverá preocupar a instituição exclusivamente em função do impacto financeiro dos seus eventos**, ao passo que os **restantes dois remetem esse impacto para a quantidade de eventos ocorridos.**

- Relativamente às ***categorias de risco***, podemos concluir que a **classe *Práticas de Mercado ou de Negócio Impróprios* tem um comportamento parecido com o *processo Valores Mobiliários*,** ou seja, é uma **classe com uma *frequência* de eventos de risco muito baixa**, **mas que quando ocorrem, têm uma *severidade* bastante mais alta.**

- Ainda, **a classe *Roubo e Fraude Externa* é a típica classe cujo efeito nefasto se justifica mais pela alta *frequênc*ia com que os seus eventos de risco ocorrem do que pelas suas *severidades* individuais.** Por outras palavras, só o facto de ocorrerem tantos eventos deste tipo é que torna esta classe alvo de reflexão para a instituição financeira.

- No entanto, **a classe mais problemática desta variável é a *Captura, Execução e Manutenção de transacções*. Esta classe não só é a responsável pelo maior número de eventos de risco com efectivas perdas da sua variável, como a sua *severidade* é também ela muito significativa.** Por outras palavras, esta é a classe cujos eventos irão ser sempre negativos, quer quando ocorrem em menor número, quer principalmente quando ocorrem em número elevado como é o caso.

- Apenas **um dos três exemplos** apresentados **deverá preocupar a instituição exclusivamente em função do impacto financeiro dos seus eventos**, ao passo que os **restantes dois, remetem esse impacto para a quantidade de eventos ocorridos.**

- Utilizando estes exemplos, que claramente se destacam dos demais, e complementando-os com as conclusões anteriormente apresentadas sobre as restantes classes, **podemos afirmar que as maiores perdas que a instituição sofre são justificadas muito mais em função do número de eventos ocorridos do que do seu impacto financeiro,** sendo que a instituição não deve descorar nenhuma delas nos seus processos de tomada de decisão.

- A análise de *Clusters* veio **confirmar a suspeita relativamente à importância das variáveis de *detecção* e de *duração* na distribuição dos eventos de risco, mais concretamente na sua influência directa sobre os *montantes perdidos*.**

- Através desta análise **concluímos que as *categorias de risco* têm maior predominância do que os *processos de risco* na forma como os eventos de risco são separados em função dos seus *montantes perdidos*. Salienta-se a classe *Roubo e Fraude Externa,*** devido **à alta *frequência* com que os seus eventos ocorrem** e também **a classe *Captura, Execução e Manutenção,*** devido **à sua alta *frequência* e *severidade***. Estas **variáveis conseguem aglomerar muito mais eventos de risco com elevadas *perdas***, sendo **estes eventos responsáveis por potenciais prejuízos a longo prazo para as instituições financeiras**.

- **As classes de *processos* também são agregadoras de eventos com elevados prejuízos**, contudo, **fazem-no maioritariamente pelo lado da variável *frequência*** e fazem-no **em número consideravelmente inferior**, o que nos leva a concluir que **estão a ser enviesadas pelas outras variáveis que caracterizam cada registo como a *duração* ou a *detecção***.

Desta forma o autor conclui que as técnicas de Data Mining podem ajudar a identificar alguns dos factores de risco mais importantes para as instituições financeiras. Na realização deste trabalho o foco recaiu sobre a **modelação descritiva**, cuja base assentou na produção de resumos sobre grandes quantidades de dados, conferindo assim um maior entendimento sobre estes mesmos dados, da mesma forma que facilitou todo o processo de criação de informação útil e de posterior conhecimento. Para tal, utilizei ferramentas de natureza eminentemente exploratória, em muitos casos, ferramentas gráficas, de imagem, ou até mesmo técnicas de redução/compressão de dados, para que, sem alienar o que de relevante se desejava reter, fosse possível aumentar o grau de compreensão e de interpretabilidade. Desta forma, foi possível atingir todas as conclusões referidas ao longo desde capítulo, respondendo a todos os objectivos específicos propostos sem nunca perder de vista o objectivo central deste trabalho. **As conclusões retiradas desta dissertação devem ser encaradas como bons indicadores sobre os factores, e áreas de risco onde a instituição financeira deverá focar o seu controlo interno e as suas medidas de mitigação.**

6. LIMITAÇÕES E INVESTIGAÇÃO FUTURA

Durante a realização deste trabalho foram alguns os constrangimentos encontrados que de certa forma limitaram as análises realizadas, quer a nível global quer a nível individual. A base de dados de risco operacional apresentava as seguintes limitações:

- Elevada quantidade de valores omissos em diferentes colunas. Por exemplo, nas colunas que dizem respeito às datas existiam 733 eventos de risco sem data de início, 1225 sem data de descoberta e 1204 sem data de termo.
- Erros da parte de quem recolheu a informação. Por exemplo, colunas de *processos* ou de *categorias* de risco sem nome em algumas classes, ou outras, nas quais na mesma coluna existia uma classe intitulada "*outro*" e uma outra intitulada "*outros*".

Importante referir que vários dos eventos de risco que constituem a fonte de dados ocorreram durante um período de dificuldades para o sector financeiro mundial e, por consequência para o sector financeiro Português. Estas dificuldades tiveram um claro impacto directo nos níveis de risco operacional das instituições financeiras, o que pode, de certa forma, enviesar alguns resultados obtidos.

Para futuros trabalhos de investigação nesta área recomenda-se a utilização de uma base de dados mais robusta, completa e diversificada, ao mesmo tempo que se deve apostar em diferentes e mais avançados procedimentos analíticos de Data Mining. Recomenda-se: (i) a utilização de diferentes técnicas de *clustering* não abordadas neste trabalho; (ii) proceder à criação de regras de associação, técnicas não supervisionadas com o objectivo de descobrir padrões e relações desconhecidas ou pouco obvias entre os dados, mas também com o objectivo de gerar um conjunto de regras que definam esses mesmos padrões e essas mesmas relações encontradas.

7. BIBLIOGRAFIA

Bação, F. L., & Painho, M. (2003). ASPECTOS METODOLÓGICOS DA UTILIZAÇÃO DO DATA MINING NO ÂMBITO DA GEOGRAFIA. *Finisterra: Revista portuguesa de geografia, 38*(75), 135-147.

Beroggi, G. E. G., & Wallace, W. (1994). Operational Risk Management: A New Paradigm for Decision Making. *Systems, Man and Cybernetics, IEEE Transactions on, 24*(10), 1450-1457.

Bhambri, V. (2011). Application of Data Mining in Banking Sector. *International Journal of Computer Science and Telecommunications, 2*(2).

Bisson, P., Stephenson, E., & Viguerie, S. P. (2010). The productivity imperative. *McKinsey Quarterly*, 1-7.

Brachman, R. J., & Anand, T. (1994). The Process of Knowledge Discovery in Databases: A First Sketch. In J. Komorowski (Ed.), *Lecture Notes in Computer Science* (pp. 1-12). Norway.

Buchelt, R., & Unteregger, S. (2004). Cultural Risk and Risk Culture: Operational Risk after Basel II. *Financial Stability Report*(6), 86-97.

Chen, M.-S., Han, J., & Yu, P. S. (1996). Data mining: An Overview from a Database Perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866-883.

Chitra, K., & Subashini, B. (2013). Data Mining Techniques and its Applications in Banking Sector. *International Journal of Emerging Technology and Advanced Engineering, 3*(8), 219-226.

Crouchy, M. (2001). *Risk Management*. New York: McGraw Hill.

Crouhy, M., Galai, D., & Mark, R. (1998). Key steps in building consistent operational risk measurement and management. *Operational Risk and Financial Institutions*. London: Risk Books.

Cummins, J. D., & Embrechts, P. (2006). Introduction: Special section on operational risk (Vol. 30).

Cummins, J. D., Lewis, C. M., & Wei, R. (2006). The Market Value Impact of Operational Risk Events For U.S. Banks and Insurers. *Journal of Banking & Finance, 30*(10), 2605-2634.

De Fontnouvelle, P. (2007). Assessing the Financial Industry's Operational Loss Experience: Interesting Findings and Some Open Questions.

De Koker, R. (2006). Operational Risk Modelling: Where Do We Go From Here? *The advanced measurement approach to operational risk* (pp. 37-57). London: Risk Books.

Di Renzo, B., Hillairet, M., Picard, M., Rifaut, A., Bernard, C., Hagen, D., . . . Reinard, D. (2007). OPERATIONAL RISK MANAGEMENT IN FINANCIAL INSTITUTIONS: PROCESS ASSESSMENT IN CONCORDANCE WITH BASEL II. *Software Process: Improvement and Practice, 12*(4), 321-330.

Dima, A. M., & Orzea, I. (2012). Risk Management in Banking: Academy Publish.

Edelstein, H. A. (1998). *Introduction to Data Mining and Knowledge Discovery*: Two Crows Corporation.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From Data Mining to Knowledge Discovery in Databases. *AI magazine, 17*(3), 37-54.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM, 39*(11), 27-34.

Friedman, J. H. (1998). DATA MINING AND STATISTICS: WHAT'S THE CONNECTION. *Computing Science and Statistics, 29*(1), 3-9.

Geiger, H. (2000). *Regulating and Supervising Operational Risk for Banks.* Paper presented at the Conference "Future of Financial Regulation", Tokyo.

Gonçalves, R. A. H. (2011). *Sistemas de informação para gestão de risco operacional em instituiçoes financeiras.* (doctoral thesis), Universidade Técnica de Lisboa - Instituto Superior de Economia e Gestão, Lisbon.

Han, J., Pei, J., & Kamber, M. (2012). *Data Mining: Concepts and Techniques, Third Edition*: Elsevier Science.

Hand, D. J. (1998). Data Mining: Statistics and More? *The American Statistician, 52*(2), 112-118.

Hand, D. J. (1999). Statistics and Data Mining: Intersecting Disciplines. *ACM SIGKDD Explorations Newsletter, 1*(1), 16-19.

Herring, R. J. (2002). The Basel 2 Approach to Bank Operational Risk: Regulation on the Wrong Track. *The Journal of Risk Finance, 4*(1), 42-45.

Hoffman, D. G. (1998). New Trends in Operational Risk Measurement and Management *Operational Risk and Financial Institutions* (pp. 29-42). London: Risk Books.

Hormozi, A. M., & Giles, S. (2004). Data Mining: A Competitive Weapon for Banking and Retail Industries. *Information Systems Management, 21*(2), 62-71.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: a Review. *ACM computing surveys (CSUR), 31*(3), 264-323.

Jednak, D., & Jednak, J. (2013). Operational Risk Management in Financial Institutions. *Management*, 66.

Jobst, A. A. (2007). The Treatment of Operational Risk under the New Basel Framework - Critical Issues (Vol. 8).

Kaiser, T., & Kohne, M. (2006). *An Introduction to Operational Risk*. London: Risk Books.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data: an Introduction to Cluster Analysis* (Vol. 344): John Wiley & Sons.

Koyuncugil, A. S., & Ozgulbas, N. (2008a). A Data Mining Model for Detecting Financial and Operational Risk Indicators of SME's. *International Journal of Social, Behavioral, Educational, Economic and Management Engineering, 2*(10), 1082-1085.

Koyuncugil, A. S., & Ozgulbas, N. (2008b). Financial Profiling for Detecting Operational Risk by Data Mining. *OC091, 17*.

Kuonen, D. (2004). Data mining and Statistics: What is the connection? *The Data Administration Newsletter, 30*, 1-6.

Lewis, C. M., & Lantsman, Y. L. (2005). What is a Fair Price to Transfer the Risk of Unauthorized Trading? A Case Study on Pricing Operational Risk. In E. Davis (Ed.), *Operational Risk: Practical Approaches* (pp. 315-356). London: Risk Books.

Lopez, J. A. (2002). What Is Operational Risk? *FRBSF Economic Letter*.

Milligan, J. (2004). Prioritizing Operational Risk. *Banking Strategies, 80*, 67.

Moin, K. I., & Ahmed, D. Q. B. (2012). Use of Data Mining in Banking. *International Journal of Engineering Research and Applications, 2*(2), 738-742.

Mooi, E., & Sarstedt, M. (2011). *A Concise Guide to Market Research The Process, Data, and Methods Using IBM SPSS Statistics*

Moosa, I., Li, L., & Naugthton, T. (2015). Operational Risk, the Legal System and Governance Indicators: A Country-Level Analysis. *Applied Economics, 47*(20), 2053-2072.

Moosa, I. A. (2007). Operational Risk: A Survey. *Financial Markets, Institutions & Instruments, 16*(4), 167-200.

Pivk, A., Vasilecas, O., Kalibatiene, D., & Rupnik, R. (2013). On approach for the implementation of data mining to business process optimisation in commercial companies. *Technological and Economic Development of Economy, 19*(2), 237-256.

Power, M. (2005). The Invention of Operational Risk. *Review of International Political Economy, 12*(4), 577-599.

Rao, V., & Dev, A. (2006). Operational Risk: Some issues in Basel II AMA Implementation in US Financial Institutions *The advanced measurement approach to operational risk* (pp. 273-294). London: Risk Books.

Rempfler, R. (2011). *Operational Risk Management: A Case Study at a Global Institution.* (master's thesis), School of Business, Economic and Law, University of Gothenburg.

Shepheard-Walwyn, T., & Litterman, R. (1998). Building a Coherent Risk Measurement and Capital Optimisation Model for Financial Firms. *Federal Reserve Bank of New York Economic Policy Review*, 171-182.

Thirlwell, J. (2002). Operational Risk: The Banks and the Regulators Struggle. *Balance Sheet, 10*(2), 28-31.

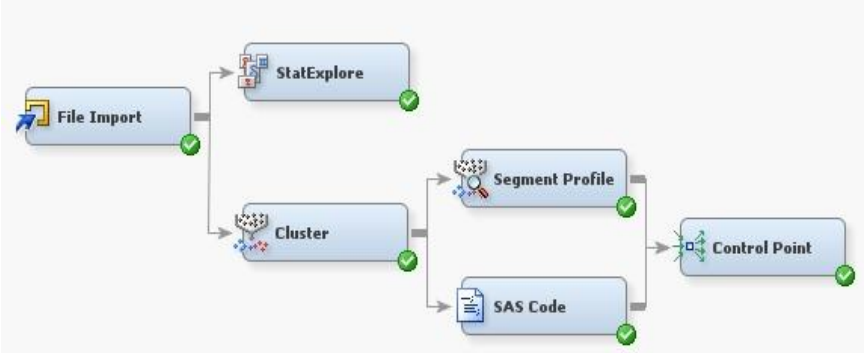
Thirty, G. o. (1993). *Derivatives: Practices and Principles* (Vol. 4): Group of Thirty.

Tukey, J. W. (1977). Exploratory data analysis.

Turing, D. (2003). The Legal and Regulatory View of Operational Risk *Advances in operational risk: Firm-wide issues for financial institutions* (pp. 253–266). London: Risk Books.

Valverde, R. (2015). Data Mining: A Tool to Increase Productivity in Supply Chain Management. *SSRN Electronic Journal*, 30.

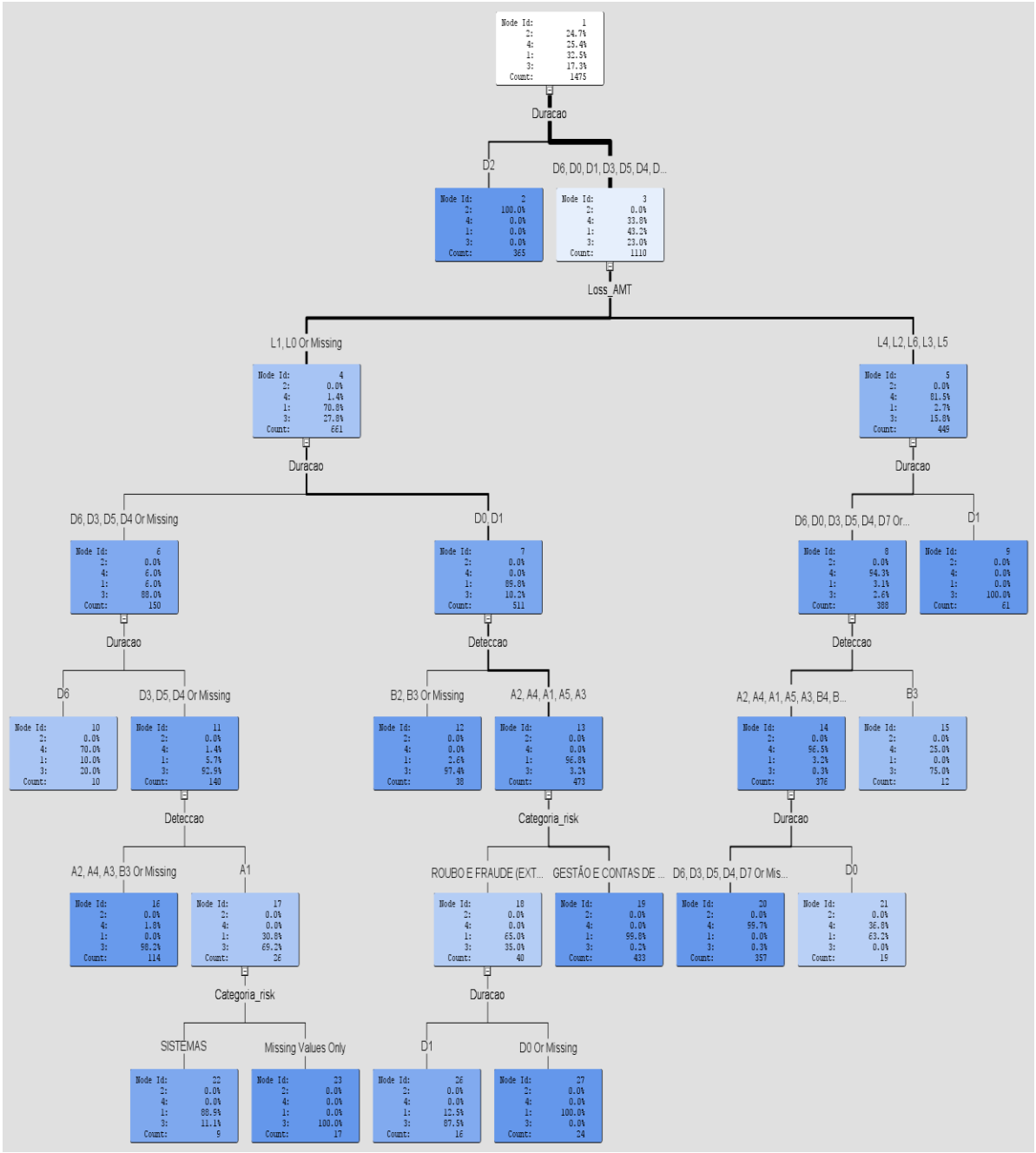
8. ANEXOS



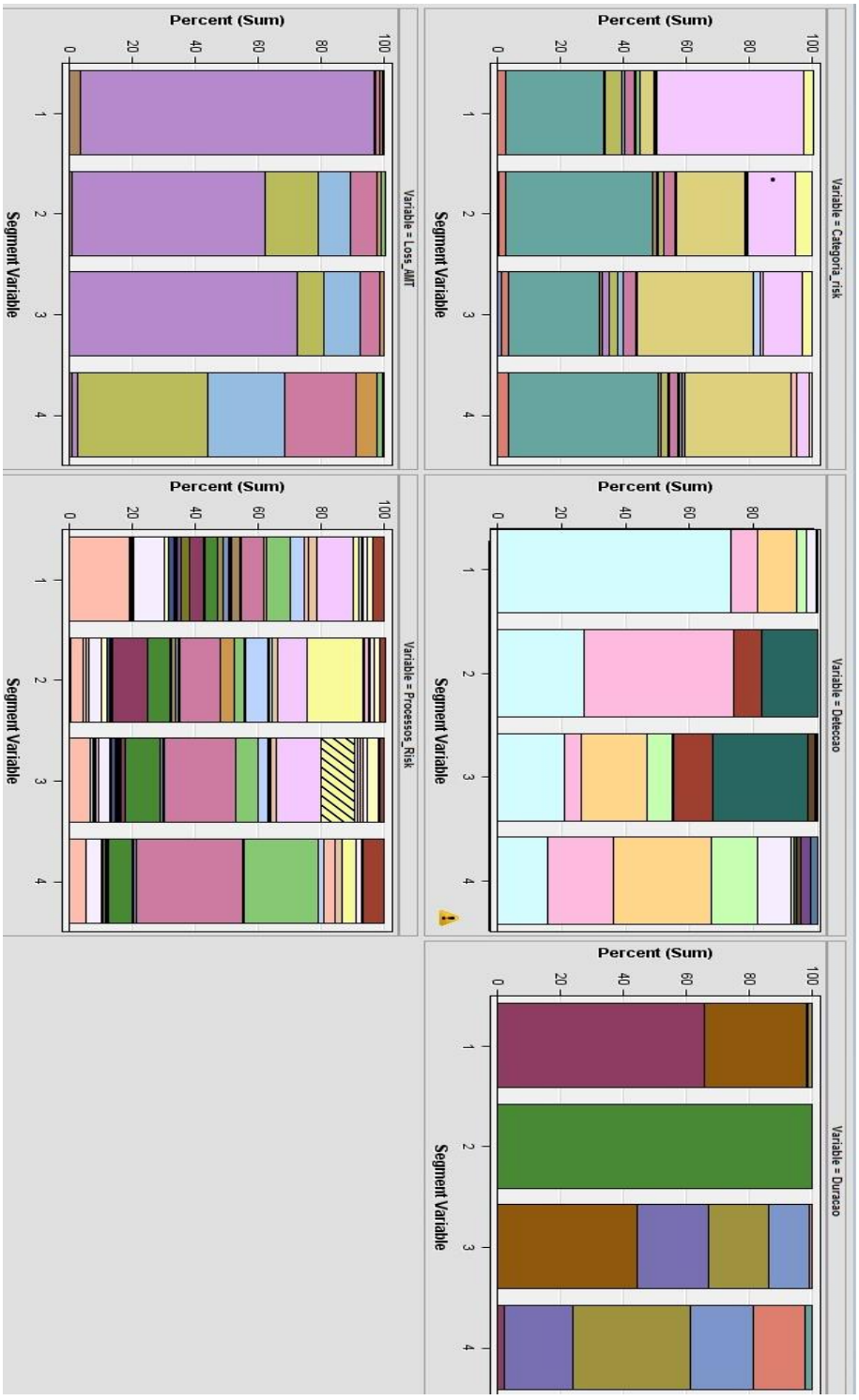
Anexo 1 - Diagrama do SAS Enterprise Miner

```
%Let CLust=cluster;
Data &EM_EXPORT_VALIDATE.;
 Set &EM_IMPORT_DATA.(Keep=%EM_ID _Segment_Label_);
      Rename _Segment_Label_=Seg_&CLust.;
Run;
%Macro Ficheiro;
Proc Sql;

 Select Sum(Distance),
        Max(_SEGMENT_) into:Distance,
                           :NCluster From &EM_Import_data.;
Quit;
%Let Distance=&Distance.;
%Let NCluster=&NCluster.;
%Put =&NCluster.=&Distance.=;
        %If %sysfunc(exist(&Em_Lib..&CLust.,DATA)) ne %str(1) or &NCluster. eq 1 %Then %Do;
           Data &Em_Lib..&CLust.;
               NumClusters=&NCluster.;
               ErroTotal=&Distance.;
               Output;
           Run;
        %End;
        %Else %Do;
           Data New;
               NumClusters=&NCluster.;
               ErroTotal=&Distance.;
               Output;
           Run;
           Data &Em_Lib..&CLust.;
            Set &Em_Lib..&CLust. New;
           Run;
        %End;
      %EM_Register(Key=SEG,Type=DATA);
      Proc Sql;
        Create Table &EM_USER_SEG As
        Select NumClusters,
               Avg(ErroTotal)   As AvgErro,


                Min(ErroTotal)   As MinErro,
                Max(ErroTotal)   As MaxErro
         From &Em_Lib..&CLust.
         Group By NumClusters;
       Quit;
     %EM_Report(Key=SEG,
                Viewtype=DATA,
                Autodisplay=Y,
                Block=Cotovelo,
                Description=Erro Total);
%Mend Ficheiro;
%Ficheiro;
```

Anexo 2- Código *SAS* para realizar o Gráfico do Cotovelo



Anexo3 - Perfil em árvore dos clusters (*Software SAS MINER*)



Anexo 4 – *Segment Profile* dos *clusters*