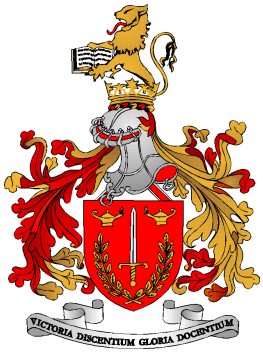
INSTITUTO SUPERIOR DE CIÊNCIAS POLICIAIS E SEGURANÇA INTERNA



Hugo **F**erreira **L**opes

Aspirante a Oficial de Polícia

O POTENCIAL DO DATA MINING PARA O ENRIQUECIMENTO DAS INFORMAÇÕES POLICIAIS NA PSP

Orientador:

Mestre **P**edro **M**oita

LISBOA, 27 DE ABRIL DE 2011

HUGO FERREIRA LOPES

Aspirante a Oficial de Polícia

O POTENCIAL DO DATA MINING PARA O ENRIQUECIMENTO DAS INFORMAÇÕES POLICIAIS NA PSP

Dissertação Final de Mestrado Integrado em Ciências Policias

XXIII Curso de Formação de Oficiais de Polícia

Orientador:

Mestre **P**edro **M**oita

INSTITUTO SUPERIOR DE CIÊNCIAS POLICIAIS E SEGURANÇA INTERNA

LISBOA, 27 DE ABRIL DE 2011

Ao meu pai Horácio

à minha namorada Filomena

e em especial a TI MÃE

AGRADECIMENTOS

A realização desta dissertação de mestrado, não teria sido possível sem a colaboração de determinadas pessoas, as quais, à sua maneira, contribuíram para elaboração da mesma. Assim, não posso deixar de expressar algumas palavras de gratidão e apreço, dirigidas aquelas que mais me ajudaram a ultrapassar esta etapa da minha vida.

À minha família, no geral, e ao meus pais em particular, sem os quais, teria sido impossível atingir este objectivo de vida.

A ti Filomena, por representares tudo para mim e por estares sempre presente quer em tempos de tempestade, quer de bonança.

Ao Exmo. Sr. Mestre Pedro Moita, meu orientador, pela sua perseverança e dedicação, tornando possível a realização desta dissertação.

À Exma. Sra. Comissário Élia Chambel e ao Exmo. Sr. Agente Principal Jorge Carvalho, sem os quais seria impossível aplicar os questionários em tempo recorde.

Aos meus ilustres camaradas do XXIII CFOP, que com um arrebatador espírito de camaradagem, nunca me deixaram ficar para trás.

Aos Exmos. Srs. e Sras. Oficiais da PSP, por terem colaborado na recolha de dados para esta dissertação.

Aos meus amigos e amigas, que mais directamente me acompanharam no decurso destes cinco anos.

A todos que de alguma maneira contribuíram para a realização desta dissertação.

O meu sincero Obrigado!

RESUMO

Actualmente, vive-se uma era informacional, onde tudo é registado em bases de dados espalhadas pelas mais diversas organizações. Esta avalanche de informação torna o tratamento manual da informação impraticável e, até mesmo, impossível. Assim, tecnologias como o *Data Mining* e o *Data Warehousing* facilitam a compreensão e o tratamento informacional dos dados.

A informação recolhida através dos Sistemas de Informação e armazenada nas bases de dados, pode conter informação e conhecimento relevante, quer a nível empresarial, quer a nível policial. E é neste último âmbito que irá incidir esta dissertação, nomeadamente na importância que as ferramentas de *Data Mining* podem ter para actuação policial.

Ou seja, este estudo versa sobre o conhecimento que os Oficiais da PSP detêm sobre as ferramentas de DM, bem como, a utilização das mesmas por corpos policiais internacionais. É ainda examinada, a possibilidade de aplicação de um sistema de Data Mining ao actual Sistema Estratégico de Informação da PSP.

Palavras-Chave: Informações, Sistemas de Informação, *Data Mining*, *Data Warehousing* e Prevenção.

ABSTRACT

Currently we are undergoing an information age where everything is recorded in databases scattered by many organizations. This avalanche of information makes the manual processing of information impractical, and even impossible. Thus, technologies such as Data Mining and Data Warehousing facilitate the understanding and treatment of informational data.

The information collected through the Information Systems and stored in databases may contain relevant information and knowledge, whether at the enterprise level or at the police level. And it is this latter context that this dissertation will focus, in particular on the importance that the tools of data mining may have to police action.

In other words, this study focuses on the knowledge that officials of the PSP have about the Data Mining tools, as well as their use by international police forces. It also is examined the applicability of a data mining system to the current Strategic Information System of the PSP.

Key Words: Intelligence, Information Systems, Data Mining, Data Warehousing and Prevention.

LISTA DE SIGLAS

APC	Autoridade de Polícia Criminal
BD	Base de Dados
BI	Business Intelligence
CDC	Comando Distrital de Coimbra
COMETLIS	Comando Metropolitano de Lisboa
CPP	Código de Processo Penal
CRISP-DM	CRoss-Industry Standard Process for Data Mining
CRP	Constituição da República Portuguesa
DCBD	Descoberta de Conhecimento em Bases de Dados
DM	Data Mining
DN	Direcção Nacional
DW	Data Wharehouse
LOPSP	Lei Orgânica da Polícia de Segurança Pública
LSI	Lei de Segurança Interna
OPC	Órgão de Polícia Criminal
PMML	Predictive Model Markup Language
PSP	Polícia de Segurança Pública
RNSI	Rede Nacional de Segurança Interna
SEI	Sistema Estratégico de Informação
SEMMA	Sample, Explore, Modify, Model, Assessment
SI	Sistemas de Informação
SIG	Sistema Informação Geográfica
UTIS	Unidade de Tecnologias de Informação de Segurança
ViCLAS	Violent Crime Linkage Analysis System
WWW	World Wide Web

ÍNDICE

INTRODUÇÃO

Este trabalho insere-se no âmbito do possível tratamento da informação contida nas bases de dados da Polícia de Segurança Pública (PSP), em particular, no que concerne às potencialidades da utilização de uma ferramenta de *Data Mining*[1] no campo de acção do SEI (Sistema Estratégico de Informação) e das respectivas informações policiais.

A Polícia de Segurança Pública utiliza uma aplicação informática, o SEI. Esta ferramenta de elevado valor operacional e estratégico carece de alguma operacionalidade no que toca à extracção e tratamento dos dados. Assim, afigurou-se necessário conhecer e estudar sistemas automáticos de tratamento de dados, capazes de produzir conhecimento a partir de dados latentes ou não, nesta base de dados (BD). Ao longo deste trabalho são referidos exemplos de ferramentas de *Data Mining* (DM), algumas com aplicação empresarial e outras com aplicações policiais. É neste último contexto que será investida a investigação, com a finalidade de abrir novas portas em termos de tecnologia, numa tentativa de acompanhar a era demarcadamente informacional e tecnológica em que se encontra a sociedade contemporânea.

Os actuais responsáveis pelo SEI, estão a estudar a aplicação de métodos como o *Data Mining* ao sistema, de forma a retirar deste, o máximo de informação possível. No entanto, existem vários factores limitadores desta vontade de evoluir os actuais sistemas de informação da PSP. Entre os quais encontram-se limitações a nível da qualidade dos dados, nível do monetário e ainda do próprio desenho do Sistema de Informação (Anexo III e IV).

Segundo B. Ewart (et al., 2004: 5)[2] este tipo de ferramentas de *Data Mining* podem gerar centenas de padrões[3] ou regras. Esta possibilidade, poderia dotar os órgãos decisores desta Polícia, de um nível de conhecimento muito mais completo e fiável, em particular, no apoio à decisão estratégica e operacional. Esta mesma ideia é transmitida por Jerry H. Ratcliffe (2007: 8) quando diz que *um produto de conhecimento informacional, é um produto que pode influenciar o pensamento do decisor*.

Por estas razões, e pelo facto de o *Data Mining* ser reconhecido como um processo de extracção de conhecimento eficiente e fidedigno, escolhe-se estudar este método de aquisição de conhecimento e de informação.

---

[1] *Etapa da Descoberta de Conhecimento, que encontra tendências e associações num grande volume de dados. Utiliza técnicas de aprendizagem automática.* (Santos & Azevedo**,** 2005: 175).
[2] Citando Han, J.; Kamber, M., (2001) *Data Mining: Concepts and Techniques.* Morgan Kaufmann.
[3] *Um padrão corresponde a um conjunto de relacionamentos não-lineares nos dados. Os padrões não devem ser confundidos com casualidade. Em DM é comum serem designados por modelos.* (Santos & Azevedo**,** 2005: 177).

Os objectivos a atingir na realização deste trabalho, passam por abordar em termos gerais os conceitos de: *Data Mining*, Informações, Conhecimento, Dados, Privacidade. Bem como, o estudo das vantagens e/ou desvantagens da implementação de ferramentas de *Data Mining* a bases de dados criminais, estabelecendo-se comparações com exemplos internacionais de outras Polícias e aferindo a opinião de Oficiais da PSP e de técnicos informáticos acerca da temática em questão.

O que se pretende alcançar com a realização deste trabalho, é a resposta à seguinte questão:

Serão as ferramentas de Data Mining (método de extracção de conhecimento) vantajosas, em termos operacionais e estratégicos, para a Polícia de Segurança Pública?

Para poder responder a esta pergunta, levantam-se as seguintes hipóteses:

1) A utilização de ferramentas de *Data Mining* por instituições policiais a nível internacional, ocorre (com bons resultados, com potencial e como um pilar decisório), principalmente, nos países tecnologicamente mais desenvolvidos;
2) Para a introdução de um sistema de *Data Mining*, sob o ponto de vista de infra-estrutura tecnológica, o actual *Data Center* do SEI, não requer qualquer tipo de alteração e/ou substituição por outras tecnologias ou infra-estruturas paralelas;
3) A qualidade dos dados existentes na base de dados do SEI, "possibilita" a introdução "directa" de uma ferramenta de *Data Mining* sem requer um tratamento exaustivo dos dados, a montante da sua introdução;
4) Parte dos Oficiais de polícia conhecem e consideram o *Data Mining* importante para a PSP[4].

Presume-se que uma ferramenta tecnológica de análise de informação, seria útil à PSP para se modernizar e para conseguir acompanhar, a cada vez mais complexa, criminalidade. Por esta razão considera-se que apesar de este ser um primeiro estudo (na vertente de dissertação de mestrado em Ciências Policiais) sobre o Data Mining, é importante desenvolver mais investigação (mesmo com recurso a parcerias) nesta área. Com o objectivo de tornar a PSP numa polícia informada, e acima de tudo, dominadora do conhecimento relativo aos fenómenos criminológicos, com os quais lida diariamente.

[4] Considerando a complexidade e a especificidade da matéria em estudo, os resultados serão considerados positivos, caso exista um conhecimento destas ferramentas de, pelo menos, 1 em cada 4 Oficiais.

Capítulo 1 - As informações e os sistemas de informação da Polícia de Segurança Pública

A Polícia de Segurança Pública (PSP) como Força de Segurança[5], tem necessariamente obrigações no domínio da Segurança Interna[6]. Por outras palavras, a PSP tem por missão *assegurar a legalidade democrática, garantir a segurança interna e os direitos dos cidadãos* (Artigo 272.º, n.º 1 da Constituição da República Portuguesa [CRP]). Esta missão, constitucionalmente consagrada, carece de alguns esclarecimentos.

Segundo Arnoldo Almeida (*et al.,* 2009:117-121), o Artigo 272.º da CRP, estabelece a existência de três dimensões de polícia. Nomeadamente, uma dimensão de polícia de ordem pública (*Polícia Administrativa Geral* [Dias, 2010:6; Clemente, 2009:105]), uma dimensão de polícia administrativa (*Polícia Administrativa Especial* [Dias, 2010:6; Clemente, 2009:105]) e uma dimensão de Polícia Judiciária.

A dimensão de polícia de ordem e tranquilidade públicas (Almeida et al., 2009: 117) têm como finalidades: 1) a protecção dos direitos liberdades e garantias dos cidadãos, através da prevenção de possíveis actos danosos aos mesmos e; 2) a garantida da manutenção da ordem e da paz pública, bem como da sua reposição, caso se afigure necessário para o normal funcionamento da sociedade.

A dimensão de polícia administrativa tem por base as actividades *de licenciamento, de fiscalização, de processamento e de sancionamento das normas jurídicas protectoras de bens jurídicos indignos de tutela penal* (Almeida et al., 2009: 118). Ou seja, é o *exercício de competências especializadas em razão da matéria* (Dias, 2010:6), baseando-se nos regimes jurídicos das contra-ordenações, intervindo assim, num âmbito de direito administrativo sancionatório e preventivo (Almeida et al., 2009: 118). Um exemplo de uma acção desenvolvida por esta polícia que se enquadra nesta dimensão, é a actividade de licenciamento e controlo do uso e porte de arma[7].

A dimensão de polícia judiciária (Almeida et al., 2009: 119) complementa as actividades de polícia administrativa, através da repressão da criminalidade e da prossecução de justiça *(A actividade dos órgãos de polícia criminal no processo penal, enquanto coadjutória das autoridades judiciárias e enquanto funcionalmente dirigidas às finalidades do processo penal, deve ser qualificada como actividade de Administração da*

[5] Cfr. Artigo 25.º da Lei 58/2008 de 29 de Agosto – Lei de Segurança Interna (LSI)
[6] Cfr. Artigo 272.º da Lei Constitucional 1/2005 de 12 de Agosto – Constituição da República Portuguesa (CRP); e Artigo 25.º, n.º 2 da LSI.
[7] Cfr. Lei n.º 5/2006 de 23 de Fevereiro – Aprova o novo regime jurídico das armas e suas munições.

Justiça [Silva, 2010: 296]), desenvolvendo respectivamente, *actos pré-processuais* (Almeida et al., 2009: 119) e/ou *actos processuais* (Silva, 2010: 294). Os primeiros são aqueles que se destinam à promoção da intervenção da autoridade judiciária, quer através de medidas cautelares e de polícia[8], quer pela figura da detenção[9], quer pela consequente aplicação do termo de identidade e residência[10] e da constituição de arguido[11]. Estes actos são considerados pré-processuais, pelo facto de serem levados a cabo imediatamente antes da existência de algum processo-crime, ou seja, são *actos praticados fora do processo, sem a direcção das entidades competentes para o inquérito ou a instrução,* logo*, não são actos processuais* (Silva, 2010:293). No entanto, carecem de aprovação e validação posterior por parte da autoridade judiciária competente (Almeida et al., 2009: 120). Por seu turno, os actos processuais, são os actos desenvolvidos pelo Órgão de Polícia Criminal (OPC) ou Autoridade de Polícia Criminal (APC), no qual o OPC é dotado de competências legais e legítimas para poder efectuar as diligências[12] delegadas pela Autoridade Judiciária (*A polícia* [OPC] *desenvolve actos próprios do poder judicial – apresentando-se desta feita como operador de justiça – não como consequência própria, mas sim indirecta, delegada ou deferida.* [Almeida et al., 2009: 120]), a fim de apurar a verdade material dos factos praticados[13] (Almeida et al., 2009: 120).

Tendo em conta o explanado anteriormente, a PSP como *uma força de segurança, uniformizada e armada, com natureza de serviço público* (Artigo 1.º n.º 1, da Lei 53/2007 de 31 de Agosto – Lei Orgânica da Polícia de Segurança Pública [LOPSP]) tem a obrigação de zelar pela segurança interna[14]. Desta imposição legislativa, pode entender-se que, uma das principais funções da PSP é a prevenção criminal. É nesta valência que esta dissertação de mestrado terá incidência, especialmente na prevenção do crime com base nas informações. Informações essas, que no subcapítulo seguinte serão abordadas, de forma a esclarecer alguns conceitos subjacentes a esta temática. Seguidamente, serão explanados alguns conceitos importantes sobre SI, de modo a que seja desenvolvido, em termos gerais (macro), um entendimento acerca deste tipo de tecnologias.

[8] Cfr. Artigo 248.º e seguintes do Código de Processo Penal (CPP).
[9] Cfr. Artigo 254.º e seguintes do CPP.
[10] Cfr. Artigo 61.º, n.º 3 al.c); e Artigo 196.º do CPP.
[11] Cfr. Artigo 58.º e seguintes do CPP.
[12] *As polícias são os … olhos e as … mãos* da autoridade judiciária (Almeida et al., 2009: 120).
[13] Cfr. Artigos 263.º; 270.º; 290.º do CPP.
[14] *"A segurança interna é a actividade desenvolvida pelo Estado para garantir a ordem, a segurança e a tranquilidade públicas, proteger pessoas e bens, prevenir e reprimir a criminalidade e contribuir para assegurar o normal funcionamento das instituições democráticas, o regular exercício dos direitos, liberdades e garantias fundamentais dos cidadãos e o respeito pela legalidade democrática".*(Artigo 1.º n.º 1 da Lei 53/2008 de 29 de Agosto – Lei de Segurança Interna [LSI])

1.1 - INFORMAÇÕES POLICIAIS

A PSP como sendo *a agência mais visível de controlo social coactivo* (Clemente, 2009: 91), tem que procurar "*o ponto de equilíbrio entre a desordem suportável e a ordem indispensável* (Clemente, 2009: 91). Por esta razão, e pelo facto de o *recuo da prática criminal depender cada vez mais da eficácia da actuação policial, guiada pelas informações* (Clemente, 2009: 92), é um imperativo policial a antecipação do risco por parte da prevenção criminal (Clemente, 2009: 93). Entende-se assim que, *a actividade policial de informações,* (…) *ao ser antecipadora,* (…) *protege a colectividade de perigos concretos* (Clemente, 2009: 100). A esta antecipação do risco pode-se designar de previsão[15].

A prevenção criminal, através das informações, requer um grande esforço de *recolha e processamento de notícias com interesse para a missão policial, (...)* permitindo *avaliar riscos específicos e orientar a acção operacional. O conhecimento de intenções ou factos favorece a gestão do risco e, por conseguinte, a tomada de providências eficazes, como o reforço do patrulhamento em certos eventos* (Clemente, 2009: 93).

Esta antecipação do risco crê-se possível através da análise da informação existente em bases de dados criminais (como falar-se-á mais adiante).

Os termos, informação e informações, podem ser por vezes confundidos e até entendidos como sendo sinónimos, por esta razão passar-se-á à explicação conceptual dos mesmos.

Segundo Paulo João (2009: 936) o termo informação está relacionado com o acto de *ficar informado* com consequente *redução da ignorância e incerteza.* Este autor, baseando-se no *The Oxford English Dictionary*, refere a existência de três campos de utilização da palavra informação:

a) *Como Processo – Quando alguém é informado, altera-se o que essa pessoa sabia. Neste sentido, "informação" é o "acto de informar", a comunicação de conhecimento ou a acção de dizer ou de nos ser dito qualquer coisa;* (João, 2009: 936)
b) *Como Conhecimento – Utiliza-se também a designação "informação" para referir o que é entendido na "informação como processo" o "conhecimento comunicado sobre um dado*

[15] *"A previsão é uma prevenção da prevenção."* (Clemente, 2009: 99)

> *facto, assunto, ou acontecimento em particular; aquilo que cada um é informado; A noção de informação como algo que reduz a incerteza pode ser considerada como um caso especial da "informação como conhecimento". No entanto, por vezes, a comunicação de informação contribui para o aumento da incerteza;* (João, 2009: 936/937)

c) *Como Coisa*[16] *– O termo "informação" usa-se também para qualificar objectos, tais como dados e documentos, na medida em que se consideram estes como sendo "informativos" ou seja como tendo a propriedade de "fazer saber" conhecimento ou de comunicarem informação.* (João, 2009: 937)

Tendo em consideração o explanado anteriormente, parece continuar a existir algumas dúvidas entre informação e informações, uma vez que a informação como conhecimento pode traduzir-se em informações.

Assim, e seguindo a linha de pensamento de Pedro Clemente (2009: 98), informação é *o conjunto de dados colocados num* determinado *contexto, relacionado com o espaço, o tempo,* e *o cenário de acção*. Sendo que informações são o resultado da *análise interpretativa de notícias*, ou seja, informações são o produto da análise da informação.

Este mesmo autor, faz uma distinção interessante entre estes termos, recorrendo para tal, aos anglo-saxónicos, no qual refere que:

> *(...) As informações assumem sempre a denominação de inteligência, quando haja, pelo menos, a análise da informação obtida pela pesquisa, a partir de fontes cobertas ou não-abertas. Em Portugal, a tradição favorece a palavra informações, como conhecimento profundo e integrado, englobando o significado de inteligência* (…) (Clemente, 2006: 395).

Sobre esta matéria, Jerry H. Ratcliffe (2007: 8) defende que, cada organização tem a sua definição de informações (*intelligence*)[17], facto este que ocorre por diversas razões. As necessidades de informações de uma grande agência federal, são bastante diferentes, das necessidades de uma polícia pequena/rural. Mesmo entre departamentos metropolitanos (subentende-se Comandos Metropolitanos) podem diferir, quer no âmbito do entusiasmo (de cada Comando) em ser guiados pelas informações (*Intelligence-Led*

[16] *"Os sistemas de informação estudam principalmente a "informação com coisa"* (João, 2009: 937)
[17] Inteligência.

Policing[18]), quer na tipologia de problemas com os quais têm que lidar. Até mesmo, quando diferentes cidades têm diferentes problemas, as organizações destas, respondem frequentemente, de modo diferente, em virtude de, terem diferentes necessidades informacionais.

Denote-se que, *um produto de conhecimento informacional pode até ser uma breve conversação telefónica, se a informação transmitida em tempo real tiver um efeito no processo de decisão do receptor da informação* (Ratcliffe, 2007: 8).

Assim, para a elaboração e melhor entendimento desta dissertação, considera-se que: informação – são os dados e/ou documentos (João, 2009: 937); e informações – são o produto da análise da informação (Clemente, 2006: 395), independentemente do formato de apresentação dos seus resultados.

Segundo Pedro Clemente (2006: 394), Portugal e qualquer outro Estado, necessita de informações capazes de assegurar não só o normal funcionamento das instituições, como a legalidade e ainda a defesa dos direitos dos cidadãos. Este autor refere ainda que, as unidades de informações das forças policiais, não se podem resumir a *um mero banco de dados*, uma vez que a *actuação policial precisa de ser proactiva e antecipadora,* e no mínimo*, redutora das oportunidades de produção de incivilidades lesivas da vida social. Sem informações, a Polícia é cega, logo inoperante* (Clemente, 2006: 384).

Segundo José Torres (cit. in. Clemente, 2009: 99), *as informações policiais são todas aquelas destinadas à prossecução directa das missões legalmente atribuídas a serviços de natureza policial, sejam elas de nível estratégico ou operativo.* É com esta finalidade que, actualmente, *as informações policiais contribuem de modo significativo, para a proactividade e eficácia da acção policial, tanto na manutenção da ordem pública, como na resposta reactiva à prática criminal* (Clemente, 2006: 386).

Assim, Pedro Clemente (2009: 99 e 2006: 399) decompõem as informações policiais em: Informações de ordem pública – *visando prevenir incidentes de ordem pública e acautelar a ocorrência de incivilidades, especialmente a produção de delitos criminais;* Informações criminais[19] – *que se inserem no âmbito da actividade criminal reportada à investigação criminal;* e Contra-informações – *que visam impedir a realização de acções de recolha indevida da informação sigilosa, (...)* através *da aplicação de*

[18] Policiamento Orientado pelas Informações.
[19] … *Informações criminais é a criação de um produto de conhecimento informacional que ajuda a tomada de decisão nas áreas de: Law enforcement, redução do crime e prevenção criminal.* (Ratcliffe, 2007: 8)

medidas de segurança passiva aos documento, classificando-os, e do controlo de acessos aos mesmos, por pessoal credenciado.

Sobre informações criminais, Jerry Ratcliffe (2007: 8-9) admite a existência de três entendimentos dominantes, no que toca às finalidades das mesmas. Primeiramente, refere que, muitas pessoas ligadas ao mundo policial vêm as informações criminais, somente como, um mecanismo para examinar o comportamento de indivíduos agressores, ou de grupos criminosos organizados. Logo, é separada, à partida, das suas principais funcionalidades ao nível do patrulhamento e da repressão criminal. Esta é uma realidade em crescimento, contudo, ao nível estratégico e operacional das polícias, um entendimento mais profundo da criminalidade em geral pode ser necessária, ou seja, muito para além do simples entendimento do comportamento do indivíduo agressor (Ratcliffe, 2007: 8). Assim, nesta definição, a tónica está nas informações que apoiam a redução do crime, a prevenção e a repressão (*Enforcement)*[20] (Ratcliffe, 2007: 9).

Por outro lado, as informações criminais são entendidas como instrumento que apenas apoia os órgãos decisores da polícia. Este facto, pode levar uma restrição na partilha de informações úteis. Ou seja, as informações apenas seriam partilhadas com os oficiais da confiança do órgão decisor. Por outras palavras, com esta ideologia, ao mencionar apenas os decisores, corre-se o risco de as informações, permanecerem num núcleo restrito de pessoas, perdendo-se a principal função destas, que é dotar de conhecimento (Ratcliffe, 2007: 9).

Por último, este autor reconhece que, alguns estudiosos entendem que, o único objectivo das informações criminais é, o uso destas para (por exemplo) efectivar uma detenção ou fornecer provas contra um determinado agressor. Contudo, boas informações criminais, podem ainda ajudar o reforço das leis criminais, e também sugerir estratégias para redução da criminalidade, ou seja, do fluxo criminal. Ou ainda, fornecer um rumo de prevenção da agressão, que conjuntamente com determinadas tácticas de policiamento orientado para o problema, possibilitam a prevenção situacional do crime, como por exemplo, a prevenção criminal pelo ornamento/design urbanístico e ambiental (Ratcliffe, 2007: 9).

[20] Actividade energética; superar através da força; influência constrangedora; convencer o cumprimento de; uma sanção (Oxford English Dictionary, 2009).

Pedro Clemente (2006: 395), acrescenta ainda que, a produção de informações policiais, obedece a um ciclo contínuo e dinâmico, com diversas fases e etapas, as quais estão representadas na figura 1.

Na opinião do presente autor, *nenhum futuro é cópia exacta do passado* no entanto, é através do estudo deste, que é possível entender fenómenos criminais. É aqui que, a análise da informação, como *um elemento decisivo no processo de produção de informações,* tem o objectivo principal de compreender determinado fenómeno, bem como, o de *prospectivar as tendências de evolução* deste (Clemente, 2006: 395).

Figura 1 – Ciclo das Informações Policiais

Fonte: Adaptado de **CLEMENTE**, Pedro (2006), *As informações Policiais – Palimpsesto* in *Estudos de Homenagem ao Juiz Conselheiro António da Costa Neves Ribeiro* (2007). Edições Almedina, SA.

Assim, é possível dizer que *a qualidade da actuação policial depende muito da produção de informações*. *O conhecimento guia a Polícia em prol da felicidade pública*, ou seja, a *inteligência policial serve o cidadão, protegendo-o de riscos maiores* (Clemente, 2006: 400).

Em suma, a acção preventiva dos corpos policiais, ao ser complementada com a sua tarefa de previsão (ou seja, antecipando a prevenção através da produção de informações, seguido da exploração dos produtos informacionais), traduz-se num policiamento mais eficaz e eficiente, onde as operações materiais de vigilância (visíveis no patrulhamento

[auto, equestre e apeado] da via pública, por agentes fardados), são muito mais do que patrulhas em giro aleatório (Clemente, 2009: 92, 100 e 104).

1.2 - SISTEMAS DE INFORMAÇÃO

Actualmente, a PSP está equipada com um sistema de informação denominado SEI. Conceptualmente, sistema de informação (SI), pode ser definido como um sistema tecnológico que manipula, armazena, processa e dissemina informação (Karwowski, et al., 2002: 185). É expectável que esta informação tenha um impacto no comportamento humano, sendo organizada mediante um determinado contexto de utilização. Os sistemas de informação são artefactos cognitivos poderosos e externos, que aumentam o nível máximo das capacidades cognitivas do Homem. Tais capacidades incluem: a codificação e a descodificação da informação, o armazenamento e procura, bem como, a recuperação e partilha da informação. Outras tarefas humanas cognitivas que são facilitadas pelo uso de SI são: o raciocínio, o pensamento, a aprendizagem e a resolução de problemas (Karwowski, et al., 2002: 185). Tipicamente, os sistemas de informação são usados para suportar tarefas de processamento de informação, que vão desde as mais simples[21], até à tarefa mais complexa imaginável. Qualquer que sejam as funções que estes oferecem ou apoiam, os sistemas de informação, são primordialmente construídos, para fornecer serviços de informação, a uma classe particular de utilizadores, de forma a satisfazer as suas necessidades informacionais (Karwowski, et al., 2002: 185).

A principal actividade desenvolvida pela interacção dos utilizadores com os sistemas de informação, é a tarefa de procura de informação, a qual é um processo guiado por um problema de informação[22]. A abordagem de um problema informacional, tem que ser iniciado por alguém activamente consciencializado da necessidade de alcançar um objectivo. Uma pessoa envolvida na procura de uma solução para um problema informacional tem um ou vários objectivos em mente, e utiliza um sistema de informação como uma ferramenta de suporte, de forma a atingir os seus objectivos. O principal objectivo do SI é providenciar a informação necessária ao potencial utilizador, da forma mais eficaz e eficiente possível. Logo, o sistema de informação tem que ser desenhado,

[21] Por exemplo, uma actividade de recuperação de informação (Karwowski, et al., 2002: 185).
[22] Exemplos de problemas de informação são: a monitorização de um determinado estado ou situação, a obtenção de informação para tomar uma decisão, a procura por um alojamento, a marcação de uma viagem de avião, a manutenção das actualizações informacionais acerca de um competidor empresarial, a satisfação da curiosidade, a publicação ou elaboração de um artigo, ou a investigação de uma nova área (Karwowski, et al., 2002: 186).

desenvolvido e utilizado, com a devida consideração pela percepção cognitiva humana e pelas capacidades, limitações e necessidades emocionais destes[23] (Karwowski, et al., 2002: 186).

Actualmente, a PSP, é ajudada por um sistema de informação denominado, Sistema Estratégico de Informação (SEI). Este sistema foi estudado a partir de 2002, sendo que, a sua implementação, apenas se deu em 2004 (aquando do Euro 2004). A PSP em conjunto com a empresa Accenture, elaborou um Plano de Estratégico de Sistemas de Informação, onde o objectivo foi elencar e definir um conjunto de necessidades desta polícia, e a partir daí, elaborar um sistema informacional capaz que responder a essas necessidades (PESI: 2002)[24]. Neste momento a PSP detém três servidores em ambiente de produção, e um servidor de base de dados. Os primeiros utilizam um software, que é o JBoss, e o segundo utiliza um software denominado SQLServer 2005. Existe ainda uma aplicação da gestão dos perfis e das permissões o ITIM. No entanto, está planeado para 2011 passar a ter para uma outra, a UTIS[25] que a Rede Nacional de Segurança Interna (RNSI) disponibilizou para todas as forças (Anexo IV).

Este sistema (SEI), foi implementado com a finalidade de dotar a PSP, de um sistema de introdução e armazenamento de dados. Ou seja, o SEI é um repositório de informação[26] (Anexo III e IV). Por esta razão, e pelo facto de ter sido concebido para este efeito, torna-se inviável a extracção de qualquer conhecimento. Actualmente, existe *um procedimento que vai ao SEI (base de dados optimizada para introdução de dados), retira os dados um pouco mais estruturados do que estão no SEI, para uma base de dados intermédia,* e *com base nessa, é que são feitos os dados estatísticos.* (Anexo III). No entanto, pelo facto de esta base de dados intermédia não ser um Data Warehouse (DW), não é viável ou eficiente a manipulação/análise dos dados aí existentes (Anexo III). Por outras palavras, qualquer tentativa de análise requer um tratamento manual (implicando um esforço humano bastante considerável) ou computacional[27] dos dados.

[23] A disciplina científica que estuda a interacção entre as pessoas (necessidades humanas) e a tecnologia, é a ergonomia, também conhecida por *human factors* (factores humanos) (Karwowski, et al., 2002: 186).
[24] O facto de ter havido uma ambição inicial de dotar este SI, com capacidades úteis para todas as valências da PSP, levou a que, nenhum dos módulos criados, fica-se convenientemente desenvolvido (pelo menos a 100%) (Anexo III).
[25] Unidade de Tecnologias de Informação de Segurança (Fonte: http://www.rnsi.mai.gov.pt/Pages/defaultint.aspx)
[26] Também denominado de *Data Center* (Anexo III).
[27] Para tratar/limpar milhões de dados existentes nas base de dados, é necessário a utilizar ferramentas de limpeza de dados (*Data Cleaning*) (ver Capítulo 2).

Este sistema possibilita à PSP o registo de todos os ilícitos criminais que chegam ao seu conhecimento. No entanto, qualquer iniciativa de elaboração de informações estatísticas ou policiais, carece da utilização de outra base de dados e de outros métodos. Assim, são efectuadas estatísticas, como ferramenta de apoio à decisão estratégica e operacional[28]. De momento, e no que concerne à produção de informações policiais, os dados têm que ser trabalhados com recurso a um programa denominado I2[29], que interliga várias bases de dados, de entre as quais, o SEI. Este programa informático, produz associações e ligações entre os critérios de pesquisa de forma gráfica[30]. Esta ferramenta ajuda a actividade de produção de informações, no entanto, é totalmente elaborada manualmente. Ou seja, o I2 apenas consegue procurar e organizar os itens de interesse mediante as relações que estes têm com outros itens de interesse. Uma vez que, ao escolhermos um determinado suspeito ("Y"), introduzindo o seu nome (como critério de pesquisa), o programa apresenta "X" ligações, mas se o objectivo for explorar cada uma destas ligações, é necessário repetir todo o processo, até se obter a quantidade de ligações desejada. Logo, não existe nenhuma ferramenta que elabore um diagrama completo de associações, autónoma e automaticamente[31]. Este programa apenas faz a apresentação/representação das ligações existentes para uma determinada pessoa, com base nos dados existentes em várias bases de dados (Anexo III). Assim, pode-se verificar que o SEI, por si próprio, não consegue produzir informações, sejam elas estatísticas ou policiais uma vez que não foi desenhado para tal.

Recentemente foi criada a Equipa Única do SEI (EUSEI)[32]. Esta equipa tem como principais funções, a supervisão do SEI, o estabelecimento de prioridades de desenvolvimento e de evolução do SEI, o supervisionamento da qualidade dos dados, entre outras. Por outras palavras, esta equipa única gere o SEI a todos os níveis, mas está fora dos Departamentos (da DN) e é transversal a toda da DN e a todos os Comandos,

[28] *Na parte da estatística, fizemos entre Dezembro de 2009/Janeiro de 2010, um relatório especial de informações onde já utilizámos algumas técnicas de estatística multivariada*, [para tentarmos] *traçar alguns perfis criminais, suspeitos e vitimas*. (Anexo III).
[29] Para mais detalhe, vide: http://www.i2group.com/us.
[30] Ao introduzir um critério de pesquisa, *por exemplo, o nome de um suspeito* o I2, *vai criar-nos de forma gráfica, todas as ligações que existem desse suspeito* e *todos os itens de interesse que estão no SEI* (*por exemplo, todos os veículos que estão relacionados com aquele suspeito no SEI*). *Se quisermos saber quem mais é que está relacionado com aquele veiculo, dizemos ao programa e ele vai-nos buscar todos os outros suspeitos e assim conseguimos criar um diagrama de associações*, o qual, *é muito utilizado diariamente pelo Departamento de Informações na analise das informações de declaração.* (Anexo III).
[31] Ou seja, todo este processo *é feito à mão, (...) não existe nada que o faça de forma automatizada* (Anexo III). Os analistas do Departamento de Informações Policiais é que fazem, manualmente, a recolha e a análise da informação (Anexo III).
[32] Criada pela Norma de Execução Permanente N.º DN/ASDDN/GEO/03/01 de 24 de Janeiro de 2011.

dependendo directamente do Director Nacional Adjunto de Operações e Segurança (Anexo III).

Segundo Carlota Fernandes, o SEI ainda não chegou ao nível da gestão pró-activa (à possibilidade de se poder prever algo no SEI). Ou seja, num determinado período, numa zona, e com os tipos de ocorrências, prever uma actividade operacional mediante os factos. Isto porque a PSP não tem ferramentas que permitam às pessoas explorar a informação a esse nível (Anexo IV).

No entanto é necessário realçar a importante ajuda que este representa para a actuação policial (Anexo III). Assim, reitera-se a ideia de que, *o recurso às novas tecnologias de informação favorece a previsão e a contenção de comportamentos ilícitos* (Clemente, 2009: 96).

CAPÍTULO 2 - DA INFORMAÇÃO AO CONHECIMENTO

O termo *Data Mining* (DM), é utilizado para descrever o processo computacional de gerar ou extrair conhecimento a partir de grandes volumes de dados (Han & Kamber, 2006: 5). A comunidade científica portuguesa mantém o termo original em inglês (*Data Mining)*, por isso, optou-se por mantê-lo ao longo de todo o trabalho, no entanto e a título de exemplo, a comunidade científica brasileira traduz este termo para *Mineração de Dados*[33].

Segundo Han & Kamber (2006: 5), o termo DM não caracteriza todo o processo de extracção de conhecimento[34], uma vez que, a designação apropriada deveria ser: extracção de conhecimento a partir de dados (*knowledge mining from data*). No entanto, este termo, tido como correcto para os autores, mesmo sendo o mais correcto, é demasiado extenso, por isso convencionou-se o termo *Data Mining*.

Estes autores reconhecem várias abordagens, admitindo que muitos autores consideram que, o *Data Mining* é sinónimo de outro termo muito usado na comunidade de Tecnologias de Informação (TI), nomeadamente, a Descoberta de Conhecimento em Bases de Dados[35] (DCBD). Por outro lado, outros defendem que o *Data Mining* é um dos passos principais em todo o processo de descoberta de conhecimento. Ou seja, é uma fase do processo de DCBD (Han & Kamber, 2006: 7). Assim, o *Data Mining* é o "coração" do processo de DCBD, que compreende a aplicação de algoritmos[36] que exploram os dados, desenvolvendo modelos e descobrindo padrões previamente desconhecidos. Os modelos descobertos são usados para compreender fenómenos (da vida real), a partir dos dados, analisando-os e prevendo-os. Actualmente, a acessibilidade e a abundância de dados, faz da DCBD e do *Data Mining*, um assunto de considerável importância, sendo necessário estuda-los e entende-los (Maimon & Rokach, 2010: 1).

Com o recente crescimento deste campo de investigação das TI, verifica-se um grande aumento da variedade de métodos, agora acessíveis a profissionais e a analistas. Em todo o caso, nenhum método é superior a outro (Maimon & Rokach, 2010: 1). Os métodos desenvolvidos antes da revolução da internet, consideravam apenas pequenos grupos de

[33] *A mineração de dados (data mining) é o processo de extrair informação de coleções de dados*. (Broolshear 2003: 359)
[34] Os autores fazem uma analogia com o termo *gold mining* (extracção de ouro) onde a ênfase é dada ao ouro e não à pedra ou à areia. Nesta perspectiva, o termo correcto deveria dar ênfase ao conhecimento e não aos dados.
[35] *Knowledge Discovery from Data* (KDD).
[36] *Fórmulas matemáticas complexas, são a parte fundamental das ferramentas de Data Mining* (Santos & Azevedo, 2005: 173).

dados, com pouca mutabilidade, quer em termos de tipologia de dados, quer termos de fiabilidade. A época informacional em que vivemos, possibilita que a acumulação de dados se torne cada vez mais fácil e mais barata[37]. No entanto, à medida que a quantidade de informação armazenada electronicamente aumenta, a nossa capacidade de a entender, e de lhe dar algum uso, é cada vez menor, logo, é impossível acompanhar o seu crescimento (Maimon & Rokach, 2010: 2). O *Data Mining* é um termo arquitectado para, descrever o processo de procura e pesquisa de padrões e relações interessantes, em grandes bases de dados. A quantidade e a disponibilidade de dados, está a aumentar exponencialmente, apesar de, o nível de processamento humano ser, na maioria dos casos, constante. Assim, o vazio criado pela incapacidade humana de analisar milhões de dados, é uma oportunidade para o campo da DCBD/DM, se revelar uma peça especialmente importante e necessária para completar o conhecimento humano (Maimon & Rokach, 2010: 2). De modo a fazer dos grandes volumes de dados uma fonte de conhecimento útil e com aplicabilidade em diversos domínios (Santos & Azevedo**,** 2005: 7).

O capítulo seguinte incorporará diversas definições e conceitos. Os quais são importantes, para a compreensão dos diversos termos utilizados pelas mais variadas comunidades científicas, que de alguma forma, estudam e aprofundam esta temática.

2.1 - OS DADOS

Os dados podem apresentar diversos formatos, o que resulta, numa grande variedade de diferentes modelos de armazenamento (Cios, et al., 2007: 27).

Ao nível mais elementar, uma única unidade de informação é o valor de um atributo, onde cada atributo pode ter um determinado número de diferentes valores. Os objectos descritos por atributos são combinados para formar conjuntos de dados, que por sua vez, estão armazenados como numa BD ou DW (Cios, et al., 2007: 27).

Existem dois tipos de valores: os numéricos e os simbólicos. Os valores numéricos são expressados por números (por exemplo: números reais [-1.09, 123.5]; números inteiros [1, 44, 125]; e números primos [1, 3, 5]). Por outro lado, os valores simbólicos, normalmente, descrevem conceitos qualitativos (por exemplo: cores [vermelho, branco] ou tamanhos [pequenos, médios, grandes]) (Cios, et al., 2007: 27).

Os atributos são descritos por um conjunto de valores correspondentes. Ou seja, por valores numéricos e simbólicos, os quais podem ser discretos (categóricos) ou contínuos.

[37] Foi estimado que a quantidade de informação armazenada duplica a cada 20 meses (MAIMON & ROKACH, 2010: 2).

Atributos são discretos quando o número total de valores é relativamente pequeno (finito). Enquanto que, os atributos são contínuos se o número total dos valores é muito grande (infinito) e corresponde a um intervalo especifico (série). (Cios, et al., 2007: 27)

Os valores de um determinado atributo podem ser organizados em conjuntos, vectores ou matrizes. Esta categorização dos dados é importante por diversas razões[38].

Os objectos (também conhecidos como registos, exemplos, unidades, casos, indivíduos, e pontos de dados) representam entidades descritas por um ou mais atributos. Quando o objecto é descrito por vários atributos utiliza-se o termo: dados multi-variados. Por outro lado, quando apenas um atributo descreve o objecto, diz-se que este detém, dados uni-variados (Cios, et al., 2007: 28).

Considerando os pacientes numa clínica de doenças cardiovasculares como exemplo, um paciente é o objecto, que pode ser descrito por um número de atributos, tais como, nome, sexo, idade, resultados dos testes de diagnóstico (pressão arterial, nível de colesterol e avaliações qualitativas como dores no peito e o seu tipo gravidade). Um assunto importante, do ponto de vista do processo de descoberta de conhecimento, é saber como manipular diferentes tipos de atributos e valores. Em particular, qualquer operação em múltiplos objectos, como a comparação do valor dos atributos ou a sua distância computacional, deve ser cuidadosamente analisada e desenhada. Por exemplo, o valor simbólico "dois", normalmente não pode ser comparado com o valor numérico "2", a não ser que conversões sejam realizadas. Embora, a computação da distância entre dois valores numéricos seja simples, realizar a mesma computação entre dois valores nominais (tais como, "branco" e "vermelho", ou "dores de peito tipo 1" e "dores de peito tipo 4") requer especial atenção. Por outras palavras, como é possível medir a distância entre cores (poderia ser feito através do uso de diagramas de cromaticidade) ou entre tipos de dores de peito (seria bastante difícil)? Em certos casos pode ser impossível calcular tal distância. (Cios, et al., 2007: 28)

Um problema importante, no que se refere aos dados, é a limitada compreensão de números por parte dos humanos, que são os utilizadores por excelência do conhecimento gerado. Por exemplo, a maioria das pessoas não compreenderá um valor de colesterol de "331,2", enquanto que estas entendem facilmente o seu significado quando este valor numérico é expressado em termos de informação agregada, como nível "alto" ou "baixo"

[38] Alguns métodos de pré-processamento e de DM, só são aplicáveis a dados descritos como atributos discretos. Nesses casos, um processo chamado discretização torna-se um passo necessário do pré-processamento, para transformar atributos contínuos em atributos discretos, e este passo tem que ser completado antes da concretização da fase do DM (Cios, et al., 2007: 28).

de colesterol. Por outras palavras, informação é muitas vezes "granulada" e representa um grande nível de abstracção (agregação). Da mesma maneira que as operações e as relações entre atributos pode ser quantificada num nível de agregação. Em geral, a granulação da informação significa a encapsulamento de valores numéricos em entidades conceptuais únicas. Como por exemplo, o encapsulamento de elementos por conjuntos, ou o encapsulamento de intervalos por números. Entender o conceito de encapsulamento, também referido como janela de conhecimento, é muito importante no quadro da descoberta de conhecimento. Continuando com o exemplo do colesterol, nós podemos estar satisfeitos com um único valor numérico, como 331,2, que expressa o maior nível de granularidade. Alternativamente, nós podemos querer definir este valor como pertencente a um intervalo (300,400), o outro nível de baixa granularidade. (Cios, 2007: 29).

2.1.1 - DADOS DINÂMICOS

As bases de dados, são muitas vezes de natureza dinâmica (exemplo: SEI), onde novos objectos/atributos podem ser adicionados, retirados, ou ainda substituídos por novos dados (Cios, et al., 2007: 40). Neste caso, os algoritmos de DM devem relacionar-se com o tempo, ou seja, o conhecimento transmitido até determinado ponto deve ser incrementalmente actualizado. Segundo Krzysztof Cios (et al., 2007: 40) a diferença entre um algoritmo de DM incremental e um não incremental, é que o primeiro, utiliza o conhecimento previamente gerado e os dados (tanto os novos como os antigos), para gerar um novo conhecimento, enquanto que, o segundo apenas utiliza os novos dados, conjuntamente com os dados já existentes. O principal desafio dos métodos de DM incremental é fundir o conhecimento gerado pelos novos dados, com o conhecimento anterior (proveniente dos dados já existentes). A fusão pode ser tão simples, quanto adicionar o novo conhecimento ao conhecimento existente, mas na maioria dos casos, requer a modificação do conhecimento existente, conservando a sua consistência (Cios, et al., 2007: 40).

Estas bases de dados relacionais, são os repositórios de dados, dos sistemas operacionais das organizações. Estes sistemas processam as transacções que fazem trabalhar as organizações. Os dados destes sistemas são, por natureza, transitórios e em constante acumulação no repositório. Um exemplo típico destes sistemas, são os sistemas de processamento de transacções bancárias de qualquer banco, onde são guardados registos de: abertura e fecho de contas bancárias, depósitos, levantamentos, balanços, e outros registos relacionados com transacções de dinheiro entre contas e clientes. Os dados

passíveis de serem extraídos destes sistemas operacionais, apresentam-se, na sua forma mais rude (no sentido em que não foram transformados, limpos ou alterados). Estes podem conter diversos erros[39], que, normalmente, estão espalhados por várias tabelas e ficheiros. (Chakrabarti, 2009: 37).

2.1.2 - QUALIDADE DOS DADOS

Muitas bases de dados, são atormentadas pelo problema de dados perdidos ou inexistentes. Este facto, pode resultar de introduções manuais incompletas ou mesmo de erros do equipamento. Tais valores são normalmente conotados como valores "NULL", "*", e "?". Em alguns domínios (por exemplo: na medicina), é comum encontrar dados, com grandes percentagens de dados inexistentes[40]. Para lidar com este "não valores", existem dois grandes métodos: a remoção[41] e a imputação[42] (Cios, et al., 2007: 41-42).

O ruído nos dados é definido como um erro no atributo medido (Cios, et al., 2007: 42). Ou seja, quando se diz que existe ruído na base de dados, pode significar que, os dados registados são válidos para o conjunto de dados mas estão incorrectamente registados[43], ou então, que os dados são inválidos[44] para o conjunto de dados (Bramer, 2007: 15). Dependendo da quantidade de dados, o ruído pode ser considerado um problema substancial, podendo mesmo, prejudicar o processo de descoberta de conhecimento[45]. A influência do ruído nos dados pode ser prevenida através da imposição de restrições nos atributos, de modo a detectar anomalias quando os dados são inseridos. Quando o ruído já está presente, este pode ser removido através do uso de determinados métodos, como por exemplo: a inspecção manual com o uso de restrições predefinidas nos

[39] Mesmo quando os dados estão na sua forma mais elementar/original, não se pode assumir que este não conte erros. No mundo real, valores erróneos nas BD, podem ser registados por uma variedade de razões, incluindo os erros de medição, pensamentos subjectivos ou mau funcionamento ou mau uso de equipamento de registo automático. (Bramer, 2007: 15)

[40] Acima dos 50%. (Cios, 2007: 41).

[41] Os objectos e/ou atributos com valores em falta, são simplesmente descartados (rejeitados). Esta abordagem apenas é eficiente, quando os atributos removidos, não são cruciais para a análise, uma vez que da sua remoção resultaria na diminuição da informação contida nos dados (Cios, et al., 2007: 41).

[42] Resume-se ao uso de diferentes algoritmos os quais desenvolvem imputações únicas ou múltiplas. Nas únicas o valor é imputado a partir de um único valor. Por outro lado, na imputação múltipla, várias opções de imputação são geradas, sendo ordenadas por ordem probabilística a fim de ser escolhida a melhor opção (Cios, et al., 2007: 42).

[43] Por exemplo: "56.81", acidentalmente pode ser inserido como "5.681", ou um atributo categórico dever ser "preto" e acidentalmente ser registado como "vermelho", o ruído deste tipo não inválida por completo o atributo, apenas é um valor errado (Bramer, 2007: 15).

[44] Por exemplo: se se introduzir "56.8X" em vez de "56.81" ou "vemelho" em vez de "vermelho" os atributos são considerados inválidos. Um atributo inválido pode ser facilmente detectado, bem como, corrigido ou rejeitado (Bramer, 2007: 15).

[45] Este é um problema perpétuo nos dados do mundo real (Bramer, 2007: 15).

valores dos atributos[46], *binning*[47], e a segmentação[48] (*clustering*) (Cios, et al., 2007: 42-44). Ver erros, mesmo que óbvios, nos valores de uma variável, quando estes estão "enterrados" em 100.000 valores é uma tarefa complicada e morosa. Ainda assim, as tentativas de limpar os dados, ajudam, especialmente, a ter uma visão global dos dados, e a detectar dados anómalos ou concentrações imprevistas de valores nos atributos (Bramer, 2007: 15).

O fenómeno dos dados incompletos, manifesta-se quando a variável não contém informação suficiente para descobrir o novo e desejado, conhecimento. Para detectar os dados incompletos, o utilizador tem que analisar os dados existentes, e determinar se os objectos e os atributos existentes, dão riqueza de representação ao problema a estudar. Um sinal comum de incompletude é quando o conhecimento gerado, é de baixa qualidade e não pode ser melhorado através do uso de outros métodos de DM. A maneira mais óbvia para melhorar os dados incompletos, é reunir e registar dados adicionais, como novos atributos ou objectos (Cios, et al., 2007: 40).

A redundância dos dados, pode revelar-se pela existência de dois ou mais atributos fortemente relacionados, ou por valores/atributos repetidos. Nalguns casos, contudo, os dados redundantes podem revelar informação útil[49]. Os dados irrelevantes criam redundância quando objectos ou atributos são insignificantes para a análise. Para a identificação de dados redundantes podem ser utilizados algoritmos de extracção ou proceder-se à selecção de atributos. Uma vez identificados, são removidos com vista a acelerar o tempo de processamento (Cios, et al., 2007: 41).

Os dados reais, muitas vezes incluem objectos de dados imprecisos[50] os quais podem levar a resultados enganosos (Cios, et al., 2007: 40; Seifert, 2007: 22). Todos os esforços de aquisição/reunião de dados, em determinada altura, sofrem problemas de precisão. Assegurar a precisão da informação pode requerer protocolos dispendiosos, os

[46] Na inspecção manual. o utilizador avalia os valores do atributo, confrontando-os contra restrições pré--definidas e remove manualmente, todos os valores que não satisfazem estas restrições. Por exemplo, um valor de colesterol de 45.0, o qual está fora do intervalo de aceitação pré-definido para este atributo ([50.0 a 600.0]), tem que ser removido (Cios, 2007: 43).

[47] Este método, primeiramente, ordena os valores do atributo com ruído e depois, substitui-os com os valores da média ou da mediana, para os bins pré-definidos. Como exemplo, e considerando como atributo, o nível de colesterol, com os seguintes valores: 45.0, 261,2, 331.2 e 407.5. Se o tamanho do bin for igual a dois, então dois bins são gerados. Para o bin 1 o valor médio é 153.1 e para o bin 2 é 369.4. Assim, os valores 45.0 e 261.2 seriam substituídos por 153.1 e os valores 331.2 e 407.5 por 369.4. Note-se que os dois novos valores estão dentro do intervalo aceitável (Cios, 2007: 44).

[48] A segmentação, encontra grupos de objectos similares, e simplesmente apaga (ou muda para valores desaparecidos) todos os valores que caiem fora dos segmentos (Cios, 2007: 44).

[49] Por outras palavras, a frequência de objectos idênticos pode fornecer informação útil (Cios, 2007: 41).

[50] Existe grande probabilidade de não se saber o valor exacto de um determinado teste médico, mas compreende-se o significado de "alto", "baixo", ou "médio" (Cios, et al., 2007: 40).

quais podem não ser rentáveis, caso os dados não tenham grande valor económico[51] (Seifert, 2007: 22).

2.2 - O DATA[52] MINING[53]

Actualmente, assiste-se a um aumento drástico de quantidade de informação armazenada, devendo-se principalmente, à constante diminuição do custo do armazenamento de dados. Assim, verifica-se um aumento da importância destes, bem como da quantidade de informação, *pois na directa proporção da sua quantidade existe uma porção de conhecimento que pode ser usado para optimizar as tomadas de decisões* (Santos & Azevedo**,** 2005: 7). O *Data Mining*, através do uso de algoritmos específicos ou de mecanismos de pesquisa, tenta descobrir padrões e tendências discerníveis nos dados, inferindo regras para os mesmos. Ou seja, a descoberta de conhecimento a partir dos dados tornou-se mais fácil (Fayyad et. al., 1996: 31). O processo de *Data Mining* fica responsável pela criação das hipóteses, dando assim, maior rapidez, aperfeiçoamento, autonomia e fiabilidade aos resultados. Neste seguimento, e segundo Fayyad (et al., 1996: 31), Maimon & Rokach (2010: 10) e Santos & Azevedo (2005: 10), o DM é uma das fazes do processo de Descoberta de Conhecimento em Bases de Dados[54] (DCBD).

Assim, e recorrendo à mais variada bibliografia na área, constata-se a existência de diversas definições, das quais se destacam as seguintes:

a) O *Data Mining* é a aplicação de algoritmos específicos para a extracção de padrões e modelos a partir dos dados (Fayyad et. al., 1996: 28);
b) O *Data Mining* é o processo de identificação de padrões, através do tratamento de grandes volumes de dados. Usando para tal, técnicas como a classificação, a previsão, a análise de clusters e a análise associativa (Holmes, et al., 2007: 330);
c) *Data Mining* é uma actividade de extracção de conhecimento, cujo objectivo é descobrir factos escondidos nas bases de dados. Noutras palavras, o DM envolve a análise sistemática de grandes conjuntos de dados usando métodos automáticos. (McCue, 2007: 25);

[51] A precisão de informação reunida através de um cartão de compras pode não ser precisa pelas seguintes razões: a falta de autenticação da identidade quando o cartão está a ser usado, o facto de as empregadas usarem os seus próprios cartões para clientes que não têm um, e/ou clientes que usam múltiplos cartões (Seifert, 2007: 22/23).
[52] *Data* deriva de *Datum* que significa algo assumido como facto; Um pressuposto ou uma premissa da qual são retiradas conclusões ou inferências. (Oxford English Dictionary, 2009)
[53] *Mining significa a acção do verbo Mine em vário sentidos. Mine significa extracção (de um item de valor) de uma fonte abundante.* (Oxford English Dictionary, 2009)
[54] Tópico desenvolvido com mais detalhe no Anexo VII.

d) *Data Mining* preocupa-se com o processo de extracção computacional de estruturas de conhecimento oculto, representadas em modelos e padrões, a partir de grandes repositórios de dados. É um campo de estudo interdisciplinar com raízes em bases de dados, estatística, aprendizagem automática e visualização de dados. (Gaber et. al., 2010: 759);

e) Data Mining é um processo que usa uma variedade de ferramentas de análise de dados, para descobrir padrões e relações nos dados, que podem ser usadas para fazer previsões válidas. (Two Crows Corporation, 1999: 1);

f) Data Mining envolve o uso de ferramentas sofisticadas de análise de dados, para descobrir padrões e relações válidas, até aí desconhecidos, em grandes conjuntos de dados. Estas ferramentas podem incluir modelos estatísticos, algoritmos matemáticos e métodos aprendizagem automática [55](Seifert, 2007: 1)

g) Data Mining envolve a produção de modelos para os dados ou a determinação de padrões a partir destes. (Fayyad et. al., 1996: 31);

h) O Data Mining consiste no desenvolvimento de modelos, que são tipicamente uma representação compacta dos padrões encontrados, usando dados históricos, e aplicando esses modelos a novos dados (Hornick et. al., 2007: 6).

Assim, e segundo Santos & Azevedo (2005: 10) o *Data Mining é a aplicação de métodos e técnicas em grandes bases de dados, para encontrar tendências ou padrões com o intuito de descobrir conhecimento*.

2.2.1 - MODELOS E MÉTODOS DE DATA MINING

Santos & Azevedo[56], (2005: 13), Maimon & Rokach (2010: 5) e Vercellis (2009:81), admitem a existência de duas estratégias fundamentais de DM, que podem ser usadas para fornecer informações relevantes:

a) *Modelo de Verificação: abordagem directa ou top-down*[57] (Santos & Azevedo, 2005: 13);

b) *Modelo de Descoberta: abordagem indirecta ou bottom-up*[58] (Santos & Azevedo, 2005: 13).

[55] Algoritmos que melhoram a sua performance automaticamente através da experiência, como por exemplo: redes neurais e árvores de decisão (Seifert, 2007: 1).
[56] Citando Michael Berry e Gordon Linoff *Mastering Data Mining: The Art and Science of Customer Relationship Management.* John Wiley & Sons, Inc. USA; 2000.
[57] *Usada quando se sabe o que se quer pesquisar* (Santos & Azevedo, 2005: 13).
[58] *Quando a pesquisa incide sobre os dados sem indicação do objectivo de pesquisa* (Santos & Azevedo, 2005: 13).

A abordagem *top-down* utiliza uma hipótese e testa a sua veracidade de acordo com os dados. A ênfase está no responsável (utilizador) em formular a hipótese e realizar as consultas aos dados para afirmar ou infirmar a hipótese. Esta abordagem não é criadora de nova informação no processo de pesquisa, no entanto, das consultas resultam uma série registos para verificar ou negar a hipótese. Neste caso, a pergunta é iterativa, uma vez que esta, pode ser modificada/melhorada. Há ainda possibilidade de serem reformuladas ou efectuadas novas perguntas ou hipóteses. (Santos & Azevedo**,** 2005: 10; Vercellis, 2009:81). Por outras palavras, os métodos de verificação lidam com a avaliação de hipóteses propostas por uma fonte exterior (por exemplo, um especialista). Estes métodos incluem os métodos mais comuns da estatística tradicional, como o teste de hipóteses e a análise de variâncias. Por esta razão, são menos associados ao *Data Mining* do que, os seus homólogos de descoberta, porque a maioria dos problemas de *Data Mining*, incidem sobre a descoberta de uma hipótese (de entre uma grande conjunto de hipóteses), em vez de testar uma já existente[59] (Maimon & Rokach, 2010: 5). A descoberta de factos acerca dos dados, é possível, através do uso uma variedade de técnicas, tais como: Consultas (*queries*), análises multidimensionais e visualização[60] (Santos & Azevedo**,** 2005: 10; Vercellis, 2009:81).

Na abordagem *bottom-up*, a ênfase está no sistema, o qual descobre automaticamente a informação importante que está implícita e "escondida" nos dados. Os dados são pesquisados de forma a encontrar padrões frequentes, tendências e generalizações, sem a intervenção ou orientação humana. Os sistemas geram hipóteses de forma autónoma. O objectivo é revelar um grande número de hipóteses, acerca dos dados, no menor tempo possível. (Santos & Azevedo**,** 2005: 10; Vercellis, 2009:81).

O modelo de descoberta identifica padrões nos dados, estando assim divididos da seguinte forma (Vercellis, 2009:90; Maimon & Rokach, 2010: 5; Giudici, 2003: 8):

a) Métodos de previsão; e
b) Métodos de descrição.

Os métodos de previsão, também denominados de métodos assimétricos, supervisionados ou directos, têm como objectivo, explicar uma ou mais variáveis, em relação a todas as outras, através da procura de regras de classificação ou previsão,

[59] Os métodos tradicionais de estatística focam-se, principalmente, na estimação de modelos, ou seja, opõem-se a um dos principais objectivos do *Data Mining*, que é, a identificação e construção de modelos, baseados na demonstração (Maimon & Rokach, 2010: 5).
[60] Tecnologias que apresentam os dados (principalmente através de gráficos) para que o utilizador os consiga compreender.

baseadas nos dados. Por outras palavras, métodos supervisionados, são métodos que tentam descobrir as relações entre variáveis independentes[61] (variáveis existentes) e um determinado atributo, por vezes denominado como variável dependente[62] (variável prevista/criada). As relações descobertas são representadas numa estrutura, denominada modelo. Normalmente, os modelos descrevem e explicam um determinado fenómeno, que está escondido num conjunto de dados. No entanto podem ser usados para prever o valor do atributo alvo (variável dependente), uma vez conhecido e criado o modelo (baseado nos valores dos atributos de entrada).

Por outras palavras, estes métodos, apontam para a criação automática de modelos comportamentais (dos dados), que, ao obterem amostras novas (até então desconhecidas), possibilitam a previsão de valores de uma ou mais variáveis relacionadas com a amostra. O conhecimento descoberto, formado a partir dos padrões desenvolvidos, é armazenado para posterior utilização. Assim, os processos de aprendizagem supervisionada, são orientados para a previsão e interpretação de um determinado atributo (Vercellis, 2009: 90; Maimon & Rokach, 2010: 5; Giudici, 2003: 8).

Os métodos de descrição, também denominados de métodos simétricos, não supervisionados ou indirectos, focam-se em compreender a forma como os dados fundamentais se relacionam com as suas partes. A aprendizagem não supervisionada versa-se primordialmente, sobre técnicas que agrupam instâncias, sem qualquer predeterminação dos atributos dependentes (Maimon & Rokach, 2010: 5; Giudici, 2003: 8). Logo, as análises de aprendizagem não supervisionada, não são guiadas pelo atributo. Assim, neste caso, as tarefas de DM estão direccionadas para a descoberta de padrões apelantes e de afinidades entre os conjuntos de dados. Na maioria das aprendizagens não supervisionadas o interesse está na identificação de *clusters*[63] (categorias/segmentos). Cada *cluster* ou segmento é constituído por valores e atributos similares entre si, e diferentes dos membros de outros segmentos (Vercellis, 2009:90).

Decompondo ainda mais este conceito, é ainda possível distinguir os métodos de descoberta (designadamente, os métodos de previsão e descrição), mediante as tarefas

[61] *As variáveis independentes (entradas) de um modelo, são regras ou variáveis utilizadas pela equação inerentes ao modelo para prever o valor das variáveis de saída (dependentes)* (Santos & Azevedo**,** 2005: 178).

[62] *As variáveis dependentes (saídas ou respostas) de um modelo, são as variáveis a prever pela equação ou regras inerentes ao modelo, utilizando as variáveis independentes (entradas).* (Santos & Azevedo**,** 2005: 178).

[63] Cluster é um *conjunto finito de categorias ou segmentos*, ou seja, *é a identificação de grupos homogéneos de objectos*. Cluster é o produto de uma tarefa de DM chamada, *Clustering* (Segmentação) (Santos & Azevedo**,** 2005: 16).

(objectivos) passíveis de ser executadas pelas ferramentas de DM, nas quais se incluem (Vercellis, 2009:91/92):

1) Caracterização e discriminação: aqui, antes de se começar a desenvolver o modelo de classificação, já existe o atributo categórico alvo, previamente escolhido. Onde é levada a cabo, uma análise exploratória com um duplo objectivo. Por um lado, o objectivo, é atingir a categorização através da comparação da distribuição dos valores dos atributos (para os registos respeitantes à mesma classe ou atributo). Por outro lado, o objectivo é detectar as diferenças entre valores e atributos, através de métodos de comparação. Ou seja, a distribuição de valores de um determinado atributo, é comparada com os registos (valores) de outra(s) classe(s) ou atributo(s). Esta tarefa de *Data Mining* é realizada através de meios de análise exploratória de dados, e portanto, é baseada em consultas (*queries*) e contagens (*counts*) que não necessitam do desenvolvimento de modelos de aprendizagem específicos. A informação assim adquirida, normalmente, é apresentada ao utilizador na forma de histogramas e noutros tipos de gráficos. Apesar de ser uma técnica simples, o valor da informação gerada é, no entanto, notável. Esta informação pode ser utilizada para direccionar a fase subsequente, que é, a selecção de atributos. Os atributos podem ser visualizados em termos de significância, realçando assim, aqueles que têm maior ou menor representação nos dados.
2) Classificação: a classificação corresponde à elaboração de uma função, que associe cada caso a uma classe[64], de entre diversas classes discretas[65] de classificação[66]. Para que seja possível classificar um novo objecto, de acordo com um modelo de classificação. Para a criação de modelos adequados à descrição das classes, são utilizados conjuntos de treino, com exemplos pré-classificados, que servem de base para a classificação posterior dos dados não classificados. A natureza categórica do objectivo determina a distinção entre classificação e regressão. Estas técnicas de classificação, podem ser usadas para a detecção de fraudes, aplicações de risco (empréstimos), tendências nos mercados financeiros e identificação de objectos em grandes bases de dados.

[64] Uma classe, é uma macro categoria, no qual se podem inserir vários atributos (Vercellis, 2009:91).
[65] Classes onde os números que os constituem são numeráveis, ou seja, números inteiros, ou conjuntos finitos de valores (Vercellis, 2009:91).
[66] Tomando como exemplo, os clientes de uma determinada operadora móvel, onde o objectivo é saber se determinados clientes ainda estão activos ou não, os atributos podem corresponder a: idade, antiguidade do cliente, volume de tráfego telefónico e os destinos das comunicações. Existindo assim, neste caso, duas classes discretas: "cliente activo" e "cliente não activo" (Vercellis, 2009:91).

3) Regressão: ao contrário da classificação, que se destina a fins discretas/descontínuas, a regressão é usada quando a variável alvo tem valores contínuos. Baseando-se nos atributos exploratórios disponíveis, o objectivo é prever o valor da variável contínua. Ou seja, é a procura de uma função que represente de uma forma aproximada os comportamentos das variáveis[67].
4) Análise de Séries Temporais: Por vezes, os atributos alvo, evoluem com o tempo, e por isso, são associados a períodos adjacentes num eixo temporal. Neste caso, a sequência de valores das variáveis alvo, diz-se que representa a série temporal. Por exemplo, as vendas semanais de um determinado produto, observadas ao logo de 2 anos, representam uma série de tempo contendo 104 observações. Os modelos para a análise de séries de tempo, investigam os dados caracterizados por uma dinâmica temporal e são direccionados para a previsão do valor da variável alvo para um ou mais períodos futuros.
5) Regras de Associação: regras de associação, também conhecidas como grupos de afinidade, são usadas para identificar associações interessantes e apelantes, entre grupos de registos em conjuntos de dados[68].
6) Segmentação (Clustering): O termo *cluster* (segmento) refere-se a um subgrupo homogéneo existente numa população. As técnicas de segmentação são, por isso, direccionadas à segmentação de uma população heterogénea num determinado número de subgrupos compostos por características similares. Os dados incluídos em diferentes segmentos têm características diferentes. Logo, o objectivo é que os dados sejam agrupados mediante a sua mútua homogeneidade. Por vezes, a identificação dos segmentos[69] representa uma fase preliminar no processo de *Data Mining*, dentro da análise explorativa de dados.
7) Descrição e Visualização: o propósito do processo de *Data Mining* é, por vezes, o fornecimento de uma representação simples e concisa da informação armazenada numa grande BD. Embora, em contraste com a segmentação e as regras de

[67] Se alguém desejar prever as vendas de um produto, baseando-se em campanhas promocionais já implementadas e nos preços dos produtos, a variável pode ter um grande número de valores discretos, podendo ser tratada como uma variável contínua (Vercellis, 2009:91).
[68] Num supermercado, é possível determinar que produtos são comprados em conjunto numa única transacção e com que frequência. Assim as empresas da indústria de retalho, recorrem às regras de associação para desenhar a disposição dos produtos quer nas prateleiras, quer nos seus catálogos promocionais. O agrupamento por elementos relacionados é também usado para promover vendas cruzadas ou para planear e promover combinações de produtos e serviços (Vercellis, 2009:92).
[69] Ao contrário da classificação, na segmentação não há classes predefinidas ou exemplos de referência, que indiquem a classe alvo (neste caso, segmento alvo) (Vercellis, 2009:92).

associação, a análise descritiva não procura um grupo ou porção particular de registos nos dados. Procura sim, uma descrição eficaz e concisa da informação útil, uma vez que, podem sugerir possíveis explicações de padrões escondidos nos dados, e levar a um melhor entendimento de determinado fenómeno[70].

As primeiras quatro tarefas (caracterização, classificação, regressão e análise de séries temporais) referem-se à aprendizagem supervisionada, uma vez que existe uma variável específica determinada, que necessita de ser explicada baseando-se nos atributos disponíveis ou ao longo da evolução destes. Os restantes três (regras de associação, segmentação e descrição e visualização), referem-se à aprendizagem não supervisionada, cujo fim, é o desenvolvimento de modelos, capazes de revelar as relações entre os atributos disponíveis (Vercellis, 2009:92).

Tendo em consideração o que foi referido anteriormente, é possível dizer-se que a grande maioria das técnicas de Data Mining, orientadas para a descoberta (em particular, as quantitativas), são baseadas na aprendizagem indutiva[71], onde um modelo[72] é construído, explícito ou implícito, pela generalização a partir de um conjunto de exemplos de treino. A acepção primordial das abordagens indutivas é o facto de os modelos treinados serem aplicáveis a exemplos futuros, nunca antes vistos e/ou constatados. A estratégia, também leva em conta, o nível *Meta-Learning* (como capacidade de aprendizagem e de interligação dos algoritmos) para um conjunto particular de dados disponíveis (Maimon & Rokach 2010: 4).

2.2.2 - TÉCNICAS DE DATA MINING

Actualmente, existe um grande número de técnicas algorítmicas para cada abordagem de DM. As características individuais destes têm que ser cuidadosamente avaliadas, sendo necessário adequar a técnica aos objectivos traçados no início da implementação do projecto de DM. É necessário que os utilizadores assegurem que a técnica algorítmica usada irá responder às suas necessidades (Giraud-Carrier: 2009: 514).

[70] É importante notar, que muitas vezes não é fácil obter uma visualização significativa dos dados. Contudo, o esforço de representação é justificado pela impressionante concisão atingida através de um gráfico bem desenhado (Vercellis, 2009:92).

[71] *Indução é uma técnica de inferência cujo objectivo é obter generalizações a partir da informação contida nos dados*. (Santos & Azevedo**,** 2005: 178).

[72] O modelo *descreve tendências e associações, permitindo entendê-las melhor. Uma das mais importantes funções do Data Mining é a produção de um modelo. Um modelo pode ser descritivo ou predictivo. Um modelo descritivo ajuda a perceber um fenómeno. Em modelo predictivo é uma equação ou um conjunto de regras que possibilitam a previsão de um valor desconhecido (da variável dependente) a partir dos valores conhecidos (variáveis independentes)* (Santos & Azevedo**,** 2005: 176).

Neste sentido existem vários algoritmos de modelação de DM, sendo que se destacam os seguintes:

a) Árvores de Decisão: é uma forma de representação de um conjunto de regras, seguindo uma hierarquia de classes ou valores. Estas representam uma lógica condicional simples (assemelhando-se a uma árvore quando representada graficamente). Esta técnica é usada para classificar instâncias, onde cada "nó da árvore" descreve um teste para os atributos da instância (variáveis), sendo que, cada ramo que deriva deste nó corresponde a um dos valores para esse atributo. O modo de funcionamento desta técnica resume-se à criação e treino de subclasses de informação, das quais são inferidas uma ou mais regras[73] (Santos & Azevedo, 2005: 45-46).
b) Redes Neurais Artificiais: *são estruturas computacionais baseadas em unidades de processamento (neurónios[74]) interligadas entre si e organizadas em grupos (camadas)* (Santos & Azevedo, 2005: 53). Esta técnica é bastante útil para resolver problemas de classificação, segmentação e previsão (Santos & Azevedo, 2005: 55).
c) Indução de Regras: *é a detecção de tendências e padrões em grupos de dados* (Santos & Azevedo, 2005: 77). *Por outras palavras, o objectivo é a encontrar dependências entre os atributos ou valores através da análise das probabilidades condicionais* (Santos & Azevedo, 2005: 77). Esta técnica consiste na descoberta de regras de previsão (Se…Então), encontrando-se muitas vezes associada às Árvores de Decisão, para representação do conhecimento gerado (Santos & Azevedo, 2005: 77-78).

É importante referir que a aplicação de um algoritmo revela-se necessária e imprescindível para a obtenção de modelos, padrões ou hipóteses (Santos & Azevedo, 2005: 44). Sendo que, cada tipo de algoritmo tem sua especificidade, quer em termos de dados, quer em termos de conhecimento gerado. Assim, é necessário o algoritmo (associado à técnica) de DM seja adequado às necessidades do utilizador e ao objectivo final do projecto.

[73] As árvores de decisão podem ser divididas em: árvores de classificação (que qualificam os registos, e associam-nos a uma determinada classe, garantindo a sua correcta classificação) e árvores de regressão (que estimam o valor de uma determinada variável) (Santos & Azevedo, 2005: 47)

[74] A estrutura de uma rede neural artificial, é baseada no cérebro humano, por esta razão, as unidades de processamento são denominadas de neurónios (tal como acontece no ser humano) (Santos & Azevedo, 2005: 52 e 53).

2.2.3 - ALGUMAS METODOLOGIAS DE DATA MINING

Actualmente, no mercado empresarial, com o desenvolvimento dos métodos e técnicas de DM, existem três metodologias de preponderantes de DM[75], nomeadamente (Santos & AZEVEDO**,** 2005: 25):

1) CRISP-DM (CRoss-Industry Standard Process for Data mining)[76];
2) SEMMA (Sample, Explore, Modify, Model, Assessment)[77];
3) PMML (Predictive Model Markup Language)[78].

As metodologias CRISP-DM e SEMMA são metodologias baseadas no modelo de Descoberta de Conhecimento em Bases de Dados proposto por Fayyad (et. al., 1996), desenvolvidas em ambientes diferentes. A primeira foi concebida, estudada e desenvolvida por um conjunto de empresas: NCR (National Manufacturing Company), DaimlerChrysler AG, SPSS (Statistical Product and Service Solutions) e OHRA. A segunda, foi desenvolvida pela empresa SAS, a qual se integra na área dos sistemas de suporte à decisão e do Business Intelligence (BI). Já a especificação PMML, foi desenvolvida por um grupo de criadores e investigadores de DM (Angoss Software Corp., Magnify, Centro Nacional de DM na Universidade de Illinois em Chicago, NCR e SPSS), onde o objectivo é reduzir os problemas de interoperabilidade, através da uniformização do formato de saída dos padrões gerados (Santos & Azevedo**,** 2005: 25).

2.3 - O DATA MINING E OS DADOS

Hoje em dia, na maioria das organizações, os dados estão armazenados em bases de dados relacionais. A qualidade e a utilidade dos dados, assim como, a quantidade de esforço necessário para transformar os dados para uma forma apropriada (com vista à aplicação do DM), depende dos tipos de aplicações que servem a base de dados (Chakrabarti, et. al., 2009: 37).

Assim, para as bases de dados existentes, uma solução lógica seria, de alguma forma, tentar limpar os dados. Isto é, explorar os dados por possíveis problemas e diligenciar para corrigir os erros. No entanto, actualmente, para qualquer BD do mundo,

[75] Para mais detalhe, vive Anexo VI.
[76] Para mais detalhe, vide: www.crisp-dm.org.
[77] Para mais detalhe, vide: www.sas.com.
[78] Para mais detalhe, vide: www.dmg.org.

fazer-se esta tarefa à mão, está completamente fora de questão, dada a quantidade de pessoas e tempo que envolveria[79] (Maimon & Rokach, 2010: 19).

Actualmente, úteis e poderosas ferramentas automáticas, ajudam grandemente no processo de limpeza de dados. Assim, revelam-se não só necessárias, mas também a única maneira prática e eficaz (em termos de custos) para atingir um nível de qualidade razoável nos dados existentes. Sem dados limpos e correctos, a utilidade do *Data Mining* e do *Data Warehouse* sairia mitigada. Assim, a limpeza dos dados é a necessária pré-condição para o sucesso um processo de DCBD (Maimon & Rokach, 2010: 20), uma vez que, o resultado do DM e do DCBD, depende fortemente, da qualidade e da quantidade de dados disponíveis (Cios, et al., 2007: 27).

2.3.1 - A REALIDADE DOS DADOS IMPERFEITOS

Independentemente da perfeição dos dados, estes têm, na maior parte das vezes, pouca rentabilidade. Em *Law Enforcement*, e na análise de inteligência, os dados e a informação, geralmente são tudo menos perfeitos. No DM científico e empresarial existe um tremendo controlo da qualidade dos dados. No entanto, em oposição, não é invulgar para um analista receber alguns dados ou informações num formato computacional impróprio para análise (McCue, 2007: 82).

A qualidade dos dados refere-se à exactidão e à plenitude/perfeição dos dados[80]. A presença de registos duplicados, a falta de normas de introdução dos dados, as actualizações inoportunas, e o erro humano[81] pode ter um impacto significativo na eficácia das mais complexas técnicas de DM, que são sensíveis a diferenças subtis que possam existir nos dados (Seifert, 2007: 21).

Para melhorar a qualidade dos dados, pode ser necessário proceder-se: à remoção de registos duplicados, à normalização dos valores usados para representar a informação da base de dados (Por exemplo: assegurar que "não" é representado com um zero (“0”), em toda a base de dados, e não representado por vezes com um zero ("0"), e outras com um "N"), à procura e contabilidade de dados em falta, à remoção de tipos de dados desnecessários, à identificação de anomalias nos dados (como é o caso de um indivíduo

[79] Um processo manual de limpeza de dados é trabalhoso, consume-se muito tempo, e este, é também susceptível a erros (Maimon & Rokach, 2010: 19)

[80] A qualidade dos dados, pode também ser afectada pela estrutura e consistência dos dados a ser analisados (Seifert, 2007: 21).

[81] É um processo monótono, e fatigante. No entanto, a introdução de alguns verificadores automáticos de fiabilidade e validade, nos sistemas de controlo de registos, ajudam a evitar este tipo de erros. (McCue, 2007: 84)

apresentar uma idade de 142 anos) e à uniformização dos formatos dos dados (por exemplo, alterar datas de forma a todas elas incluírem Dia/Mês/Ano) (Seifert, 2007: 21).

Qualidade dos dados é assim, uma questão multifacetada que representa um dos maiores desafios para o DM. (Seifert, 2007: 21).

2.3.2 - PREPARAÇÃO DOS DADOS PARA O DATA MINING

A preparação dos dados para o DM é a arte de espremer os dados mais valiosos disponíveis na BD, para possibilitar que o DM, como arte de descobrir padrões significativos dos dados, gere conhecimento. Essencialmente, DM não é mais do que a demanda de padrões (Ye, 2003: 366).

As ferramentas e os algoritmos de DM são, entidades totalmente diferentes, mas para o objectivo de preparação dos dados, e porque todas as ferramentas aplicam um ou mais algoritmos, mostra-se necessária a preparação destes. O principal objectivo da preparação dos dados é manipular os dados, de forma a que, qualquer padrão de input e output presentes nos dados, seja feita da forma mais obvia possível. Assim, diferentes tipos de detecção de padrões, através da aplicação de vários algoritmos, passam por diferentes técnicas de preparação (Ye, 2003: 366). Por vezes, os algoritmos, mostram diferenças demarcadas nos seus requisitos de apresentação dos dados. Por exemplo, os algoritmos de redes neurais, requerem que os dados sejam apresentados pelo menos de forma ordinal e numérica. Por outro lado, a maior parte, se não a maioria, dos algoritmos de árvores de decisão, requerem dados categóricos. Os primeiros são extremamente sensíveis a valores desaparecidos, visto que, criam estimativas contínuas das relações entre os padrões de input e os padrões de output. Enquanto que, os segundos, não o são, principalmente, por criarem uma categorização descontínua das relações entre atributos. (Ye, 2003: 367).

Assim, é possível concluir que a preparação de dados não é uma actividade opcional, mas sim obrigatória (Ye, 2003: 370).

2.3.3 - MÉTODOS GERAIS DE LIMPEZA DOS DADOS

Muitos problemas relacionados com os dados, têm impactos significativos na qualidade do resultado do processo de descoberta de conhecimento. De entre eles destacam-se os seguintes (Cios, et al., 2007: 37):

1. Grandes volumes de dados;
2. O natureza dinâmica dos dados, que estão constantemente a ser actualizados/mudados;

3. Problemas relacionados com a qualidade dos dados, como imprecisão, incompletude, ruído, valores inexistentes e redundâncias.

Estes problemas têm que ser tidos em conta uma vez que, por exemplo, só alguns métodos de DM, é que podem ser usados em dados que contenham ruído ou dados inexistentes. Neste caso, é necessário ter a certeza que se escolhe o método apropriado para analisar os dados. Ao desenvolver um sistema de descoberta de conhecimento, é necessário seleccionar um sistema de DM que consiga analisar os dados, gastando para tal, um determinado tempo. Se o objectivo for reduzir o tempo de análise, pode reduzir-se o tamanho dos dados, por exemplo (Cios, et al., 2007: 37).

Para Maimon & Rokach (2010: 23), a limpeza de dados é vista como um processo. Este processo está directamente ligado com a aquisição ou definição dos dados, ou, por outro lado, é aplicado para melhorar a qualidade dos dados existente num sistema. As três fases seguintes definem o processo de limpeza dos dados (Maimon & Rokach, 2010: 23):

1. Definir e determinar os tipos de erro;
2. Procurar e identificar as instâncias com erros;
3. Corrigir os erros descobertos.

Cada um destes passos constitui, em si próprios, um problema complexo, e uma grande variedade de métodos e tecnologias especializadas que podem ser aplicadas a cada um deles. O foco aqui está nos dois primeiros passos[82]. O último passo é muito difícil de automatizar fora de um domínio restrito e bem definido. Enquanto a análise de integridade dos dados pode descobrir um determinado número de possíveis erros nos dados, este não se direcciona a erros mais complexos. Estes tipos de erros requerem uma análise e inspecção mais profundas, facto este, que pode ser visto como um problema de detecção de *outlaiers*[83]. Ou seja, se uma grande percentagem (99,9%) dos elementos dos dados estão conformes de modo geral, então os restantes (0,1%) elementos dos dados são, muito provavelmente, candidatos a erros[84] (Maimon & Rokach, 2010: 23).

Assim, os métodos a seguir descritos, podem ser utilizados para a detecção de erros (Maimon & Rokach, 2010: 23-24):

[82] A intenção aqui é automatizar o processo de limpeza de dados, fora do âmbito do domínio do conhecimento e das regras do negócio (Maimon & Rokach, 2010: 23).
[83] São valores atípicos para um determinado conjunto de dados (tipo de dados inesperado, valor fora dos limites normais) (Santos & Azevedo, 2005: 177).
[84] Estes elementos dos dados são considerados dados atípicos (Maimon & Rokach, 2010: 23).

1. Estatístico: identificação de campos e registos atípicos. Esta abordagem pode gerar muitos falsos positivos, no entanto, é simples e rápida, e pode ser usada em conjunto com outros métodos.
2. Segmentação: identificar registos atípicos usando as técnicas de segmentação baseadas em distância Euclidian[85]. Alguns algoritmos de segmentação podem dar um apoio para a identificação de valores atípicos. No entanto, estes métodos, são de uma complexidade computacional elevada.
3. Baseado em padrões: técnicas combinadas de partição, classificação e segmentação, são utilizadas para identificar padrões que se apliquem à maioria dos registos. Um padrão é definido por um grupo de registos que têm características ou comportamentos similares, ou seja, procura-se uma percentagem "p"[86] de campos similares no conjunto de dados.
4. Regras de Associação: as regras de associação, definem diferentes tipos de padrões. Como no caso anterior, os registos que não sigam a regra, são considerados atípicos. O poder das regras de associação é que estas podem lidar com dados de diferentes tipos. Este método pode ser estendido para encontrar outros tipos de associações entre grupos de elementos de dados (Exemplo: correlação estatística)

2.4 - PRIVACIDADE VS DATA MINING

O medo normalmente acompanha o progresso. Historicamente, a ameaça de invasão da privacidade pessoal de alguém, era mais uma possibilidade do que uma realidade. Contudo, com o aumento do uso de comunicações electrónicas e da World Wide Web (WWW), tem-se tornado, fácil e barato, a troca de informação entre parceiros comerciais. Antes dos meados dos anos 90, existiam barreiras técnicas e económicas que desincentivavam a partilha de informação. Mas, à medida que estas barreiras foram caindo, o potencial uso e abuso do DM tem aumentado (Wang, 2003: 397).

O DM por si próprio, não é eticamente problemático. Os dilemas éticos e legais, levantam-se quando a extracção é executada em dados de natureza pessoal. Talvez, o mais imediatamente visível desta, é invasão da privacidade. Completa privacidade não é parte integrante de muitas sociedades, uma vez que a participação numa sociedade necessita de comunicação e negociação, que torna a absoluta privacidade irrealizável/inexequível.

[85] A distância euclidiana é uma métrica estatística utilizadas para comparar os desvios individualizados e identificar o comportamento mais próximo entre eles, ou seja, a menor distância (Cunha, et al., 2004: 30).
[86] "P" é um valor definido pelo utilizador, normalmente acima de 90 campos (Maimon & Rokach, 2010: 24).

Assim, os membros individuais de uma sociedade, desenvolvem uma única e independente percepção da sua privacidade. Assim, a privacidade existe entre a sociedade apenas porque esta existe como uma percepção dos membros da sociedade. Esta percepção é crucial, uma vez que, parcialmente determina se e em que extensão, a privacidade de uma pessoa está a ser violada. (Brazdil, et. al., 2009: 1158).

Um indivíduo pode manter a sua privacidade, ao limitar a acessibilidade dos outros a esta. Em alguns contextos, isto é alcançável através da restrição do acesso a informação pessoal. Se uma pessoa considera que o tipo e a quantidade de informação conhecida acerca deste é inapropriada, então estes podem pressentir que a sua privacidade está em risco. Assim, a privacidade pode ser violada quando, a informação de um indivíduo é obtida, usada e disseminada, especialmente se ocorrer sem o seu conhecimento ou consentimento (Brazdil, et. al., 2009: 1158)

Como a privacidade é um assunto de percepção individual, uma solução infalível e universal para esta dicotomia é inviável. Contudo, existem mediadas que pode ser empreendidas para melhorar a protecção da privacidade. Comunalmente, um indivíduo deve adoptar uma atitude pró-activa e assertiva de forma a manter a sua privacidade, normalmente, tendo que iniciar comunicações com os detentores dos seus dados para aplicar quaisquer restrições que considere apropriadas. Para a maioria, os indivíduos não sabem da extensão da sua informação pessoal armazenada pelos governos e pelas empresas privadas.

A partilha de dados corporativos pode ser benéfico para as organizações, mas permitir o acesso total às bases de dados para extracção de conhecimento, pode ter resultados prejudiciais (Brazdil, et. al., 2009: 1159).

A quantidade de dados, reunidos acerca dos clientes de determinadas empresas, é simplesmente alucinante[87] (Wang, 2003: 397).

Dados acerca de qualquer pessoa, podem ser reunidos mesmo nos locais mais improváveis[88] (Wang, 2003: 397). Os profissionais das empresas, têm que explorar como é que a tecnologia criada irá ser usada. Ética pode ser definida como normas de conduta acordadas por culturas e organizações. Neste caso, o que realmente importa para o consumidor, é que a informação reunida para um determinado fim, não ser analisada para

[87] Por exemplo, o detalhe contido no histórico de compras, de um indivíduo que use cartões VIP (Very Important Person), cartões de compras, ou cartões de crédito, para obter descontos nas lojas. Com estes cartões de sócio, as companhias são capazes de identificar as compras, e possivelmente, deduzir os interesses das pessoas (Wang, 2003: 397).

[88] Como por exemplo: informação demográfica, satisfação dos clientes, antecedentes com a justiça, registos de seguros, preferências de compras, informações bancárias, assim como, perfis médicos (Wang, 2003: 397).

outro fim secundário, a não ser que esse seja claramente compatível com o fim original (Wang, 2003: 398).

Os avanços tecnológicos possibilitam encontrar, com grande detalhe, o que determinada pessoa faz na sua vida privada. Com estes detalhes pessoais, vem uma obrigação ética substancial para resguardar/salvaguardar estes dados da divulgação a indivíduos não autorizados (Wang, 2003: 399). Os empresários têm que ser sensíveis à percepção do público em geral, em relação às suas práticas negociais, o que inclui o DM. Numa tentativa de elevar a privacidade ao nível de importância que deve ter, algumas organizações criaram uma posição de gerente conhecida como o *chief privacy officer*[89]. O qual apenas se preocupa com questões relacionadas com a aquisição e armazenamento dos dados e o DM (Wang, 2003: 400).

É importante notar que os benefícios do DM são inúmeros, não só para os empresários e para governos, mas para todos os indivíduos e para a sociedade como um todo. Infelizmente, as preocupações éticas, sociais e legais criadas em torno do DM e os medos a este associados, frequentemente levantam questões acerca do seu uso. Se um meio-termo for alcançado, onde a privacidade e a segurança dos clientes é protegida, sem limitar o poder do DM, ambos consumidores e empresários poderiam ganhar dos seus benefícios substanciais. Com crescente facilidade, as companhias são capazes de reunir vastas quantidades de dados de actuais e de potenciais clientes (Wang, 2003: 417).

O aparecimento do DM abre um número interessante de perspectivas para aumentar a competitividade de uma firma. Nestes tempos voláteis, para permanecer competitiva, uma organização tem que gerir estrategicamente a sua informação, para reagir mais rápido que os seus competidores. Contudo, esta informação tem que ser mantida segura e ao mesmo tempo acessível. Cada organização tem que se responsabilizar por assegurar que os dados estão a ser usados de forma ética e legalmente correcta. Ao mesmo tempo, as organizações têm que permanecer competitivas através da transformação os seus dados em conhecimento como forma de melhor servir os seus clientes. À medida que se torna incrementalmente fácil para os empresários reunirem informações pessoais, os indivíduos têm que exigir e usar contra-medidas de informação, para controlar quem tem acesso à sua informação, e como é que ela está a usada. (Wang, 2003: 418).

[89] Oficial chefe de privacidade.

O objectivo principal das abordagens de DM é a desenvoltura de conhecimento generalizado, em vez da identificação da informação acerca de indivíduos específicos[90] (Vaidya, et al., 2006: 1). O conhecimento generalizado não constitui na prática um risco para a privacidade (Vaidya, et al., 2006: 2). O problema pode estar não no DM, mas na infra-estrutura usada para o suportar[91]. Quanto mais completos e precisos forem os dados, melhor serão os resultados deste. A existência de bases de dados completas, compreensíveis e precisas, é que levanta problemas relacionados com a privacidade, independentemente do uso ser intencionado ou não (Vaidya, et al., 2006: 2). Embora a maior parte dos dados estejam acessíveis, o facto de estes serem distribuídos por múltiplas bases de dados (cada uma com uma autoridade responsável diferente), torna a obtenção de dados para outros fins mais difícil. Os mesmos problemas surgem com a construção de DW's para o DM. Mesmo que o DM por si próprio possa ser benigno, ter acesso a um DW para utilizar os dados para outros fins, é mais fácil do que ter acesso à fonte original (Vaidya, et al., 2006: 2).

Várias técnicas têm sido desenvolvidas para atenuar este problema (evitar o potencial abuso da informação contida num DW). Estas técnicas permitem a extracção mesmo quando não existe autorização para ver os dados. Nomeadamente: Perturbação dos dados[92] e Segurança de Computação Multipartidária[93]. Os objectivos destas técnicas de DM de preservação da privacidade[94], é possibilitar uma relação de *win-win*[95], onde o conhecimento presente nos dados é extraído para utilização, a privacidade individual é protegida, e o detentor dos dados é protegido contra possíveis abusos ou divulgações dos seus dados (Vaidya, et al., 2006: 2).

Para aliviar as possíveis deficiências ao longo desta linha, três directivas têm que ser perseguidas activamente (Cios, et al., 2007: 489; Vaidya, et al., 2006: 18-27; Liu, et. al., 2008: 420-434; Kargupta, 2009: 127-128):

[90] As regras de associação (aplicadas pelas grandes superfícies comerciais), identificam relações entre os itns comprados (exemplo: As pessoas que compram leite, e ovos também compras manteiga), onde a identificação dos indivíduos que fazem estas compras não faz parte do resultado (Vaidya, et al., 2006: 1).
[91] A maior parte dos casos de mau uso de dados privados, não derivam do DM. Pois, são os problemas de segurança das bases de dados que levam a este facto (Clifton, et al., 2009: 361).
[92] A perturbação dos dados é baseada na ideia de não providenciar dados reais ao utilizador (Vaidya, et al., 2006: 2).
[93] A segurança de computação multipartidária, é baseada na reparação da autoridade, ou seja, os dados são, presumivelmente, controlados por diferentes entidades, e o objectivo é que estas cooperem e obtenham resultados válidos, sem revelar os seus dados a outros (Vaidya, et al., 2006: 2).
[94] *Privacy-Preserving Data Mining* (Vaidya, et al., 2006: 2).
[95] Designa circunstancias em que cada parte envolvida pode beneficiar, ou que, todos os que resultados possíveis são favoráveis (Oxford English Dictionary, 2009).

a) Higienização dos dados: aqui o ponto-chave é a modificação dos dados para que os dados tidos como sensíveis, não possam ser directamente explorados. É antecipado que tal modificação de dados não irá ter um impacto significativo nas descobertas, dado o grande volume de dados disponíveis;
b) Distorção dos dados: também conhecida como perturbação/randomização dos dados, oferece privacidade através de algumas modificações nos dados pessoais. Enquanto a distorção afecta os valores de um registo individual, o impacto na descoberta e na quantificação dos relacionamentos dos dados é considerada insignificante;
c) Métodos criptográficos: diferentes técnicas de criptografia são consideradas, de modo a que os dados originais não sejam revelados durante o processo de DM[96]. As técnicas criptográficas[97] são normalmente utilizadas para dar segurança à computação multipartidária, de forma a permitir que múltiplas partes colaborem em computação conjunta, onde não aprendem nada, a não ser o resultado final da sua actividade combinada.

Assim é possível constatar que, os métodos computacionais utilizam uma determinada forma de transformação dos dados, de forma a desencadear uma "preservação da privacidade". Tipicamente, estes métodos reduzem a granularidade das representações, de forma a reduzir a privacidade. Esta redução em granulação resulta em algumas perdas de eficiência da gestão dos dados e dos algoritmos de DM, no entanto, esta é a troca natural entre a perda de informação em detrimento da privacidade.

[96] Tais técnicas escondem efectivamente os dados do acesso não autorizado, mas permitem o uso inapropriado por parte utilizadores autorizados (Brazdil, et. al., 2009: 1159).
[97] Embora seja uma actividade atractiva, os métodos criptográficos aparecem como sendo métodos de grande comunicação (interacção), mas o custo da sua computação é bastante caro. Estes custos podem mesmo ser proibitivos, especialmente quando se lida com grandes quantidades de dados (Cios, et al., 2007: 489).

CAPÍTULO 3 - A PERSPECTIVA POLICIAL DO DATA MINING

O capítulo terceiro versa-se sobre a visão policial do DM. Aqui serão abordadas várias questões como, a opinião dos oficiais da PSP relativamente às ferramentas de DM, a viabilidade da implementação destas ferramentas ao actual sistema de informação da PSP (SEI), e finalmente, algumas ferramentas de DM que estão a ser usadas por diversas polícias internacionais. Estes vários pontos a abordar visam principalmente, confirmar ou infirmar as hipóteses delineadas no início deste trabalho.

Os tradicionais sistemas policiais focam-se em pequenas partes dos dados disponíveis[98], para atingir um determinado fim. Sem DM, a quantidade de dados usados na análise é delimitada pelas limitações inerentes ao ser humano, as quais se reflectem no tempo de duração da análise. É praticamente inviável analisar todos os dados potencialmente úteis "à mão"[99]. No entanto, é importante utilizar grandes quantidades de dados, para ser possível explicar, entender, relacionar e prever. A explicação de um determinado fenómeno, normalmente, está em pequenos detalhes[100]. Assim, para ajudar os analistas a chegar às conclusões correctas, é necessário utilizar grandes quantidades de dados, para que os padrões ofereçam informação mais contextualizada (Veer et. al., 2009: 1).

Para entender o crime, os dados têm que ir para além dos aspectos simples de identificação de um incidente ou de uma pessoa[101]. O reconhecimento automático de padrões é necessário para transformar o avolumar de dados, num fluxo tratável de informações, que realmente interessem. Até porque, padrões complexos podem ser representados por poucos elementos de dados (Veer et. al., 2009: 2). As técnicas de DM também ajudam a lidar com a natureza dinâmica e com a complexidade do comportamento criminal (Veer et. al., 2009: 2). Muitas organizações, são surpreendidas pela quantidade de conhecimento adquirido através do DM, mesmo que aplicado a uma pequena parte dos seus dados. Ainda assim, quanto mais dados, melhor será a profundidade e o contexto do conhecimento gerado (Veer et. al., 2009: 2).

Usadas intensamente na comunidade empresarial, as novas ferramentas de DM não requerem: orçamentos enormes, pessoal especializado ou treino avançado em estatística.

[98] Como por exemplo, o ano, o mês e o tipo de crime (Veer et. al., 2009: 1).
[99] O DM ajuda a resolver um problema comum: quanto mais informação mais difícil e mais tempo se consome a analisar eficazmente e tirar algum sentido dos dados (McCue & Parker, 2003).
[100] Como por exemplo, com a realização de festivais de rua, existe um aumento de potenciais vítimas de furto por carteirista a circular nas ruas (Veer et. al., 2009: 1).
[101] Como é o caso do tipo de bairro, do *Modus Operandi*, das descrições das testemunhas, dos bens furtados, dos veículos envolvidos e do historial das pessoas envolvidas (Veer et. al., 2009: 2).

Por sua vez, estes produtos são muito intuitivos, relativamente fáceis de usar, baseados em computadores, e bastante acessíveis à comunidade policial (McCue & Parker, 2003).

O DM pode ser usado pela polícia, para descobrir novos padrões ou para confirmar padrões ou tendências previamente concebidas/idealizadas. Uma das forças do DM quando aposta com métodos estatísticos mais tradicionais, é que não é preciso saber-se exactamente o que se procura quando se começa (McCue & Parker, 2003).

Um dos maiores desafios do uso do DM e da análise predictiva na polícia, é que a maioria, se não todos os dados, nunca foram intencionalmente criados para a análise. Assim, desafios significativos associados ao formato dos dados, ao conteúdo, à fiabilidade e à validade[102] têm que ser constantemente avaliados e estudados (McCue & Parker, 2003).

Assim, em teoria, o DM permite à polícia, um melhor entendimento e previsão do crime, uma vez que, muitas fontes de dados podem ser analisadas e padrões complexos podem ser encontrados. Segundo um relatório interno da Polícia Holandesa o DM permite uma tomada de decisão mais eficaz e guiada por objectivos, ao nível estratégico, táctico e operacional.

3.1 – O CONHECIMENTO DO DM POR PARTE DOS OFICIAIS DA PSP

Actualmente a PSP não detém nenhum tipo de ferramenta de Data Mining ou de Descoberta de Conhecimento em Bases de Dados. Estas ferramentas de apoio à decisão podem revelar-se uma mais-valia para a actuação policial, através da criação de conhecimento útil respeitante às várias áreas de actuação da PSP. Contribuindo assim para o aumento dos níveis de eficiência e eficácia desta polícia. Este facto motivou uma investigação (através da aplicação de questionários) junto dos Oficiais da PSP, com o intuito de averiguar se os mesmos conhecem estas tecnologias. Este questionário pretendeu ainda, quantificar (segundo a opinião dos inquiridos conhecedores do DM) o grau de importância que as ferramentas de DM teriam para os vários Departamentos Policiais desta polícia.

Assim, foram aplicados inquéritos por questionário de administração indirecta (através de email), aos Oficiais da PSP, com vista à obtenção de dados quantitativos. O questionário foi aplicado ao nível do Comando Metropolitano de Lisboa (COMETLIS) e

[102] Ao contrário dos padrões empresariais onde as variáveis predictivas são baseadas em informação pouco mutável (como a informação demográfica). As tendências e padrões criminais podem mudar rapidamente e com alguma frequência em qualquer comunidade (McCue & Parker, 2003).

do Comando Distrital de Coimbra (CDC), durante o mês de Abril do presente ano. O tempo disponível para a aplicação dos mesmos não foi o ideal, factor este, que motivou a escolha dos supracitados Comandos, quer pelas suas manifestas diferenças em termos de área de competência, quer pela possibilidade de obtenção respostas rápidas por parte dos mesmos.

Para este estudo, o universo é constituído pelos Oficiais em pleno desempenho de funções, independentemente da natureza das mesmas[103] (691 Oficiais em pleno desempenho de funções). Desta forma, foi seleccionada uma amostra representativa deste universo, que compreende os Oficias adstritos ao Comando Metropolitano de Lisboa e ao Comando Distrital de Coimbra. Assim, são englobados os Comandantes de Comando, de Divisão e de Esquadra, e ainda Oficiais adstritos aos diversos núcleos.

Em termos de representatividade, o número total de Oficiais destes dois Comandos é de 216 (duzentos e dezasseis), sendo esta a amostra escolhida para representar apenas a dos Oficiais da PSP inquiridos. Em termos de adesão à investigação, obtiveram-se 125 respostas, perfazendo uma taxa de resposta de 58,1 %. Assim, obteve-se um erro de 2,94%[104] para um nível de confiança de 95%.

O questionário é composto por cinco perguntas (no total), as quais se relacionam com o tema em estudo e com as funções desempenhadas pelos inquiridos[105]. Assim, as perguntas foram elaboradas com o fim de avaliar se os Oficiais desta polícia conhecem algum tipo de ferramentas de DM. Os inquiridos conhecedores destas ferramentas, são convidados a quantificar o seu conhecimento e a pronunciarem-se sobre a necessidade das mesmas, para os diversos Departamentos Policiais. Onde se inclui: a Direcção Nacional, os Comandos, as Divisões e as Esquadras. O verdadeiro objectivo é perceber até que ponto os Oficiais da PSP sentem necessidade de esta polícia deter tais ferramentas.

Foi garantido o anonimato da população inquirida, sendo impossível identifica-los.

[103] A escolha de Oficiais da PSP, em vez de outros elementos policiais pertencentes a outras categorias, deveu-se ao facto de estes exercerem funções de comando e direcção, o que implica que estes sejam os órgãos decisores desta polícia. O DM, tal como foi referido no Capítulo 2, tem como principal função, ajudar os órgão decisores a tomar a decisão certa no momento certo, recorrendo para tal ao conhecimento gerado por estas ferramentas.
[104] Cálculo obtido através do sitio: http://www.measuringusability.com/wald.htm#wilson.
[105] Cfr. Anexo I.

Os questionários foram numerados e sujeitos a tratamento estatístico, através do programa informático, SPSS.

Passando agora à análise de resultados, dos 125 Oficiais da PSP inquiridos, 34% (n=43) afirmam saber da existência de ferramentas de DM, contrapondo contra os restantes 66% (n=82) que não conhecem[106] (conforme Figura 2).

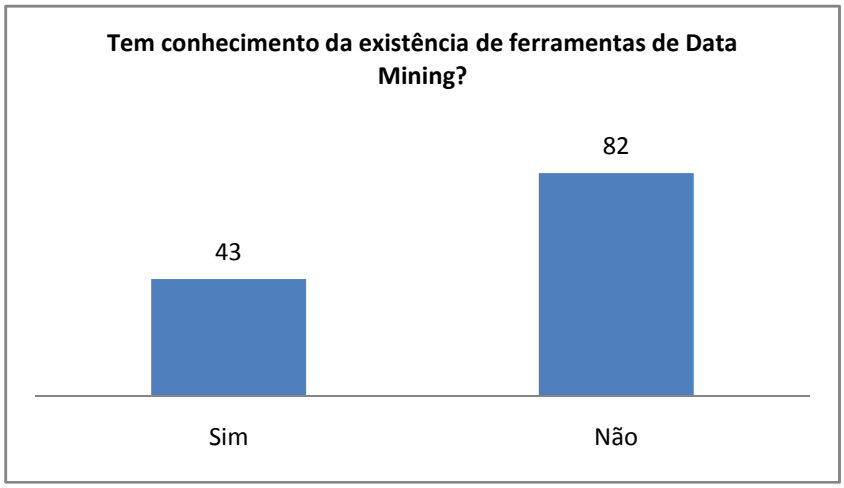


Figura 2: Gráfico 1 - Conhecimento do DM pelos Oficiais inquiridos

Este resultado, indica que existe uma quantidade considerável de Oficiais que não conhecem as ferramentas de DM. Contudo, face à especificidade da tecnologia em questão, considera-se o rácio de 34% de Oficiais que admitem conhecer esta tecnologia, como bastante positivo e optimista.

Ao distinguir-se os inquéritos por Comando, verifica-se que, apesar de o CDC apenas terem respondido 12 Oficiais, 5 destes (42%[107]) conhecem as Ferramentas de DM, enquanto que, os restantes 58% (n=7)[108] não conhecem. Por outro lado, dos Oficiais inquiridos do COMETLIS (113), 34% (n=38)[109] conhecem as ferramentas de DM, e 66% (n=75)[110] não conhecem (conforme Figura 3).

[106] Vide Anexo V: Tabela 1, Gráfico 1 e 2.
[107] Vide Anexo V: Tabela 10, Gráfico 12.
[108] Vide Anexo V: Tabela 10, Gráfico 12.
[109] Vide Anexo V: Tabela 10, Gráfico 13.
[110] Vide Anexo V: Tabela 10, Gráfico 13.

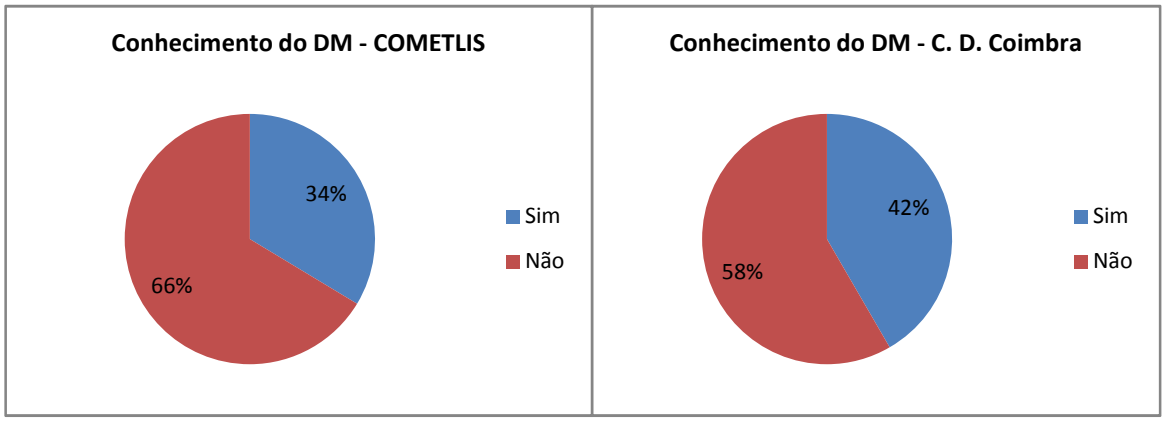


Figura 3: Gráfico 12 e 13 - Conhecimento do DM por Comando.

Quanto ao nível de conhecimento dos Oficiais inquiridos, foi-lhes proposto, que quantificassem numa escala de 1 (um) a 5 (cinco) o domínio conceptual detido acerca da matéria (Pergunta 4[111]). Neste seguimento, pôde-se verificar que a maior parte dos inquiridos, ou seja, 44% (n=19)[112] qualifica o seu conhecimento como "Neutro"[113] (conforme Anexo VI, Gráfico 3). Sendo que não se obtém nenhum resultado no nível 5 de domínio conceptual ("Domina Muito"). Não deixa de ser interessante que, apesar do nível de complexidade da matéria em questão, alguns inquiridos admitem "Dominar"[114] conceptualmente o DM (12% [n=5]) (conforme Figura 4).

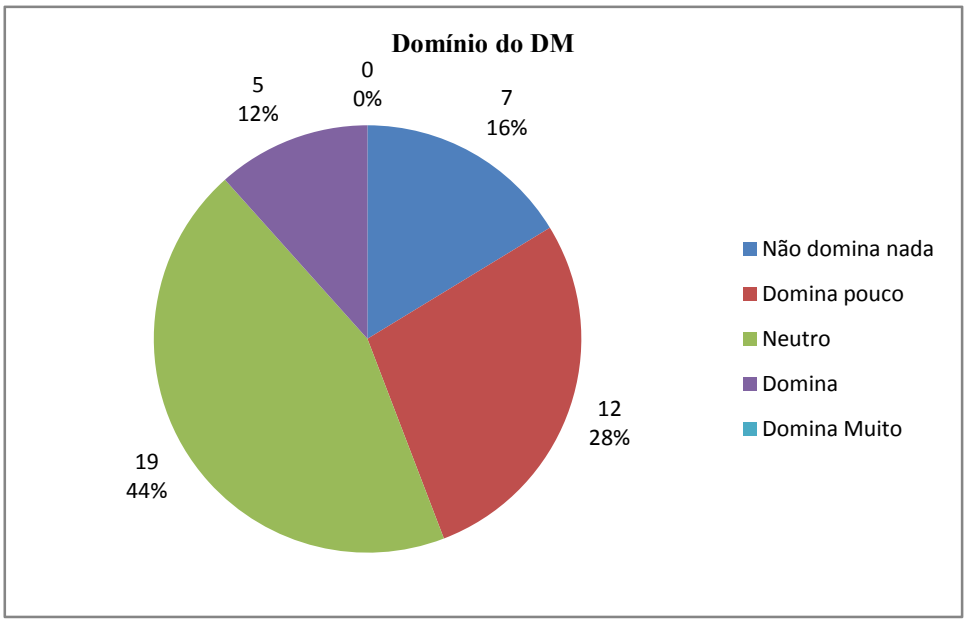


Figura 4: Gráfico 3 - Domínio Conceptual do DM para todos os Comandos

[111] Cfr. Anexo I.
[112] Vide Anexo V: Tabela 2, Gráfico 3.
[113] Não domina muito nem domina pouco.
[114] Nível de conhecimento 4.

Numa tentativa de relacionar a Pergunta 2 e a Pergunta 3[115], pretendendo-se relacionar o tipo de função desempenhada (Funções Operacionais e Funções de Apoio Operacional) com o conhecimento das ferramentas de DM, descobriu-se uma relação interessante. A maioria dos Oficiais inquiridos que conhecem o DM, desempenham funções operacionais (86% [n=37])[116]. No entanto, esta percentagem, deve-se, também, ao facto de haver mais Oficiais a desempenhar funções operacionais, isto porque, são também estes que têm maior percentagem de desconhecimento do DM (73% [n=60])[117] (conforme figura 5).

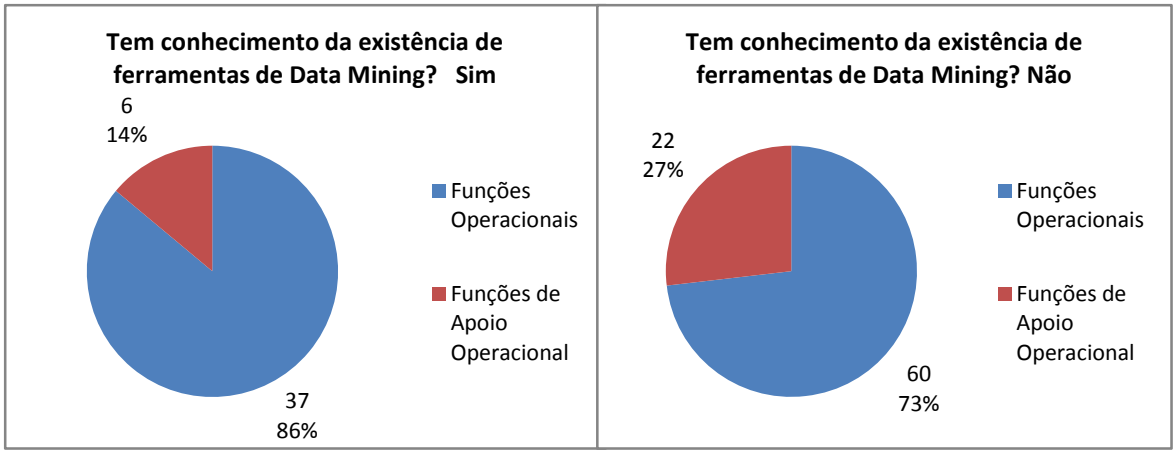


Figura 5: Gráfico 4 e 5 - Conhecimento do DM por tipo de função

A Pergunta 5[118] foi elaborada com o objectivo de saber em que medida, é que os inquiridos acham que o DM é importante, para os diversos Departamentos Policiais. Da análise dos resultados obtidos constata-se o seguinte[119] (Conforme Figura 6):

1) 38% (n=77)[120] considera o DM "Muito Importante";
2) 37% (n=76)[121] considera o DM "Importante";
3) 21% (n=43)[122] considera que o DM tem um grau de importância "Neutro";
4) 2% (n=5)[123] considera o DM "Pouco Importante;

[115] Vide Anexo I.
[116] Vide Anexo V: Gráfico 4
[117] Vide Anexo V: Gráfico 5
[118] Vide Anexo I.
[119] Aqui estão incluídos todos os departamentos.
[120] Vide Anexo V: Tabela 4, Gráfico 6.
[121] Vide Anexo V: Tabela 4, Gráfico 6.
[122] Vide Anexo V: Tabela 4, Gráfico 6.
[123] Vide Anexo V: Tabela 4, Gráfico 6.

5) 2% (n=3)[124] considera o DM "Nada Importante.

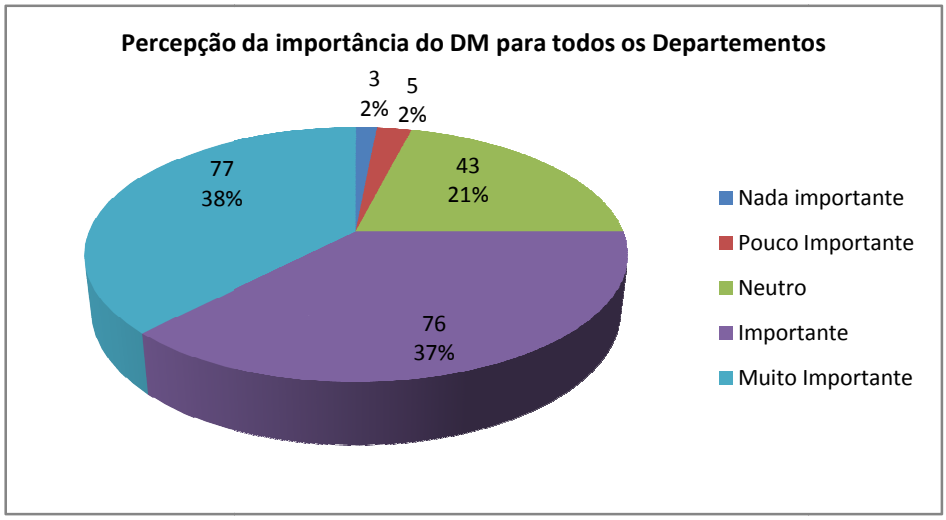


Figura 6: Gráfico 6 - Importância do DM para todos os Departamentos

Estas percentagens sugerem que a maior parte dos inquiridos consideram o DM relativamente importante. Da mesma forma, mas recorrendo a outra de ferramenta de visualização, o gráfico seguinte[125] demonstra a dispersão da opinião dos inquiridos, no que se refere à importância do DM para os mais diversos departamentos.

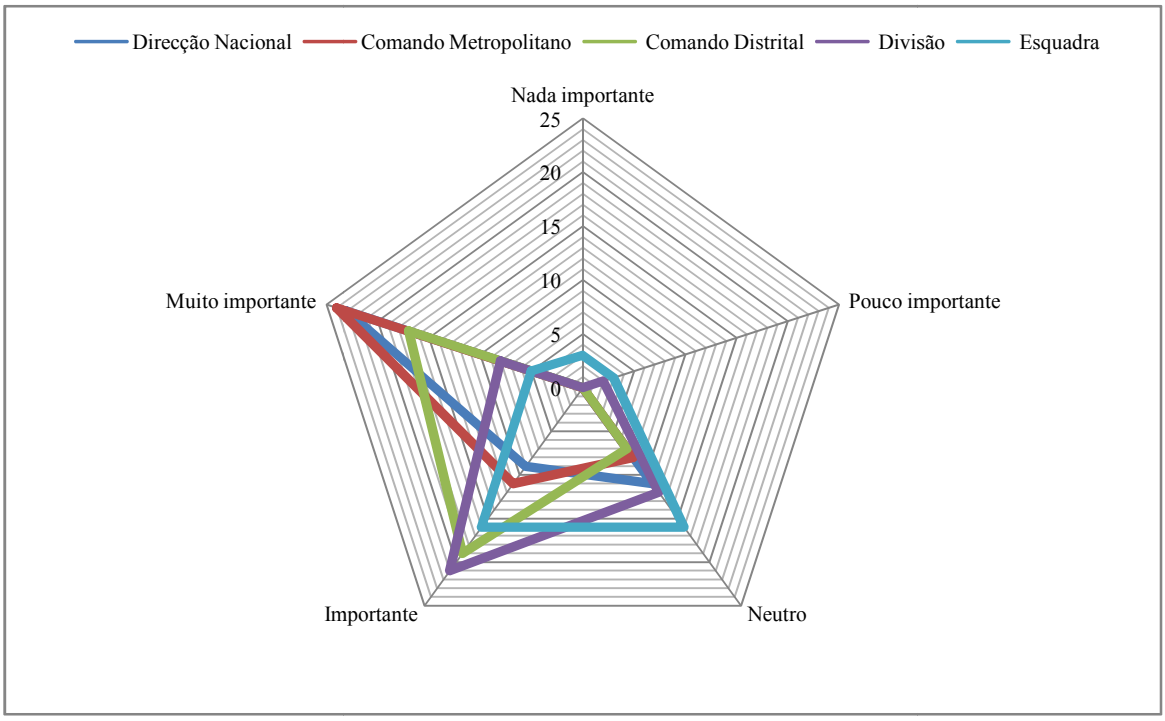


Figura 7: Gráfico de dispersão da opinião, na totalidade dos Departamentos

Este gráfico possibilita visualização das tendências de resposta dos inquiridos, onde se constata que, a maioria das respostas se direccionam para os qualificativos: "Neutro",

[124] Vide Anexo V: Tabela 4, Gráfico 6.
[125] Este gráfico foi elaborado com recurso às Tabelas 5, 6, 7, 8 e 9.

"Importante" e "Muito Importante". Este facto revela que para os inquiridos o DM tem considerável importância para os diversos Departamentos Policiais, e consequentemente, para esta Polícia.

A análise das Tabelas 5, 6, 7, 8 e 9, (e os gráficos 7, 8, 9, 10 e 11, respectivamente) faz realçar o facto de os quantificadores mais baixos ("Nada Importante" e Pouco Importante") apenas se manifestam nos seguintes Departamentos Policiais, com as subsequentes percentagens (conforme Figura 6):

1) Divisão:
   i) 5% (n=2) – "Pouco Importante";
2) Esquadra:
   i) 7% (n=3) – "Pouco Importante";
   ii) 7% (n=3) – "Nada Importante".

Estes valores percentuais, revelam que para alguns dos inquiridos (4% [n=8])[126], as Divisões e as Esquadras, seriam os Departamentos Policiais com menor necessidade de ferramentas de DM.

Conclusivamente, e tendo em consideração a hipótese de investigação (*"Parte dos Oficiais de polícia conhecem e consideram o Data Mining importante para a PSP?"*), que deu origem à utilização deste método, infere-se que os inquiridos (Oficiais da PSP) conhecedores das ferramentas de Data Mining[127] consideram-nas importantes para a PSP[128]. A nível do conhecimento do DM, por parte dos Oficiais inquiridos, verifica-se que efectivamente existe algum conhecimento da matéria em estudo (34% dos inquiridos conhecem o DM).

3.2 - O DATA MINING APLICADO AO SEI

A PSP detém, como já foi referido anteriormente, um sistema de informação denominado, SEI. Em consonância com as hipóteses levantadas, pretendeu-se descobrir, se este sistema tem capacidade para integrar uma ferramenta de DM. E ainda, se esta integração seria possível, sem efectuar alterações a nível das infra-estruturas tecnológicas

[126] Vide Anexo V: Tabela 4, Gráfico 6.
[127] Conforme foi referido no capítulo introdutório, uma percentagem acima dos 25% é considera positiva, tendo em conta a complexidade do tema em estudo.
[128] Da matriz de respostas possíveis, verifica-se que 75 % inquiridos conhecedores do DM, classificam-no como "Importante" e "Muito Importante"para a totalidade dos Departamentos Policiais

em uso. Para tal, escolheu-se entrevistar dois elementos da Direcção Nacional, mais concretamente, o Exmo. Sr. Comissário Bruno Mora do Departamento de Informações Policiais, e a Exma. Sra. Dra. Carlota Fernandes do Núcleo de Sistemas de Informação, pertencente ao Departamento de Informática.

Neste seguimento, a metodologia adoptada para a obtenção de dados qualitativos, foi a denominada, entrevista semi-estruturada. Este método tem como objectivo a recolha de opiniões junto de especialistas numa determinada matéria. O guião da entrevista foi desenvolvido com dois objectivos distintos. Primeiramente no sentido de saber em que estádio de desenvolvimento realmente se encontra o SEI, e posteriormente saber se estes membros conhecem o Data Mining bem como, se existe ou não viabilidade de aplicar estas ferramentas ao actual sistema de informação da PSP, o SEI.

O SEI, neste momento, é um repositório de informações (Anexo III) e é um sistema de informação muito completo, visto quem já abarca as diversas áreas de "negócio" da PSP (Anexo IV). Neste momento, é utilizado um software de apoio à base de dados denominado SQLServer 2005, estando está agendado para este ano, a migração dos dados da base de dados, para uma versão mais recente, nomeadamente, para o SQLServer 2008 R2[129] (Anexo IV).

Este sistema foi concebido, primordialmente, para a introdução de dados, onde a extracção dos mesmos, é bastante difícil[130], até porque o mesmo não foi concebido para tal (Anexo III). No entanto, existe uma base de dados intermédia (que contém todos os dados inseridos no SEI) que é actualizada diariamente. É com base nesta última que actualmente se efectuam as análises estatísticas, visto que a informação está um pouco mais estruturada do que no SEI. No entanto para além da análise estatística não se consegue fazer nenhum outro tipo de análise da informação lá existente, uma vez que, quando a informação é exportada, é automaticamente direccionada para a tabela respectiva, sem que os dados possam ser modelados/alterados/agrupados/desnormalizados. Não se pode manipular a informação lá existente porque esta base de dados intermédia não é um Data Warehouse (Anexo III).

Segundo os entrevistados, o verdadeiro problema do SEI, é que este não foi estruturado para a análise de dados, mas sim, para servir de base à actuação policial, ou seja, é uma base de dados transaccional (operacional) desenhada para introdução e

[129] Esta versão já contem algumas ferramentas de DM (Anexo IV).
[130] Além dos problemas inerentes à própria base de dados, existem ainda problemas da introdução, ou seja, por falta de formação do pessoal policial. (Anexo III)

catálogo dos dados (Anexo III e Anexo IV). É uma base de dados complexa, com centenas de tabelas relacionadas, não facilitando a extracção de informação (Anexo IV).

De acordo com o que se pretende com uma aplicação/base de dados operacional, o SEI contempla o registo de dados/itens de interesse, organizados em tabelas relacionadas, visando um eficaz registo e processamento de transacções. Desta forma, as tabelas e as respectivas relações estão desenhadas para que não haja redundância nem inconsistência de informação (Anexo IV). Por outras palavras, quando se efectua o registo de um processo, os dados são organizados correctamente em tabelas (ou seja, numa base de dados normalizada) logo, não existe redundância.

Por outro lado, se o objectivo for explorar a informação, os dados têm de estar preparados, ou seja, agrupados de determinada forma[131] (Anexo IV). Para explorar a informação a base de dados tem que estar construída de forma completamente diferente da actual filosofia do SEI, ou seja, é necessário construir uma base de dados desnormalizada (Anexo IV). O que implica fazer-se o trabalho contrário (ao actualmente desenvolvido no SEI) que é criar redundância[132]. O criar-se redundância significa, a mesma informação será repartida por várias tabelas. Num Data Warehouse as tabelas têm que estar desnormalizadas e os dados têm que estar agrupados. Esta filosofia de base de dados torna a exploração de dados bastante mais fácil e acessível (Anexo IV).

Entrando agora no âmbito da qualidade dos dados do SEI, denota-se que os dados lá existentes[133] têm muitas incongruências (Anexo III). Esta ideia é transmitida pelo Comissário Bruno Mora, quando refere um estudo efectuado por um elemento da Microsoft (durante seis meses). O objectivo foi estudar a base de dados e os dados nela contidos, com a finalidade de uma possível integração de um *Data Warehouse* e de um sistema de *Business Intelligence*. Segundo o mesmo, o resultado do relatório foi que não valeria a pena considerar tal evolução, visto que o SEI tem muitas incongruências dificultando a extracção dos dados (Anexo III). Igualmente, a empresa *QlikTech* através de um estudo piloto, utilizou os dados do SEI, no seu programa denominado *QlikView*[134]. Este estudo demorou vários meses, devido às incongruências nos dados retirados do SEI. O resultado foi positivo, mas tal como no exemplo anterior, o mais complicado é a adequação do SEI a este tipo de Tecnologias de Informação (Anexo III).

[131] Na base de dados existem tabelas técnicas, que são desenhadas para a inserção de dados e para garantir a integridade dos dados, as quais teriam que ser excluídas da exploração dos dados (Anexo IV).
[132] Para tal, é necessário um servidor bastante potente, porque o volume de dados é enorme (Anexo IV).
[133] SQLServer 2005 (Anexo IV).
[134] É um programa de *Business Intelligence* sem os métodos de DM, apenas utiliza estatística descritiva (Anexo III).

A análise das entrevistas, permite concluir que, os dados são incongruentes porque o sistema que os sustenta, também contém incongruências. Ou seja, os entrevistados fazem transparecer que o grande problema reside na própria base de dados[135], bem como na (in)disciplina de introdução dos dados[136]. Assim, é necessário construir uma nova base de dados (com base na actual), mas sem a componente de inserção dos dados (o Data Warehouse), para que possibilita a utilização de ferramentas que permitam explorar os dados e tirar partido destes, como por exemplo, o DM (Anexo IV).

Tendo em conta o que foi referido anteriormente, a Hipótese 2[137] é infirmada, visto que, para a introdução directa de um sistema de DM é necessária a existência de uma base de dados optimizada para a extracção de dados (por exemplo, um Data Wharehouse). Logo, alterações profundas a nível da infra-estrutura da base de dados são imprescindíveis para uma futura introdução de um sistema de DM.

Já no que diz respeito à Hipótese 3[138], os dados provenientes do SEI contêm demasiadas incongruências para que se introduza uma ferramenta de DM[139], sem se proceder a um tratamento exaustivo dos mesmos. Para se conseguir atingir um certo grau de viabilidade dos dados a utilizar num sistema de DM, seria necessário implementar mecanismos de limpeza de dados[140], com a finalidade de aumentar o grau de confiança e a credibilidade dos modelos/tendências/padrões gerados. Logo, esta hipótese não se confirma.

[135] O grande problema do SEI é mesmo a estrutura de dados, a forma como está construída, é muito confusa e muito complicada de tirar de lá os dados (Anexo III).
[136] Também é dada relevância à falta de formação dos elementos policiais, no que diz respeito ao SEI (Anexo III)
[137] Para a introdução de um sistema de *Data Mining*, sob o ponto de vista de infra-estrutura tecnológica, o actual *Data Center* do SEI, não requer qualquer tipo de alteração e/ou substituição por outras tecnologias ou infra-estruturas paralelas.
[138] A qualidade dos dados existentes na base de dados do SEI, "possibilita" a introdução "directa" de uma ferramenta de *Data Mining* ou requer um tratamento exaustivo dos dados, a montante da sua introdução.
[139] Principalmente pelo facto de, as ferramentas de DM, requerem dados "limpos" e no minimamente adequados a estas.
[140] No Subcapítulo 2.4, são dados alguns exemplos de métodos de limpeza de dados. É também necessário ter em conta, a técnica de DM a utilizar, visto que, diferentes técnicas têm diferentes requisitos quanto ao formato dos dados.

3.3 - EXEMPLOS INTERNACIONAIS DE FERRAMENTAS POLICIAIS DE DATA MINING

Nos dias que correm, o Data Mining é utilizado mundialmente por empresas de todos os ramos. No entanto, várias polícias no mundo adaptaram esta tecnologia à realidade policial.

Neste subcapítulo serão elencados alguns sistemas de DM utilizados no contexto policial. Nomeadamente o Data Detective, o Clementine na Polícia de Richmond e outras ferramentas policiais de DM. O objectivo principal é conhecer estas ferramentas, com vista a provar ou infirmar a hipótese respeitante às ferramentas policiais de DM utilizadas internacionalmente.

O processo analítico no seio policial e empresarial, requer um elevado grau de conhecimento técnico (nas áreas da estatística e do DM) o que dificulta ainda mais, encontrar polícias que dominem estas ciências (Veer et. al., 2009: 1). No entanto, como será referido posteriormente, existem diversas tecnologias que diminuem esta necessidade de detenção (por parte dos analistas) de conhecimentos técnicos profundos acerca do DM.

Nesta era digital, os profissionais de polícia necessitam de um acesso rápido aos dados. A natureza dinâmica, a complexidade do comportamento criminal e os grandes volumes de dados, "esboçam o cenário perfeito" para introdução de aplicações de DM à realidade policial (Veer et. al., 2009: 1).

3.3.1 – DATA DETECTIVE

O *Data Detective* é um sistema holandês, que foi desenvolvido através de uma parceria entre a empresa de software de DM (*Sentient)* e algumas forças policiais holandesas (nomeadamente, a *Politie Amsterdam-Amstelland,* a *Politie Midden en West Brabant* e a *Politie Brabant-Noord*). Esta parceria consistiu na construção de uma ferramenta integrada de DM, o *Data Detective*, de um extenso Data Warehouse, contendo dados de vários sistemas policiais e de fontes externas[141]. Este sistema está constantemente em evolução, pois tem oito anos de existência[142] (Veer et. al., 2009: 1).

Segundo Veer (et. al., 2009: 1), a principal chave para o sucesso é a simplificação. Uma vez que, os polícias seleccionados, depois de breves sessões de formação, rapidamente descobrem padrões e tendências, fazem previsões, encontram relações e

[141]Os dados de fontes externas incluem-se: dados sócio-demográficos, dados geográficos e dados meteorológicos (Veer et. al., 2009: 1).
[142] A polícia de Amesterdão começou a usar a primeira versão em 2001 (Veer et. al., 2009: 4).

possíveis explicações, mapeiam redes criminais e identificam possíveis suspeitos. Um conhecimento profundo e técnico em estatística e DM não é necessário.

As ferramentas de DM dificilmente têm uma continuidade útil para os peritos da polícia. Neste sentido a filosofia do desenho do *Data Detective*, tem como objectivo, solicitar que os utilizadores conheçam o domínio e tenham aptidões estatísticas, não sendo necessários mais conhecimentos. Este sistema incorpora várias técnicas de Business Intelligence, estatística, aprendizagem automática, Sistemas de Informação Geográfica (SIG's), de forma a tornar a infra-estrutura de DM compreensiva. Esta infra-estrutura inclui um DW, um módulo de *Reporting* e uma ferramenta de *Desktop*[143], que fornece aplicações de fácil consulta, de correlação, de visualização de dados, de estatística básica, de segmentação, de modelação para previsão, de modelação para explicação, de análise correlacional, de perfis geográficos e de visualizações geográficas (Veer et. al., 2009: 2).

O Data Detective detém das seguintes características:

1) Uma base de dados pré-preparada: é uma base de dados abrangente, onde todos os dados foram extraídos, correlacionados, limpos e aumentados[144]. A base de dados extensiva funciona com um único ponto de verdade, ou seja, todos os analistas usam definições e dados padronizados (Veer et. al., 2009: 3);
2) Um Data Mining automatizado[145]: baseado na tarefa e nos dados, a ferramenta escolhe a técnica indicada e optimiza os parâmetros baseados nos dados. A ferramenta assiste o utilizador ao procurar por armadilhas típicas, como por exemplo, padrões inviáveis e valores inexistentes[146]. Este é o ajustamento que tem que ser feito, para possibilitar pessoas que não são peritas na matéria, poderem tratar os dados[147] (Veer et. al., 2009: 3);
3) Uma interface amigável para o utilizador: existe uma interface gráfica intuitiva, com configurações baseadas na tarefa, em vez se basear na concepção da técnica de DM (Veer et. al., 2009: 3);

[143] Desktop é a área de trabalho do ecrã de um computador (Oxford English Dictionary, 2009). Assim infere-se, que os autores se querem referir à interface humana que é disponibilizada e simplificada e optimizada para a realização de tarefas de DM.
[144] O objectivo é que a base de dados cubra 99% da informação necessária. Onde os restante 1% necessitam de ser extraídos e preparados (Veer et. al., 2009: 3).
[145] A necessidade de conhecimento especializado em estatística e em DM, é reduzido pela selecção e configuração automática das técnicas de DM (Veer et. al., 2009: 3).
[146] É necessário referir, que em determinados casos, o perito em DM poderia superar a escolha e a selecção automática da técnica a utilizar (Veer et. al., 2009: 3).
[147] O utilizador não precisa de se preocupar com a definição de parâmetros, com o treino, com a resolução do problema de valores inexistentes, com selecção de variáveis, nem com a descodificação de variáveis em formatos utilizáveis (Veer et. al., 2009: 3).

4) Uma análise interactiva: o sistema funciona como um instrumento analítico interactivo, em que cada porção dos resultados pode ser "aumentada"[148]. Para suportar este processo intuitivo, a visualização é um importante aspecto para a interface humana. Estas possibilidades interactivas suportam o processo de descoberta e aumentam a criatividade e os instintos do analista. É ainda possível, realizar-se um processo interactivo com os requisitantes da informação[149] (Veer et. al., 2009: 3);
5) Rastreabilidade: embora o utilizador trabalhe interactivamente, é importante manter um registo dos passos que foram tomados para atingir o objectivo, especialmente porque tal documentação, pode ser necessária em sede de julgamento[150]. Assim, o sistema guarda os registos do histórico de cada resultado (Veer et. al., 2009: 3);
6) Flexibilidade dos dados: a memória associativa é usada no sistema de DM[151], como a principal técnica para: previsão, segmentação, e combinação. A preparação de dados é praticamente desnecessária, uma vez que a memória associativa pode suportar um grande número de tipologia de dados[152]. O requisito a que deve obedecer a tipologia dos dados é a possibilidade de cálculo da similitude entre variáveis. Os valores inexistentes não precisam de ser removidos ou estimados uma vez que, métricas de similitude conseguem excluir do cálculo estes "não-valores". O facto de a tecnologia poder sustentar um grande número de tipologias de dados, muito mais informação é incluída no processo de descoberta[153]. As memórias associativas são capazes de explicar como se atingiu o resultado através da apresentação ao utilizador. Tais como casos ou pessoas relevantes a partir da memória, ou seja, é uma forma intuitiva de explicar uma decisão ao utilizador (Veer et. al., 2009: 3);

[148] Desta forma, o utilizador pode simplesmente "embarcar numa viagem analítica", sem precisar primeiro de desenhar o processo. Tal como acontece nas ferramentas típicas de DM (Veer et. al., 2009: 3).
[149] Como por exemplo, o responsável por uma investigação. É possível desencadear um trabalho em conjunto, em momentos críticos do processo de análise, onde questões podem ser redefinidas, novas questões podem ser respondidas imediatamente e os padrões podem ser seleccionados baseando-se na sua relevância para a investigação (Veer et. al., 2009: 3).
[150] Denota-se que na Holanda, a nível de legislação, existe admissibilidade e viabilidade na aplicação destes métodos.
[151] As memórias associativas são robustas contra parâmetros pouco ideais (degradação graciosa) e por isso são admissíveis de serem usadas em situações onde os parâmetros são seleccionados automaticamente e o utilizador não é um perito para optimizar a técnica (Veer et. al., 2009: 3).
[152] Tais como, dados simbólicos, ordinais cíclicos (ex: dia da semana), listas, textos e categorias de muitos valores (Veer et. al., 2009: 3).
[153] Tal seria inviável com outras técnicas, pelas suas acentuadas restrições de dados (Veer et. al., 2009: 3).

7) Integração: ao integrar a maioria das técnicas numa só técnica, existe mais consistência e usabilidade. O utilizador não tem que instalar e aprender várias ferramentas, nem precisa de trocar resultados entre ferramentas de DM (através da exportação e importação). Os recursos do sistema, são mais do que as técnicas de DM, variando também entre simples navegações nos dados até análises avançadas de OLAP. Para determinados tipos de resultados, foram incorporadas ferramentas para que estes possam ser trocados/difundidos automaticamente[154] (Veer et. al., 2009: 3);
8) Análise Geo-Espacial: O aspecto espacial do crime é obviamente importante e por isso o sistema de DM permite visualizar resultados em mapas e consegue usar aspectos espaciais (coordenadas, informações sobre o solo e dados demográficos) nos modelos (Veer et. al., 2009: 4).

Estas características fazem deste sistema, uma mais-valia para a actividade policial. No entanto, esta ferramenta sairia um pouco aquém se não possui-se uma rotina de trabalho automatizada. Ou seja, o *Data Detective* tem um módulo de relatórios, o qual cria um relatório para cada Distrito[155], contendo os seguintes elementos(Veer et. al., 2009: 4):

a) Mapas de *Hot Spots*[156] (Pontos Quentes), de períodos recentes;
b) Mapas de *Hot Spots*[157] temporais que mostram o que mudou;
c) Mapas de previsão[158] do período vindouro;
d) Uma análise onde/quando[159] com a descrição dos segmentos (*clusters*) encontrados;
e) Uma distribuição semanal de previsão do crime[160], num determinado período: em que dias e horas é esperada maior incidência criminal;
f) Gráficos de incidência criminal[161], com estatística básica, tendências, e indicadores--chave de actuação.
g) Listas de *Hot Shot*[162], dos infractores mais frequentes, com as suas respectivas redes sociais e fotografias.
h) Mapas demonstrativos das residências e das áreas de actividade dos infractores.

[154] De entre estas aplicações incorporadas, destacam-se as seguintes: Excel™, Microsoft Word™, Cognos Reportnet™, Analys's Notebook™, Weka e Google Maps™ (Veer et. al., 2009: 3).
[155] É ainda possível definir um determinado crime como prioritário e este módulo criará o mesmo tipo de relatório, mas direccionado para a tipologia criminal escolhida (Veer et. al., 2009: 4).
[156] Para mais detalhe, vide: *Kernel density estimation* (Veer et. al., 2009: 7).
[157] Para mais detalhe, vide: *Temporal hot spots* (Veer et. al., 2009: 7).
[158] Para mais detalhe, vide: *Associative spatial prediction* (Veer et. al., 2009: 5).
[159] Para mais detalhe, vide: *Spatio-temporal clusters: where, when* (Veer et. al., 2009: 5).
[160] Para mais detalhe, vide: *Geographic profiling* (Veer et. al., 2009: 8).
[161] Para mais detalhe, vide: *Cluster series of crimes* (Veer et. al., 2009: 7).
[162] Para mais detalhe, vide: *Link analysis* (Veer et. al., 2009: 8).

Todos estes relatórios ajudam os órgãos decisores da polícia, a tomar decisões informadas, quer em termos estratégicos quer em termos operacionais. O sistema permite ainda o armazenamento das suas melhores práticas na forma de "receitas", onde constam: as descrições dos problemas encontrados e dos passos tomados para os resolver. Estas boas práticas podem ser procuradas e reutilizadas por todos os utilizadores.

Segundo Veer (et. al., 2009: 3) existem algumas desvantagens, nomeadamente no que diz respeito ao tempo de execução dos modelos de previsão associativa, os quais, não são tão rápidos como outras técnicas. Outras técnicas podem produzir modelos mais precisos, no entanto, em alguns casos, a memória associativa supera os restantes. Mesmo assim, os autores não consideram que as diferenças na qualidade dos modelos supere as vantagens de poupar tempo de análise, a possibilidade de utilizar dados ricos (diferenciados) e a possibilidade de "não peritos" poderem aplicar o DM.

3.3.2 – CLEMENTINE NA POLÍCIA DE RICHMOND

O Departamento de Polícia de Richmond[163] (Virginia) está a usar o DM e a análise predictiva para uma variedade de aplicações (informações e aplicações policiais). Nestas aplicações, inclui-se: a análise táctica do crime, a avaliação do risco e da ameaça, a análise comportamental de crimes violentos e estratégias de policiamento pró-activo.

O referido departamento seleccionou a plataforma incorporada de DM, denominada *Clementine* da SPSS - IBM.

Esta plataforma é utilizada para uma variedade de aplicações, de entre estas destacam-se: a distribuição do patrulhamento, o patrulhamento baseado no risco, a segurança dos polícias e Análise 24 horas. Segundo McCue e Parker (2003), o processo de descoberta associado ao DM fornece ao Chefes de Polícia e comandantes, a oportunidade de identificar padrões pouco usuais e subtis, em grandes conjuntos de dados, os quais não estariam tão prontamente disponíveis, sem as metodologias avançadas incorporadas no DM e na análise predictiva (McCue & Parker, 2003).

A capacidade de eficazmente distribuir os elementos policiais, onde e quando estes provavelmente serão necessários, ajuda a assegurar uma eficaz segurança pública, o objectivo primordial do policiamento[164]. O DM possibilita o desenvolvimento de modelos

[163] Richmond Police Department (McCue & Parker 2003).
[164] O desafio para os gestores e Comandantes de Polícia é tomar correctas decisões de patrulhamento, num esforço de dispor os recursos humanos disponíveis, para onde e quando são necessários (McCue & Parker 2003).

precisos e fiáveis que significativamente enriquecem as decisões de distribuição do patrulhamento (McCue & Parker, 2003).

A disposição das patrulhas, geralmente, está associada às chamadas dos cidadãos ou à antecipação de reclamações. O Departamento de Polícia de Richmond, utiliza unidades tácticas especializadas, que pro-activamente são destacadas para áreas associadas a problemas específicos, particularmente áreas associadas à venda de narcóticos e a crimes violentos. Um dos pressupostos subjacentes ao uso destas unidades tácticas, é que estas sejam posicionadas nos locais ou perto dos locais onde é provável que possam vir a ser necessários, para que possam responder rapidamente quando chamados. Esta disposição, apenas é possível com recurso às ferramentas de DM, as quais prevêem os locais de maior incidência criminal (McCue & Parker, 2003).

O uso do DM e da análise predictiva, permite desenvolver modelos, que prevêem as áreas de maior risco de crime violentos. Incluindo a identificação crimes que indiciam outros crimes (McCue & Parker, 2003).

O *Clementine*, avalia o risco das áreas associadas ao aumento dos ilícitos criminais, facilitando o posicionamento concentrado das unidades tácticas nas áreas mais necessitadas. Esta informação é descarregada em mapas, que são usados pro-activamente para dispor as unidades tácticas, na antecipação de aumento de probabilidade de ocorrência de criminalidade violenta (McCue & Parker, 2003).

A plataforma também desempenha análises tácticas do crime. Nesta vertente específica, os crimes ou séries de crimes podem ser caracterizados, correlacionados, e até antecipados. Recorrendo para tal à hora do dia, ao dia da semana, ao local, ao *modus operandi*, e a muitas outras variáveis. Este trabalho também é feito numa matriz temporal, que é importantíssima para identificar e deter suspeitos antes que estes cometam outros crimes[165] (McCue & Parker, 2003).

Ao longo de pesquisas automáticas de grandes bases de dados correccionais (elaboradas pelo *Clementine*), os autores afirmam ter encontrado uma relação interessante entre violadores e ladrões. Ou seja, estes descobriram que a maior parte dos violadores, têm antecedentes criminais por crimes contra a propriedade. Por outras palavras, um

[165] Esta análise predictiva oferece um valor acrescentado, de selectivamente, medir os factores mais prováveis para prever eventos futuros (McCue & Parker 2003).

violador, na maioria dos casos, tem antecedentes de furtos e roubos[166](McCue & Parker, 2003).

O *Clementine*, oferece a possibilidade de identificar e caracterizar eventos ou atributos (itens de interesse), associando-os com o aumento ou diminuição do nível de risco. Desta forma as agências policiais passam a ter uma espécie de "bola de cristal" analítica, para planear o policiamento e operações especiais e ainda para melhorar a prevenção criminal. Desta avaliação do risco é ainda possível efectuar previsões, ou seja, prever possíveis variações no nível de risco associado a um determinado local ou evento (McCue & Parker, 2003).

Um aspecto importante, que o DM e a análise predictiva (através da plataforma *Clementine*) trouxeram para a actividade policial, foi o enriquecimento da segurança, como resultado directo do crescente conhecimento da actividade criminal, dos padrões criminais e das suas tendências. Entender e caracterizar como é que diferentes factores podem interagir para criar ambientes pouco seguros para os polícias, pode resultar na alteração de estratégias operacionais, com a finalidade de aumentar a segurança dos elementos policiais[167]. Este tipo de informações, dota os profissionais de polícia de uma percepção situacional enriquecida, de modo a que estes abordem e interajam de forma diferente para com diferentes grupos de criminosos (McCue & Parker, 2003). Este facto, em última instância, resultará num reforço da segurança do polícia, é uma intervenção baseada no risco.

Por fim, uma das características mais inovadoras desta plataforma, é o facto de possibilitar análise de padrões e tendências vinte e quatro horas por dia e sete dias por semana. Esta característica assenta na ideia do perigo na demora. Por outras palavras, grande parte dos crimes são cometidos durante a noite e aos fins-de-semana. Esperar por pessoal especializado (analistas) pode comprometer a resolubilidade do caso. Assim, a forma encontrada para resolver este problema, é o fornecimento de regras de decisão, geradas pelo DM e pela análise predictiva, directamente para polícias autorizados. Usando uma ferramenta baseada na Internet, os decisores, os oficiais e outro pessoal autorizado, podem introduzir um pequeno conjunto de informações relevantes, e receber uma análise

[166] De facto, muitos predadores sexuais extremamente violentos como Timothy Spencer (o primeiro caso do uso de ADN como prova em tribunal) tiveram antecedentes de furtos e roubos (McCue & Parker 2003).
[167] Por exemplo: os vendedores de droga frequentemente usam armas para fins defensivos, enquanto que criminosos violentos tendem a usar armas para fins ofensivos. Ou seja, os criminosos que usam armas para fins defensivos, preferem armas fidedignas e que possam ser fáceis de esconder/ocultar. Enquanto que, criminosos mais violentos preferem uma arma particularmente ameaçadora ou popular (McCue & Parker 2003).

numa questão se segundos, independentemente do dia e da hora. Desta forma, assim como os criminosos não conhecem limites/fronteiras, o alcance deste tipo de ferramenta é extraordinário, onde a exploração centralizada de capacidades analíticas, é partilhada além das fronteiras geográficas e jurisdicionais (McCue & Parker, 2003).

3.3.3 - OUTRAS FERRAMENTAS

3.3.3.1 - VICLAS

O *ViCLAS*[168] é uma base de dados desenvolvida pelo Governo Canadiano[169], na qual está armazenada informação acerca de homicídios, de violações, de desaparecimentos e de restos humanos. Esta ferramenta identifica informação que associa os crimes resolvidos aos crimes por resolver[170] (Holmes, et al., 2007: 332). O *ViCLAS* está a ser usado por vários países[171], pelo facto de a *Royal Canadian Mounted Police* disponibilizar esta ferramenta a quem a solicitar e sem a cobrança de qualquer quantia monetária.

Segundo Holmes (et. al., 2007: 333), existem três desvantagens neste sistema:

1. Cada investigador ao introduzir um determinado caso de crime violento, tem que responder a 263 perguntas;
2. Os investigadores hesitam em introduzir dados sobre as suas principais provas do crime, o que reduz a probabilidade de encontrar casos idênticos;
3. A legislação sobre privacidade de determinado país pode atrasar o uso deste sistema de combate ao crime violento.

Assim, o *ViCLAS* tem sido usado, largamente, para identificar suspeitos e para resolver crimes especialmente violentos (Holmes, et al., 2007: 334).

3.3.3.2 - SHERPA

O sistema de informação *SHERPA* foi desenvolvido pela Divisão de Narcóticos de Wisconsin, nos Estados Unidos, e é usada para investigar crimes cometidos no Estado de Wisconsin. Este sistema foi desenhado para ser uma ferramenta de apoio à decisão, a qual

[168] *Violent Crime Linkage Analysis System* (Holmes, et al., 2007: 332).
[169] Esta base de dados demorou três anos a ser desenvolvida (Holmes, et al., 2007: 332).
[170] O facto de os "*serial killers*" praticarem frequentemente o mesmo tipo de crime, leva a que o armazenamento da informação de homicídios não resolvidos permita a comparação com homicídios cometidos recentemente, aumentando assim a probabilidade de encontrar estes criminosos reincidentes (Holmes, et al., 2007: 332).
[171] Nomeadamente, a Bélgica, a Áustria, a Austrália, a Holanda, o Reino Unido e os Estados Unidos (este último através dos Estados de Tennessee e Indiana) (Holmes, et al., 2007: 333).

pode combinar diversos dados de diversas fontes[172], para ajudar na investigação de processo de droga (Holmes, et al., 2007: 331).

Este sistema detém três módulos de resolução de problemas, nomeadamente: o *Language System*[173], o *Problem-Processing System*[174] e o *Metal-Level Knowledge System*[175] (Holmes, et al., 2007: 331).

O objectivo principal deste sistema, é aumentar a eficácia e a eficiência do processo de análise dos dados relacionados com os processos de droga, fornecendo mais provas e identificando mais criminosos (Holmes, et al., 2007: 331). Este facto, faz desta ferramenta um importante recurso no combate à droga, o que demonstra, uma área bastante limitada de utilização (Holmes, et al., 2007: 334).

Conclusivamente, verifica-se que a utilização deste tipo de ferramentas policiais de DM, se verifica em países desenvolvidos, como os Estados Unidos, o Reino Unido, a Holanda, a Bélgica e a Austrália. Ficando assim provada a Hipótese 1.

[172] Tais como, dados financeiros, dados de portagens, dados de vigilâncias, entre outras (Holmes, et al., 2007: 331).
[173] Este módulo lida com a interacção entre o Homem e o computador, sendo utilizada para a introdução dos dados a para a posterior obtenção de informação (Holmes, et al., 2007: 331).
[174] Aqui é comparada a classificação da informação já concebida, com a classificação em desenvolvimento pelo sistema (Holmes, et al., 2007: 331).
[175] Neste módulo o sistema fornece conhecimento sobre conhecimento, ou seja, é determinada a importância das diferentes fontes de conhecimentos para um problema em concreto que está a ser estudado (Holmes, et al., 2007: 331).

CONCLUSÃO

A PSP está a caminhar para no sentido da evolução. Vários projectos foram pensados, outros estão em desenvolvimento, mas todos com o objectivo de tornar esta polícia, numa polícia de futuro.

O futuro não se resume apenas inovação tecnológica, o ser humano também tem que se predispor a aprender e a mudar algumas ideologias fortemente enraizadas e ultrapassadas. Só assim é possível evoluir.

Descobrir conhecimento latente e escondido nas infindáveis bases de dados das polícias, não pode ser uma miragem, tem que ser uma realidade. Uma das formas de atingir essa realidade, passa pela adopção de ferramentas de DM e de DCBD, para que haja mais uma arma de combate à criminalidade cada vez mais violenta e organizada.

A realização deste trabalho centra-se na finalidade de dar a conhecer à PSP uma ponta de tecnologia relativamente acessível, e demarcadamente necessária. Assim, através da revisão bibliográfica demonstrou-se as mais variadas aplicações do DM, incluindo a sua possível inclusão no mundo policial, visto que, a quantidade de informação actualmente disponível para as policiais, é demasiada para ser tratada manualmente.

Os resultados e as conclusões obtidas neste trabalho são bastante positivas, mas também reveladoras de que há ainda um longo e árduo percurso a desenvolver. Por um lado, constatou-se que existe algum conhecimento, por parte dos Oficiais da PSP, acerca do DM e das suas potenciais aplicações. Sendo que, estes dão bastante relevo à necessidade de os níveis hierárquicos mais altos, em termos de Departamentos Policiais (Direcção Nacional, Comando Metropolitano e Comando Distrital), deterem ferramentas de DM, com vista, a dotar de conhecimento, todas as decisões estratégicas e operacionais tomadas a estes níveis.

Desta forma, é possível afirmar que, existe uma grande percentagem de conhecimento acerca do DM, bem como, das vantagens que um sistema deste tipo poderia trazer à PSP.

Por outro lado, tal como esperado, a infra-estrutura tecnológica do sistema de informação da PSP não consegue suportar directamente a introdução de um sistema de DM. Este facto deve-se principalmente, à filosofia inicial de criação do mesmo, que foi, a criação de um sistema optimizado para o registo de dados e a realização de transacções. Para o posterior tratamento da informação, seria necessário a criação de um DW, com vista a facilitar a análise da informação e a produção de conhecimento útil.

A proliferação das técnicas e das ferramentas de DM e a percepção da sua utilidade para os corpos policiais, despoletou inúmeras iniciativas de adopção e adaptação das mesmas, à realidade e necessidade policial.

Neste trabalho são elencados alguns exemplos internacionais de ferramentas policiais de DM. Reconhece-se que grande parte destas iniciativas, manifestam-se primordialmente, nos países desenvolvidos. Este facto, pode dever-se, à fácil acessibilidade, por parte dos corpos policiais destes países, às mais recentes tecnologias informacionais. Como é o caso dos Estados Unidos, do Reino Unido, da Holanda, da Austrália e da Bélgica.

Assim, constata-se que das quatro hipóteses formuladas no início deste trabalho, duas foram provadas (Hipótese 1 e 4) e as restantes foram refutadas (Hipótese 2 e 3). Neste seguimento, e com base na investigação desenvolvida, é possível afirmar que, as ferramentas de DM são, realmente, vantajosas a todos os níveis, para a Polícia de Segurança Pública. Vários esforços de modernização têm que ser feitos e várias barreiras informacionais têm que ser ultrapassadas, para que o DM passe a fazer parte do quotidiano desta polícia. Com o objectivo máximo de conhecer o "inimigo" para poder "derrotá-lo". Este conhecimento necessário e adicional, não serve apenas para fundamentar as decisões estratégicas e operacionais, pode servir ainda, para proteger e salvaguardar a segurança dos elementos policiais, para determinar o fim de uma investigação complexa, e em última instância, para prever ocorrências criminais. Assim, considera-se necessário estudar estas ferramentas, considerando assim, a usa utilização no âmbito policial.

No decorrer da realização desta dissertação de mestrado, verificaram-se algumas limitações, entre as quais, na selecção da amostra optou-se pelos Comandos supracitados, que mesmo sendo representativos da população, o espaço de tempo para a aplicação dos inquéritos não foi o desejado. O tempo disponível para a realização desta dissertação, revelou-se bastante curto. Uma outra limitação, prendeu-se com alguma dificuldade de obtenção de artigos e referências acerca dos sistemas policiais de DM.

Não obstante a estas limitações, considera-se que o trabalho desenvolvido, permitiu tirar algumas conclusões e abrir alguns caminhos no sentido de tornar a PSP numa polícia: inovadora em termos tecnológicos; geradora de conhecimento referente à toda a actividade criminal que chega ao seu conhecimento; e principalmente antecipadora de todo o tipo criminalidade.

A possibilidade de a Polícia estar um paço à frente dos criminosos, é de facto, uma mais-valia para de conseguir efectivamente, exercer a principal função da PSP que é

prevenir crimes, impedindo que estes aconteçam, criando assim mais segurança para toda a sociedade. Com estas ferramentas de DM, a presença de um polícia no local certo à hora certa, é mais do que um mero objectivo, é uma certeza e uma realidade.

Lisboa e ISCPSI, Abril de 2011

Hugo **F**erreira **L**opes
Aspirante a Oficial de Polícia

BIBLIOGRAFIA

LIVROS:

Almeida, Arnoldo Mozart Costa de; **Cézar** Gilberto Tadeu Vieira; **Assis**, Ivan Rosa de; **Valente**, Manuel Monteiro Guedes; **Pereira**, Mário Gonçalves; **Oliveira**, Paulo César de; **Martinelli**, Regiane; Boarin, **Reinaldo** Ragazzo; Oliveira, **Ricardo** Munhoz de (2009), *Sistema Policial Português* in *Estudos Comemorativos dos 25 anos do ISCPSI em Homenagem ao Superintendente-Chefe Afonso de Almeida* (2009). Edições Almedina, SA.

Bidgoli, Hossein (2002) *Encyclopedia of Information Systems – Volume 2*. Elsevier Science.

Bramer, Max (2007), *Principles of Data Mining*. Springer

Brazdil, Pavel; **Giraud-Carrier**, Christophe; **Soares**, Carlos; **Vilalta**, Ricardo, *Meta-Learning* in Wang, John (2009), *Encyclopedia of Data Warehousing and Mining, Second Edition*. IGI Global.

Broolshear, J. Glenn, (2003) *Ciência da Computação: Uma visão abrangente*. Artmed Editora S.A.

Caldeira, C. Pampulim (2008), *Data Warehousing – Conceitos e Modelos*. Lisboa: Edições Sílabo, Lda.

Chakrabarti, Soumen; **Cox**, Earl; **Frank**, Eibe; **Güting**, Ralf Hartmut; **Han**, Jaiwei; **Jiang**, Xia; **Kamber**, Micheline; **Lightstone**, Sam S.; **Nadeau**, Thomas P.; **Neapolitan**, Richard E.; **Pyle**, Dorian; **Refaat**, Mamdouh; **Schneider**, Markus; **Teorey**, Toby J.; **Witten**, Ian H. (2009) *Data Mining: Know It All*. Elsevier Inc.

Cios, Krzysztof J.; **Pedrycz**, Witold; **Swiniarski**, Roman W.; **Kurgan**, Lukasz A (2007), *Data Mining: A Knowledge Discovery Approach*. Springer

Clemente, Pedro (2006), *As informações Policiais – Palimpsesto* in *Estudos de Homenagem ao Juiz Conselheiro António da Costa Neves Ribeiro* (2007). Edições Almedina, SA.

Clemente, Pedro (2009), *Polícia - O Caminho* in **Valente**, M. M. Guedes (2009), *Estudos Comemorativos dos 25 Anos do ISCPSI em Homenagem ao Superintendente-Chefe Afonso de Almeida.* Coimbra: Edições Almedina. SA.

Clifton, Chris; **Jiang**, Wei; **Murugesan**, Mummoorthy; **Nergiz**, M. Ercan (2009), *Is Privacy Still an Issue for Data Mining?* in Kargupta, Hillol; Han, Jiawei; Yu, Philip S.; Motwani, Rajeev; Kumar, Vipin (2009), *Next Generation of Data Mining*. Taylor & Francis Group, LLC.

Cunha, Eleonora Schettini Martins; **Corrêa**, Edison José; **Carvalho**, Aysson Massote, (2004) *(Re)conhecer diferenças, construir resultados*. UNESCO, Brasília.

Dias, Hélder Valente (2010), *O Mundo Passa e a Polícia Passa Também: Metamorfoses da Polícia no Contexto de Estado Pós-Social – Lição Inaugural do Ano Lectivo 2010/2011 do Instituto Superior de Ciências Policiais e Segurança Interna.* ISCPSI.

Gaber, Mohamed M.; **Zaslavsky**, Arkady; **Krishnaswamy**, Shonali (2010), *Data Stream Mining* in *Data Mining and Knowledge Discovery Handbook, Second Edition*. Springer.

Giraud-Carrier, Christophe, (2009) *Data Mining Tool Selection* in **Wang**, John (2009), *Encyclopedia of Data Warehousing and Mining, Second Edition*. IGI Global.

Giudici, Paolo (2003) *Applied Data Mining: Statistical Methods for Business and Industry*. Chichester. John Wiley & Sons Ltd.

Han, Jiawei; **Kamber**, Micheline (2006), *Data Mining: Concepts and Techniques, Second Edition*. Elsevier Inc.

Hornick, Mark F.; **Marcadé**, Erik; **Venkayala**, Sunil (2007) *Java Data Mining: Strategy, Standard, and Practice - A Practical Guide for Architecture, Design, and Implementation. San Francisco*. Elsevier Inc.

João, Paulo A. A. (2009), *Gestão do Conhecimento nas Organizações* in **Valente**, M. M. Guedes (2009), *Estudos Comemorativos dos 25 Anos do ISCPSI em Homenagem ao Superintendente-Chefe Afonso de Almeida.* Coimbra: Edições Almedina. SA.

Kamel, Magdi, *Data Preparation for Data Mining* in **Wang**, John (2009), *Encyclopedia of Data Warehousing and Mining,* Second Edition. IGI Global.

Kargupta, Hillol, *Thoughts on Human Emotions, Breakthroughs in Communication, and the Next Generation of Data Mining* in **Kargupta**, Hillol; **Han**, Jiawei; **Yu**, Philip S.; **Motwani**, Rajeev; **Kumar**, Vipin (2009), *Next Generation of Data Mining*. Taylor & Francis Group, LLC.

Karwowski, Waldemar; **Rizzo**, Francesca; **Rodrick**, David, Ergonomics in **Bidgoli**, Hossein (2002) *Encyclopedia of Information Systems*, Volume II. Academic Press.

Liu, Kun; **Das**, Kamalika; **Grandison**, Tyrone; **Kargupta**, Hillol, *Privacy-Preserving Data Analysis on Graphs and Social Networks* in **Kargupta**, Hillol; **Han**, Jiawei; **Yu**, Philip S.; **Motwani**, Rajeev; **Kumar**, Vipin (2009), *Next Generation of Data Mining*. Taylor & Francis Group, LLC.

Maimon, Oded; **Rokach**, Lior (2010), *Data Mining and Knowledge Discovery Handbook Second Edition*. Springer.

McCue, Colleen (2007), *Data Mining and Predictive Analysis Intelligence Gathering and Crime Analysis*. Oxford. Elsevier Inc

Ratcliffe, Jerry H. (2007), *Integrated Intelligence and Crime Analysis: Enhanced Information Management for Law Enforcement Leader*. Washington: Police Foundation.

Santos, M. Filipe; **Azevedo**, Carla (2005), *Data Mining – Descoberta de Conhecimento em Bases de Dados*. FCA – Editora de Informática, Lda.

Seifert, Jeffrey W. (2007), *Data Mining and Homeland Security: An Overview*. RL31798 Congressional Research Service Report for Congress

Silva, Germano Marques da (2010), *Curso de Processo Penal – Volume I, 6ª Edição*. Verbo.

Two Crows Corporation (1999) *Introduction to Data Mining and Knowledge Discovery, Third Edition.* Potomac. Two Crows Corporation.

Vaidya, Jaideep; **Clifton**, Chris; **Zhu**, Michael (2006) Privacy Preserving Data Mining. Springer.

Vercellis, Carlo (2009)*, Business Intelligence: Data Mining and Optimization for Decision Making.* John Wiley & Sons Ltd.

Wang, John (2003), *Data Mining: Opportunities and Challenges*. Idea Group Inc.

Ye, Nong (2003) *The Handbook of Data Mining*. Mahwah, Lawrence Erlbaum Associates.

ARTIGOS:

Ewart, B.W; **Oatley**, G.C.; **Zeleznikow**, J., (2004) *Decision Support Systems For Police: Lessons From The Application of Data Mining Techniques To 'Soft' Forensic Evidence.* in *Artificial Intelligence and Law*, Volume 14, April 2006. Kluwer Academic Publishers Hingham, USA.

Fayyad, Usama; **Piatetsky-Shapiro**, Gregory; **Smyth**, Padhraic, (1996) *The KDD Process for Extracting Useful Knowledge from Volumes of Data* in Communications of the ACM, Volume 39, N.º 11, November, 1996.

Holmes, Monica C.; **Comstock-Davidson**, Diane D.; **Hayen**, Roger L., (2007), *Data Mining and Expert Systems in Law Enforcement Agencies* in *Issues in Information Systems*, Volume VIII, N.º 2. Retirado de:
http://www.iacis.org/iis/2007_iis/pdfs/holmes_davidson_hayen.pdf
Consultado em 15/01/2011

McCue, Colleen; **Parker**, Colonel Andre, (2003) *Connecting the Dots: Data Mining and Predictive Analytics in Law Enforcement and Intelligence Analysis* in The Police Chief, vol. 70, n.º 10, Outubro 2003. Retirado de:
http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display_arch&article_id=121&issue_id=102003
Consultado em 06/12/2010

Veer, Rob van der; **Roos**, H. T.; **Zanden**, A. van der (2009), *Data mining for intelligence led policing* in Proceedings *of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Retirado de:
http://www.sentient.nl/docs/data_mining_for_intelligence_led_policing.pdf
Consultado em 02/01/2011

LEGISLAÇÃO:

Lei Constitucional 1/2005 de 12 de Agosto – Constituição da República Portuguesa

Lei n.º 5/2006 de 23 de Fevereiro – Regime Jurídico das Armas e suas Munições

Lei n.º 58/2008 de 29 de Agosto – Lei de Segurança Interna

DICIONÁRIOS:

Oxford English Dictionary *– Second Edition on CD-ROM (v. 4.0.0.3)*, Oxford University Press, 2009

ANEXOS

ANEXO I

Instituto Superior de Ciências Policiais e Segurança Interna

QUESTIONÁRIO DE INVESTIGAÇÃO

Está a ser levada a cabo a Dissertação de Mestrado em Ciências Policiais pelo Aspirante Hugo Ferreira Lopes que versa sobre "*O potencial do Data Mining para o enriquecimento das informações policiais na PSP*". É um estudo que pretende avaliar, se a aplicação de técnicas de Data Mining, seriam ou não vantajosos para o enriquecimento das informações desta polícia.

Desta forma, com a aplicação do presente questionário, pretende-se perceber se, os oficiais da PSP, têm noção das potencialidades de uma ferramenta desta natureza. Mais concretamente no que diz respeito à tomada de decisões estratégicas e operacionais. Assim, este questionário aplica-se apenas aos oficiais da PSP.

Não assine o questionário. Este é **anónimo e confidencial** e destina-se exclusivamente para fins de investigação. Pede-se por isso a sua **colaboração** e máxima **sinceridade** nas respostas.

Muito Obrigado.

Considerando as actuais funções que desempenha na Polícia de Segurança Pública, indique:

1. Qual o seu posto?
   - ☐ 1.1 - Superintendente/ Intendente/ Subintendente
   - ☐ 1.2 - Comissário/ Subcomissário

2. Qual a natureza das funções que desempenha?
   - ☐ 2.1 - Funções Operacionais
   - ☐ 2.2 - Funções de Apoio Operacional

3. Tem conhecimento da existência de ferramentas de Data Mining?

   ☐ 3.1 - Sim

   ☐ 3.2 - Não

Nota: No caso de ter respondido negativamente, dê este questionário por terminado.

4. Numa escala de 1 (um) a 5 (cinco), domina esta tecnologia do ponto de vista conceptual? (1 = Não domina nada; 3 = Neutro; 5 = Domina muito)

	1	2	3	4	5
4.1 - DOMÍNIO CONCEPTUAL	☐	☐	☐	☐	☐

5. Na sua opinião, classifique (em termos de grau de importância) a necessidade de uma ferramenta de Data Mining, ao nível das funções desempenhadas pelos oficiais da PSP (1 = Nada Importante; 3 = Importância neutra; 5 = Muito importante). Assinale apenas uma das hipóteses com uma cruz.

	1	2	3	4	5
5.1 - DIRECÇÃO NACIONAL	☐	☐	☐	☐	☐
5.2 - COMANDO METROPOLITANO	☐	☐	☐	☐	☐
5.3 - COMANDO DISTRITAL	☐	☐	☐	☐	☐
5.4 - DIVISÃO	☐	☐	☐	☐	☐
5.5 - ESQUADRA	☐	☐	☐	☐	☐

MUITO OBRIGADO PELA SUA COLABORAÇÃO

ANEXO II

Guião da Entrevista

1. Qual é a sua opinião, em termos gerais, acerca do SEI?

2. Actualmente, o SEI está a atingir os objectivos para os quais foi concebido?

3. Que tipo de informação é retirada do SEI?
    a. Informação Estatística?
    b. Informações Policiais?

4. A nível das tecnologias de informação, o SEI, é um *Data Center*/Repositório de Informação/Bases de dados/*Data Warehouse*?

5. Na sua opinião, um sistema de apoio à decisão, gerador de conhecimento (a nível estratégico e operacional), é vantajoso para a PSP?

6. Na sua opinião, e a nível tecnológico, o que é que a PSP realmente precisa, para poder obter informações criminais/policiais, através das Tecnologias de Informação?

7. Está familiarizado com o termo Data Mining e Descoberta de Conhecimento em Bases de Dados? E que entende por estes conceitos?

8. Na sua opinião, seria benéfica a implementação de um Sistema de DM ou de DCBD para orientar a actuação policial?

9. Tendo em conta a actual conjuntura (policial, económica, criminal e social), acha necessário e pertinente o investimento num sistema de DM ou KDD?

10. Neste momento, e na sua opinião, a PSP, conseguiria implementar um sistema de DM/DCBD sem recorrer a parcerias (quer a nível Tecnológico, quer a nível Técnico ou Especializado)?

11. Actualmente, está a ser estudado ou está planeado algum estudo acerca de um sistema de apoio à decisão como o DM/DCBD?

12. Acha que as capacidades predictivas das ferramentas de DM, são uma mais-valia para a actuação policial?

ANEXO III

Entrevista com o Exmo. Sr. Comissário Bruno Mora do Departamento de Informações Policiais.

1 - Qual é a sua opinião, em termos gerais, acerca do SEI?

É estratégico, neste momento, era uma coisa que precisávamos quando foi implementado em 2004, tem vindo a melhorar. Foi implementado um pouco, porque havia uma janela de oportunidade em termos de investimento e portanto foi implementado. Quisemos implementar o máximo possível com o dinheiro que havia e demos a formação necessária. Em termos de evolução até agora (2011) tem havido muita evolução no sistema, mas quase nenhuma em termos de formação, o que é que acontece? O SEI foi concebido como um sistema para nós introduzirmos informação e de agora a um ano e meio mais ou menos, por acaso estamos a começar a retira-la e estamos agora a verificar que há alguns problemas na informação que foi introduzida, umas vezes porque o próprio sistema não foi concebido para extracção mas para introdução de dados, outra por falta de informação do pessoal. No entanto, aquilo que temos, já é muito bom, como eu disse é estratégico, e ainda á cerca de um mês e meio entrou em vigor uma norma de execução permanente, que criou a equipa única do SEI, uma equipa fora dos departamentos, transversal a toda a DN e a todos os Comandos, directamente dependente do Director Nacional de Operações e Segurança para gerir efectivamente o SEI a todos os níveis. Neste momento é uma ferramenta essencial e penso que a polícia já não podia voltar atrás.

2 - Actualmente, o SEI está a atingir os objectivos para os quais foi concebido?

É assim, os objectivos inicias foram, o SEI ser um repositório de informações e esse objectivo está a ser cumprido, nós estamos efectivamente a introduzir lá informação. O grande problema do SEI é a informação que lá está ser aproveitada da melhor forma. Quando a estamos a tentar retirar estamos a verificar que há alguns problemas, ou seja, o objectivo tinha duas fases, uma fase de introdução e outra de exploração porque a informação está na base de dados se não for aproveitada não serve para nada. A segunda parte, que é a exploração desses dados fica um bocado à quem, como já disse não só em problemas de introdução dos dados por falta de formação do pessoal, porque no Euro2004

toda a gente teve formação mas depois perdeu-se um bocado essa linha, e o SEI já não é nada daquilo que era em 2004, não tem nada a ver, já tem muito mais módulos de funcionamento, já tem muito mais filtros de informação. O grande problema do SEI é mesmo a estrutura de dados da forma como está construída é muito confusa e muito complicada de tirar de lá os dados.

3- Que tipo de informação é retirada do SEI?

a. Informação Estatística?

b. Informações Policiais?

O SEI não consegue só retirar estatística, nós tiramos vários tipos de informação, tiramos informação estatística básica, ou seja contabilizações, crimes, a suspeitos, vítimas e etc. E tentamos com base, neste momento em estatística descritiva fazer aquela estatística normal, que é a apresentada, mais 5% menos 5%, como apoio à decisão estratégica e operacional. Isto na parte da estatística fizemos em Dezembro de 2009/Janeiro de 2010, fizemos um relatório especial de informações onde já utilizámos algumas técnicas de estatística multivariada, tentamos traçar alguns perfis criminais, suspeitos e vitimas, o relatório foi apresentado não sei se terá sido aproveitado em termos estratégicos mas, foi apresentado pelo Departamento de Informações. Isto em termos estatísticos, em termos de informações, também é retirada informação, e de que forma? Nós temos um programa de computador que é o I2, que é um programa que está assente sobre o SEI, tem ligação a essa base de dados e a outras bases de dados, e o que faz basicamente é, nós produzimos um critério de pesquisa por exemplo o nome de um suspeito, e ele vai-nos criar de forma gráfica, todas as ligações que existem desse suspeito todos os itens de interesse que estão no SEI, por exemplo, todos os veículos que estão relacionados com aquele suspeito no SEI, depois se nós quisermos saber quem mais é que está relacionado com aquele veiculo, dizemos ao programa e ele vai-nos buscar todos os outros suspeitos e assim conseguimos criar um diagrama de associações que é muito utilizado diariamente pelo Departamento de Informações na analise das informações de exploração.

Mas isto é feito à mão, ou seja nós temos um suspeito, introduzimos o nome do suspeito, ele vai-nos dar um x de ligações, depois se quisermos vamos explorar cada uma das ligações e ele vai dar mais ligações até nós acharmos que temos informação suficiente,

não existe nada que o faça de forma automatizada. Passamos assim por várias etapas, nós é que fazemos a recolha, e a análise manualmente.

4- A nível das tecnologias de informação, o SEI, é um *Data Center*/Repositório de Informação/Bases de dados/*Data Warehouse*?

Neste momento é só repositório de informações. A informação estatística que é retirada, é retirada com base numa base de dados intermédia que devia ser uma *Data Warehouse*, mas não o é. Uma *Data Warehouse* é como sabe, optimizada para retirar informação, então o que é feito? Todos os dias de manhã há um procedimento que vai ao SEI, base de dados optimizado para introdução de dados, retira os dados um pouco mais estruturados do que estão no SEI, para uma base de dados intermédia, com base nessa é que são feitos os dados estatísticos. Mas essa intermédia não é uma *Data Warehouse*, só tem aquela informação que foi definida como indo para ali, mas não conseguimos manipular a informação que lá está.

5- Na sua opinião, um sistema de apoio à decisão, gerador de conhecimento (a nível estratégico e operacional), é vantajoso para a PSP?

Claramente, não é só vantajoso é essencial, ou seja Portugal está sempre um pouco atrasado relativamente aos outros países da Europa. Mas nós temos exemplos ainda à pouco tempo penso que foi na Holanda, que tinham um caso (isto saiu num semanário) vários crimes de um Serial Killer que à cerca que 20 anos que andavam à procura dele, tinham milhares e milhares de peças de expediente e não conseguiam tratar daquilo, não conseguiam descobrir nada, criaram um sistema de apoio à decisão, e uma base de dados. Introduziram lá os dados e através de algumas técnicas de Data Mining em cerca de uma semana descobriram o autor dos crimes e isso veio beneficiar todo o Mundo. Portanto neste momento hoje em dia a quantidade de informação que está disponível para as polícias é demasiada para ser tratada de forma leve/manual. Portanto o SEI como só repositório de informações também não ajuda muito, a única coisa que faz é nos termos ali a informação disponível mas se formos à mão pesquisar e fazer as relações no SEI é quase o mesmo que irmos às peças de expediente antigas à mão, aos dossiers à procura das relações, nos precisamos é de efectivamente um sistema automatizado de apoio à

informação, que me diga por exemplo, quando eu comandante de um Comando chego de manhã ao meu gabinete e ligo o meu computador eu quero saber automaticamente como está a criminalidade no meu Comando. Se subiu de desceu, quais os crimes que mais subiram quais os que mais desceram, etc. Em que zonas é que estão a ocorrer, a que horas do dia, quais os perfis dos suspeitos, quais são as vitimas. Só assim é que ao nível da DN, podemos ter decisão estratégica, ao nível do comando é a decisão operacional.

6 - Na sua opinião, e a nível tecnológico, o que é que a PSP realmente precisa, para poder obter informações criminais/policiais, através das Tecnologias de Informação?

O que a PSP devia avançar era…Eu sou um bocado critico do SEI, por causa da forma como ele foi implementado, já me justificaram e compreendo de certa forma, que é como disse no inicio, havia uma janela de oportunidade em termos de investimento muito pequena e nós tínhamos de aproveitar ao máximo, e eu tenho uma opinião contraria e muito própria minha que é, penso se aquilo que era essencial para a policia era o repositório de informações catálogo com os itens de interesse. Portanto registo de processos de expediente e dos itens de interesse, pessoa, local, veiculo e etc, era nisso que nós tínhamos que investir tudo e garantir que esse modulo ficava perfeito de inicio e a partir dai quando esse estivesse a funcionar a 100% começar então a utilizar outros módulos que explorassem esse modulo central. Como disse como a janela de investimento era muito curta, tentamos apanhar vários âmbitos com o mesmo dinheiro, resultado: nenhum módulo ficou a 100%. Portanto neste momento aquilo que deveria ser feito era reformular o SEI tentar de alguma forma melhorá-lo estamos a fazer, essa equipa única do SEI que foi criada é esse o objectivo ir melhorando continuamente a partir daí criar uma verdadeira Data Warehouse e em cima aplicar um Business Intelligence, Data Mining quer em estatística descritiva de apoio normal, quer em termos de Data Mining encontrar informação que está lá, padrões que estão escondidos e que não é possível ver com tratamento manual.

7- Está familiarizado com o termo Data Mining e Descoberta de Conhecimento em Bases de Dados? E que entende por estes conceitos?

É assim, o objectivo é descobrir informação que está escondida, relações que estão escondidas, com recurso a diversos algoritmos. Há algoritmos de todo o tipo e descobrir informação que não é possível, dada à quantidade de informação descobri-la manualmente. E com base nesses algoritmos encontrar relações que possam apresentar esse conhecimento, portanto nós temos os dados introduzimos na base de dados, mas o objectivo final é retirar conhecimento. É nós podermos decidir com base na informação que lá está. O objectivo do Data Mining é descobrir esse conhecimento e apresenta-lo de forma que ajude a decisão.

8-Na sua opinião, seria benéfica a implementação de um Sistema de DM ou de DCBD para orientar a actuação policial?

Esta questão não foi colocada, pelo facto de o entrevistado já ter respondido ao longo das perguntas anteriores.

9- Tendo em conta a actual conjuntura (policial, económica, criminal e social), acha necessário e pertinente o investimento num sistema de DM ou KDD?

Necessário sim sem dúvida, pertinente também, se é exequível, neste momento não. Porque um sistema destes não é só o custo do sistema, porque há sistemas que já são um grande salto em termos de descoberta de conhecimento de base de dados que até não são muito caros os próprio sistema, o que é caro é a adequação desse sistema ao SEI. Nós tivemos à certa de 6 meses talvez aqui um elemento da Microsoft que esteve a analisar o SEI exactamente com esse objectivo, de eventual colocação de um Data Warehouse e de um sistema de Business Intelligent e o relatório final foi que neste momento "nem pensar", não vale a pena porque há tantas incongruências na base de dados que não é possível extrair informação dessa forma. Portanto o sistema é barato, o problema é adequar o SEI, e o sistema ao SEI, e isso são muitas horas (Homem), são meses a trabalhar a fundo, e a assistência técnica depois, porque há sempre erros na implementação de um sistema deste e isso é que sai caro. Que são os contratos de manutenção.

10 - Neste momento, e na sua opinião, a PSP, conseguiria implementar um sistema de DM/DCBD sem recorrer a parcerias (quer a nível Tecnológico, quer a nível Técnico ou Especializado)?

Seriam necessárias parcerias, claramente. É assim, nós temos aqui bons informáticos, temos aqui bons técnicos em termos de programação, que já utilizamos muito no desenvolvimento do SEI, neste momento o GSI (Gabinete de Sistemas de Informação) tem apenas dois elementos a trabalhar da *Accenture* que foi a empresa que implementou o SEI, todos os outros técnicos que estão a trabalhar no SEI são da PSP, são técnicos superiores da PSP. Mas uma coisa é trabalhar com o SEI em termos da programação do SEI, e Trabalhar com o SQL, etc. Outra coisa são por exemplo mecanismos de Data Mining os algoritmos não são propriamente do conhecimento geral, portanto já implicam conhecimentos muito mais técnicos, normalmente algoritmos estatísticos tem que haver algum conhecimento nessa base. E portanto, não é possível neste momento à PSP, quer em termos de técnicos humanos, quer em termos de dinheiro, porque é sempre o maior problema. Neste momento não é possível nós avançarmos para um sistema assim sozinhos. Portanto, tem que ser sempre com apoios, uma empresa externa que venha cá implementar e que eventualmente depois forme os nossos técnicos para manter o sistema sozinho sem contracto de manutenção. Mas a implementação inicial terá sempre que ser externa.

11 - Actualmente, está a ser estudado ou está planeado algum estudo acerca de um sistema de apoio à decisão como o DM/DCBD?

Está, tivemos já um piloto, por uma empresa, a *QlikTech*, que tem um programa chamado *QlikView* que é um Business Intelligence sem a parte Data Mining, ou seja , tem apenas a parte da estatística descritiva. Eles fizeram um piloto com dados de fardo tratados, tiveram uns quantos meses a tratar os dados do SEI, por causa dessas incongruências, e apresentaram um projecto-piloto que era exactamente o que eu estava a dizer antes, que era: Eu chegava como comandante de esquadra, comandante de divisão, comandante de um comando. Se eu fosse comandante de Comando, chegava de manhã abria o Clickview e tinha ali a informação do meu comando, podia ir ver ao nível da divisão, a nível da esquadra, as horas que queria, os tipos de processos que queria, se estavam a subir se

estavam a descer. Tinha todo o tipo de informação, todo o tipo de gráficos etc. Eles fizeram esse piloto, o problema aqui é sempre a implementação que é caro e nós neste momento não temos dinheiro para isso. A vontade existe, de evoluir e temos aqui muitas pessoas que querem evoluir e que conhecem estas ferramentas e sabem que seria uma mais valia para a decisão da PSP, mas não é possível, há outras prioridades neste momento, como pagar os ordenados por exemplo.

12 - Acha que as capacidades predictivas das ferramentas de DM, são uma mais-valia para a actuação policial?

Foi feito em 2009 também, mais ou menos em Outubro ou Novembro antes daquele tal relatório que eu disse, foi feito um outro relatório anterior feito pelo Director do Departamento de Informações Sr. Intendente Feães Fernandes, exactamente um relatório de analises predictiva da criminalidade, quem utilizou os dados de 2 anos anteriores utilizou um algoritmo Arima12, que é um algoritmo de regressão que prevê efectivamente qual, neste caso o relatório que ele fez, qual é que iria ser a criminalidade de ocorrer a nível nacional nos 3 meses seguintes e depois a 12 meses. A 12 meses ainda está agora a decorrer o período mas os 3 meses a margem que foi dada cerca de 0,5%, caiu lá certinha e não falhou nada. E essas técnicas são utilizadas em todo o mundo. Quer a nível global como foi feito para o pais inteiro, quer a nível local ou quer ao nível de rua. Há algoritmos que são utilizados nos Estados Unidos de Data Mining, que, também por causa da forma como eles trabalham, que eles utilizam, por exemplo nós temos um módulo no SEI de gestão de ocorrências, é suposto as centrais rádio utilizarem esse sistema para registarem todas as chamadas, alguma estão a usar outras não, a nível nacional. Nos Estados Unidos eles usam e o que é que eles conseguem fazer com alguns algoritmos de Data Mining? Conseguem descobrir crimes que indiciam outros crimes, ou seja, eles conseguiram descobrir num dos "papers" que eu vi, que sempre que havia durante um certo período um aumento no número de ocorrências que envolviam armas de fogo, (basicamente eram chamadas da policia para tiros que tinham havido), e depois chegava-se lá e não se encontrava nada, mas que passava um certo período de aumento, verificava-se um aumento em 2 ou 3 tipos de crime muito específicos isto durante vários anos. E é isto que é preciso para a PSP. É isto que é preciso para as policias, porque a criminalidade não é assim tão variada ao longo do tempo, é uma coisa que se mantem muito depois pode variar em

termos de conjuntura económica, etc. mas normalmente um bocado em termos de tendências de padrões. Se nós conseguirmos, com base na informação anterior, prever o objectivo máximo da polícia e de todas teorias do crime dizem isso: o objectivo é ter o polícia no local certo. Uma das condições para ocorrer um crime é não haver ali um "Guardião da Paz" presente. Se nós conseguirmos estar no local onde iria ocorrer o crime ele obviamente não vai acontecer e é esse o nosso grande objectivo. Isto é que é a verdadeira polícia de prevenção, porque a da investigação nós fazemos depois e fazemos com o SEI agora. Como eu disse com o I2 nós vamos lá, procuramos ali os suspeitos as relações deles, isto com coisas que já estão registadas no SEI. Mas se nós conseguirmos com aquilo que está registado prever o que poderia acontecer e evita-lo, isso seria o ideal.

ANEXO IV

Entrevista com a Exma. Sra. Dr. Carlota Fernandes do Núcleo de Sistemas de Informação

1 - Qual é a sua opinião, em termos gerais, acerca do SEI?

Eu acho que é um sistema de informação muito completo. Para já, porque abarca as várias áreas de negócio da Polícia, onde as informações e os vários processos policiais, estão integrados, ou pelo menos teoricamente estão integrados. Mas, estão sim. Porque tem uma base comum, com os itens de interesse. Estes são a base de todo o sistema de informação, como: as pessoas, os veículos, os locais, as armas. É, realmente, conseguir obter a informação do que se passa com um desses objectos. No fundo, é uma das grandes vantagens, saber-se sobre uma coisa o que passa sobre ela. Isto é um nível de conhecimento muito bom. Porque nesta altura, e ao longo deste tempo todo, desde 2004 têm-se efectuado pequenas melhorias no sentido afinar, de a polícia poder tirar mais partido dessa informação. Porque é crucial para a Polícia saber o que se passa sobre uma pessoa. Acho que o sistema ao longo deste tempo todo, não esteve parado, esteve em constante evolução e, aliás, é de tal maneira o mundo da informação, que diariamente é inserido, que por exemplo: temos o sistema a funcionar e os servidores, que suportam o SEI, não caiem. Temos três servidores aplicacionais que suportam o processamento da aplicação, e também, um servidor de base de dados. Ou seja, são quatro. O nível de manutenção tem de ser constante. E porquê? Posso ter a aplicação a funcionar, e tenho, mas passado um ou dois meses o sistema degrada-se, devido ao grande volume de informação inserido diariamente. Porque não estamos a falar de meia dúzia de registos. Há muita informação a ser inserida neste sistema de informação que constantemente temos que estar a melhorar questões técnicas, de performance, até à forma como os dados são obtidos da base de dados para fazer optimizações. A manutenção constante que é precisa assegurar significa que o sistema tem muita utilização. Porque a nível de base de dados as coisas não se degradam de um dia para o outro. Mas, neste caso, em relação ao SEI acontece. Eu acho que, nomeadamente, nas minhas funções que tenho estado a acompanhar o SEI desde o início não há monotonia neste sistema. Asseguro-lhe que todos os dias aparecem situações novas que é preciso ultrapassar. Por isso, acho que se tem evoluído muito positivamente. Têm-se melhorado muito o sistema. Estão sempre a aparecer novas velocidades, como é normal. Tem havido um bom investimento a nível do desenvolvimento. Agora, se estamos a ir de encontro com o utilizador, ou não, se calhar uns estão contentes porque as coisas da sua área vão sendo melhoradas e para outros estará

mais deficiente. Mas, isto é como tudo. As alterações que nos pedem estão identificadas, organizadas, e vai ser à medida que há disponibilidade. Acho que está num nível de utilização e as pessoas podem tirar muito partido da informação que têm.

2 - Actualmente, o SEI está a atingir os objectivos para os quais foi concebido?

Eu acho que sim. Embora tenha sido um processo um bocadinho doloroso. Porque no início, 2004, desenvolveram-se módulos para todas as áreas de negócios. Não sei se mal ou bem, foi uma boa estratégia. Porque tentou-se haver um investimento enorme inicialmente, mas não se conseguiu absorver toda a camada de software que foi desenvolvida. Existem módulos no SEI como os módulos das informações, a gestão das celas e detidos, a coordenação de meios das centrais do 112 e centrais telefónicas. Não se conseguiu tirar o partido desses modos. Porque, acho que foi muita coisa ao mesmo tempo, muita informação ao mesmo tempo, e uma mudança muito grande. As pessoas tiveram de se adaptar a um sistema novo. É como não ter nada e passar a ter tudo, o utilizador tem de ter um certo tempo para utilizar o sistema, para ter maturidade, para tentar aperceber-se de qual o partido que pode tirar da informação que vai sendo gerida no SEI. Já se atingiram os objectivos? Acho que não, porque é um sistema iterativo. Mas, temos ido de encontro a melhorar as coisas dentro do possível para que, cada vez, termos um sistema melhor e com mais qualidade. Atingir ou não, acho que ainda não se pode responder a isso, temos sempre de melhorar.

3 - Que tipo de informação é retirada do SEI?

a. Informação Estatística?

b. Informações Policiais?

A informação retirada é uma das partes do SEI. Temos uma base de dados complexa, com centenas de tabelas que se cruzam umas com as outras. Não é fácil tirar a informação de um sistema de informação como este. No entanto, para aquelas áreas em que é uma necessidade, isto é, uma necessidade real, em que é preciso de totais, de processos-crime, de contra-ordenações, de acidentes. Quer dizer, há estatísticas que já estão perfeitamente identificadas, números digamos. Por isso, já está a ser dada resposta a esse nível de gestão. Na gestão pró-activa, eu isso ainda não estou a ver poder acontecer, que é eu poder prever no SEI, por exemplo, num determinado período numa zona os crimes que estão a ocorrer ou os tipos de ocorrências, e tentar prever de uma actividade

operacional que se possa fazer um investimento, mediante os factos. Ainda não chegamos a esse nível, porque também não temos ferramentas que permitam às pessoas explorar a informação a esse nível. Portanto, o que está a ser feito são totais. É saber o que é que existe. Para onde vamos? Podem-se fazer algumas previsões com base no histórico, comparando a estatística anual dos vários anos, mas agora, acho que teremos de fazer um investimento maior para chegar a esse nível. Isto é, ter ferramentas DATA MINING que nos permita com base em dados, num determinado número de dados serem analisados e poder prever qualquer coisa no futuro, isso é muito importante para a polícia (actividade operacional), mas ainda não chegamos a esse nível.

4- A nível das tecnologias de informação, o SEI, é um *Data Center*/Repositório de Informação/Bases de dados/*Data Warehouse*?

É uma base de dados centralizada, ou seja, temos uma base de dados que é SQLServer 2005. Agora estamos num processo a arrancar durante este mês de Março, num processo de migrar os nossos dados para uma nova versão de base de dados que é o SQLServer 2008 R2. À partida todas as novas versões de software têm uma melhor performance e, portanto, temos de migrar naturalmente, é o caminho. E essa migração está prevista, portanto, durante o ano de 2011. Não temos um Data Warehouse, porque isso é uma outra base de dados. Enquanto uma base de dados a nível operacional envolve o registo de dados, portanto, em que as tabelas são desenhadas para que não haja redundância, para que cada uma guarde os dados específicos. Numa Data Warehouse tenho de ter tabelas desnormalizadas, tenho de ter dados agrupados, ou seja, é uma base de dados com outra filosofia e não deixa de ser complexa. Ainda não se conseguiu implementar chegar ao desenho dessa base de dados. É ambição minha que este ano se conseguisse fazer isso, começar a reestruturar a base de dados em 2011. Não sei se vai ser possível, a disponibilização de uma base de dados desnormalizada para utilizar como PowerPivot Tables para o Exel 2010. Porque a última versão do Excel é possível trabalhar com grande volume de dados. Eu acho que quando nos pedem informação ad hoc, ou seja, para um determinado tipo de crime, numa determinada zona, um suspeito com determinadas características. Ainda estão totalmente dependentes de desenvolvimento à medida. Ou seja eu tenho de ter programadores que me desenvolvam à medida, ou seja, se nós conseguimos ter uma base de dados, não digo um Data Warehouse a prever todas as situações. A base de dados é tão complexa que é difícil criar um Data Warehouse de início já a prever a

interacção da informação. Era bom que nós juntamente o DIP (Departamento de Informações Policiais), porque ele tem a maior necessidade de pro-activamente saber dados para tomar decisões e disponibilizar informação. Eu gostaria que este ano fosse desenvolvida, é uma ambição minha, porque vejo que é uma limitação para quem quer informação não poder ter tabelas construídas, (porque as tabelas na base de dados são códigos) para que eles pudessem ter acesso a tabelas já com os campos descodificados, para que, quando eles precisassem de alguma coisa acedessem a uma base de dados e depois trabalhassem no Excel. Portanto o output seria o Excel ou o Front-End, mas que tivessem essa base de dados já construída de forma a que conseguissem ler os dados. Porque acedendo à base de dados operacional, de todo é impossível porque os dados estão codificados. O Data Warehouse não temos, mas vamos ver se este ano conseguimos começar a construir uma base de dados que permita, pelo menos, dar alguma independência, não ao utilizador comum, mas a utilizadores que já tenham algum grau de conhecimento em trabalhar dados. Temos servidores aplicacionais, temos três em ambiente de produção. Utilizamos um software que é freeware, é o JBoss. É o nosso software dos servidores aplicacionais. Também temos uma aplicação da gestão dos perfis e das permissões, mas durante 2011 vamos ter de passar a ter o SEI integrado em vez de estar integrado com o ITIM, que actualmente é a aplicação de gestão de perfis, e passar a ter para uma outra, a UTIS[176] que é que a RNSI disponibilizou para todas as forças. Está previsto para este ano essa mudança.

5- Na sua opinião, um sistema de apoio à decisão, gerador de conhecimento (a nível estratégico e operacional), é vantajoso para a PSP?

Sim. Era muito vantajoso, porque já temos muitos anos de informação. Na minha opinião temos informação que não estamos a utilizar, nem sabemos o que temos. É como se fosse uma caixa negra (dos dados que estão aqui registados) e que não se consegue fazer uma análise porque não há ferramentas para isso. Mesmo fazer uma análise de suspeitos, há tantas áreas que a polícia, dentro dos dados registados, podia explorar, e não se consegue porque não há ferramentas para isso.

[176] Unidade de Tecnologias de Informação de Segurança (Fonte: http://www.rnsi.mai.gov.pt/Pages/defaultint.aspx)

6 - Na sua opinião, e a nível tecnológico, o que é que a PSP realmente precisa, para poder obter informações criminais/policiais, através das Tecnologias de Informação?

Precisa de uma nova base de dados construída com base na actual. Mas, esta que é o nível mais baixo, precisa de uma que não tenha os dados a nível de inserção. Quando eu vou fazer o registo de um processo, eu registo os dados correctamente em tabelas de uma base de dados normalizada. Isto é, eu não tenho redundância, portanto, tenho os dados, digamos, em caixinhas. Em que cada caixinha tem a sua função, corresponde a determinado tipo de informação. Ao passo que, se eu pretendo explorar a informação, não posso explorar a informação a um nível tão baixo porque os dados têm de estar preparados, isto é, agrupados de determinada forma. Além disso, há tabelas técnicas que não interessam, são desenhadas para a inserção e para garantir a integridade dos dados. Se eu quiser explorar a informação, tenho de fazer uma base de dados construída de forma completamente diferente. Tenho de pegar em tudo isto e desnormalizar. Fazer um trabalho completamente ao contrário que é criar redundância. Para isso, tenho de ter um servidor bastante potente, porque o volume de dados é enorme. Porque se eu criar redundância significa que uma informação vai estar repartida em várias tabelas, a mesma informação. Mas, preciso dela porque é mais fácil para explorar os dados. É preciso de ferramentas por exemplo de DM. Eu acho que o versão actual do SQLServer que vamos instalar já tem ferramentas DM que permitem explorar os dados. Mas falta-nos a construção da "tal" base de dados, que é o DW que permita usar o DM, ou seja, ferramentas que permitam explorar os dados e tirar partido.

7- Está familiarizado com o termo Data Mining e Descoberta de Conhecimento em Bases de Dados? E que entende por estes conceitos?

Eu nunca utilizei. Embora tenha estudado já vão uns anos, na universidade. Nunca utilizei nenhuma ferramenta para fazer isso. Mas, eu acho que nesta área de negócio, que é importantíssimo, atendendo à actividade policial que cada vez mais impõe uma actividade pró-activa, saber e tentar conseguir perceber tendências. Eu acho que era uma grande melhoria para o sistema, os polícias terem acesso a essas previsões que nos permitisse tomar decisões ou seja, que fossem assertivas, que conseguissem desempenhar melhor a sua actividade. Se conseguirmos construir a base de dados, ou começar a construir a base de dados porque ela é enorme, mas eu penso que a ferramenta é secundária, porque as

ferramentas arranjam-se. Agora o problema aqui, é começar a construir a base de dados para dar acesso à informação. É o mais importante para a Polícia, porque a Polícia não pode andar a reagir às coisas, ou seja, conseguir ter uma pro-actividade seria o ideal.

8-Na sua opinião, seria benéfica a implementação de um Sistema de DM ou de DCBD para orientar a actuação policial?

Para orientar a decisão, o planeamento das decisões a tomar e onde reforçar os recursos. Nesse sentido, ia-lhe falar da geo-referenciação. Para um comandante ou para quem a função de decisão, olhar para um mapa e conseguir ver e analisar os crimes, o tipo, em que ruas/locais em que eles estão a acontecer. Uma coisa é ler números e ler informação alfanumérica, outra é puder olhar para um mapa e verificar que numa determinada rua eu tenho um maior número de ocorrências que noutras ruas paralelas ou nos seus limites, por exemplo. Portanto, isso ajuda à decisão de reforçar a actividade operacional, mas isso conto eu, durante este ano, arrancar com esse projecto. Espero eu que amanhã, vai ser a primeira reunião de análise para começarmos a utilizar uma tecnologia que está disponível no Google, que são aplicações que eles disponibilizam para ver se conseguimos registar a ocorrência e o utilizador registar o local/rua da ocorrência e conseguir geo-referenciar no mapa. Isto para começarmos a guardar as coordenadas de geo-referenciação para depois apresentar as ocorrências geo-referenciadas num mapa. Isto vai ajudar no planeamento das actividades operacionais. Actualmente é impossível ir ao nível da rua. Ao passo que se estiver representado num mapa, visualiza-se logo através das coordenadas. Vamos ser se este ano conseguimos implementar esse projecto.

9- Tendo em conta a actual conjuntura (policial, económica, criminal e social), acha necessário e pertinente o investimento num sistema de DM ou KDD?

Eu acho que cada vez mais é necessário. Há uma grande preocupação que é a segurança no aspecto da informação. Não cabe a mim restringir a parte policial, porque há aqui uma componente técnica que se integra, nomeadamente, quem decide as regras, quem é que há-de aceder ao quê. Ao longo deste tempo todo em que o SEI tem estado activo, seis ou sete anos, as pessoas que acedem ao SEI (os utilizadores) não têm preocupação nenhuma, primeiro porque há um grande dinamismo em termos de colocações e de funções. Estas não têm preocupação quando abandonam a função/mudam de Comando em alterar/actualizar os seus perfis, em dizer que já não devem ter aqueles perfis e devem ter

outros. Isto é bastante preocupante, porque a informação disponível é tão grande, que é preciso cada vez mais. Acho que quando não tínhamos informação, ninguém se preocupava em actualizar os perfis e ao longo destes anos todos ao não actualizar os perfis, possivelmente, alguns utilizadores já têm os perfis todos e mais alguns, por isso, é preciso haver muito cuidado nesse aspecto. Eu acho que deve ser dada grande atenção é a segurança, ou seja, quem é que deve aceder ao quê. Mas, também andamos a pensar em algumas soluções, por exemplo, aos itens de interesse que qualquer pessoa pode aceder dentro dos processos policiais, estou a falar do perfil mais *soft*, digamos que para uns há permissões para tudo e há outros que não, as coisas são controladas. Quando um perfil base, que permite registar processos e ele ao começar a registar um processo pode aceder aos itens, (ele não pode aceder aos itens directamente) mas dentro de um processo policial ele pode aceder aos itens que pretender. E também, temos detectado (a propósito disto) que os itens sofrem muitas alterações, ou seja, tenho um item registado a determinada altura e depois a evolução natural é ele ser completado há medida que vai sendo utilizado e as pessoas obtêm mais informação sobre aquele item. Mas o que estamos a fazer e vamos implementar numa das próximas versões do SEI é começar a registar quem actualizou o item, que matricula, quando, e onde é que ela pertence (a matricula); para as pessoas começarem a ver que há auditorias e que estão disponíveis a toda a gente. Porque tem que se começar a ter determinados cuidados que uma base de dados que nasce, porque quando arranca inicialmente não é preciso, porque ela não tem informação, e à medida que há mais informação, há que ter outros cuidados. E neste caso, o SEI de futuro tem de ter um investimento a nível da segurança dos dados.

10- Neste momento, e na sua opinião, a PSP, conseguiria implementar um sistema de DM/DCBD sem recorrer a parcerias (quer a nível Tecnológico, quer a nível Técnico ou Especializado)?

Acho que sim. Sem recorrer a ajuda, aliás grande parte do desenvolvimento que já é feito, já é interno, E portanto, é assim, este sistema não se compadece com 1 ou 2 recursos, ele precisa constantemente de manutenção, eu estou a falar de manutenção, não é só manter o que está, é manutenção evolutiva também, porque há sempre necessidades novas, e de maneira que, já há um grau de conhecimento da base de dados, da forma como o SEI, tecnicamente está implementado, o que permite termos bastante autonomia. Agora é assim, como lhe digo e já lhe disse á bocado era uma das intenções deste ano começar a construir

uma base de dados, portanto, que depois irá evoluir, mas para preparar e para tornar a dar alguma autonomia, a quem precisa de ter acesso aos dados, a quem precisa de ler os dados. O SQLserver 2008 já tem já tem ferramentas de Data Mining para poder explorar os dados, então resta-nos construir ou começar a construir, e ir desenvolvendo a base de dados. Portanto, se não houver investimento para adquirir fora uma equipa de desenvolvimento para fazer isso, a polícia, através dos técnicos que tenho aqui no gabinete, termos que andar para a frente. Porque é bom, para que quem tem que ter acesso a essa informação se sinta confortável e autónomo para a explorar, e é bom para nós técnicos porque libertamo-nos de tarefas que actualmente são penosas, porquê desenvolver tudo à medida (à mão).

11 - Actualmente, está a ser estudado ou está planeado algum estudo acerca de um sistema de apoio à decisão como o DM/DCBD?

Não, porque agora vamos arrancar com um projecto da geo referenciação, e este ano também temos, várias necessidades de implementação mesmo a nível técnico, que é nomeadamente actualização de software. Ou seja nós vamos actualizar o software da base de dados SQLServer, temos que actualizar o software dos servidores aplicacionais que são novas versões do JBoss, em que se não o fizermos começamos a andar para trás, e nomeadamente isto é importantíssimo porque cada vez mais temos mais integrações com entidades externas, e ao integrarmos com entidades externas a tecnologia vai evoluindo e se nós não actualizamos os softwares dos nossos servidores, cada vez é mais penoso porque temos que desenvolver à medida e se calhar se tivéssemos versões actualizadas do software nos servidores, já escusávamos de desenvolver à medida. Portanto, já haviam objectos e classes já pré construídas que nos permitiam reutilizar, de maneira que este ano é uma das grandes prioridades mas, já te respondendo, como vejo que sinto que é uma necessidade realmente começar a fazer qualquer coisa nesse aspecto de disponibilizar dados eu acho que este ano isso vai acontecer

12 - Acha que as capacidades predictivas das ferramentas de DM, são uma mais-valia para a actuação policial?

Eu acho que sim até porque atendendo ao grau da crise social que estamos, e que se vai impor daqui para a frente, e cada a vez mais, que o nível de agressividade e instabilidade vai ser cada vez maior. E eu acho que para a policia, conseguir prever caminhos ou tendências acho que era uma forma de proteger, e de estar à espera de qualquer coisa que

possa vir a acontecer para estar alerta e para estarem melhor preparados para poder agir em conformidade com algo. Ou que há uma tendência que foi analisada, e portanto se calhar não são apanhados desprevenidos, e vão conseguir com certeza responder, com maior segurança, aquilo que possa vir a acontecer. Porque o que se vê, é cada vez mais manifestações, distúrbios organizacionais, de maneira que seria bom para a policia ter uma previsão e estudar tendências. Fugindo agora um pouco à pergunta, na quinta feira eu fui à reunião, que se dá de 15 em 15 dias com todos os comandantes de Comando por vídeo conferência, e fiquei muito agradada, porque vi que o Departamento de Informações, que é uma coisa que eu não via na policia, também estou cá à uns anos, que é o poder prever a propósito da manifestação da geração à rasca. E que realmente fizeram um estudo, seguindo o facebook os grupos, as pessoas que se iam inscrevendo para a manifestação, que acho que isso é um bom trabalho, gostei desse nível de investigação. Eu acho que o caminho é esse, porque o estar a agir assim sem saber muito bem (isto agora vai haver uma manifestação, mas nem se sabe quantos são, nem de onde é que vêm), e acho que fazer esse estudo e fazer esse levantamento, só pode ser positivo. Eu acho que a capacidade para prever as coisas é muito importante para a polícia, e só pode ser esse o caminho. Porque no fundo, quando a polícia retira os dados do sistema (totais), é estatística que no fundo, até à bem pouco tempo, servia para disponibilizar às entidades externas, e agora eu noto que neste últimos 3 anos, as coisas têm mudado, eu vejo outro tipo de atitude. Atendendo também ao facto de a informação não estar tão disponível, porque se esta estivesse disponível para essas ferramentas trabalharem os dados, então seria muito bom. Mas entendendo ao que existe, já há uma grande utilização da informação.

ANEXO V

ANÁLISE DE RESULTADOS: TABELAS E GRÁFICOS

TABELA 1

Tem conhecimento da existência de ferramentas de Data Mining?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Sim	43	34,4	34,4	34,4
	Não	82	65,6	65,6	100,0
	Total	125	100,0	100,0	

GRÁFICO 1

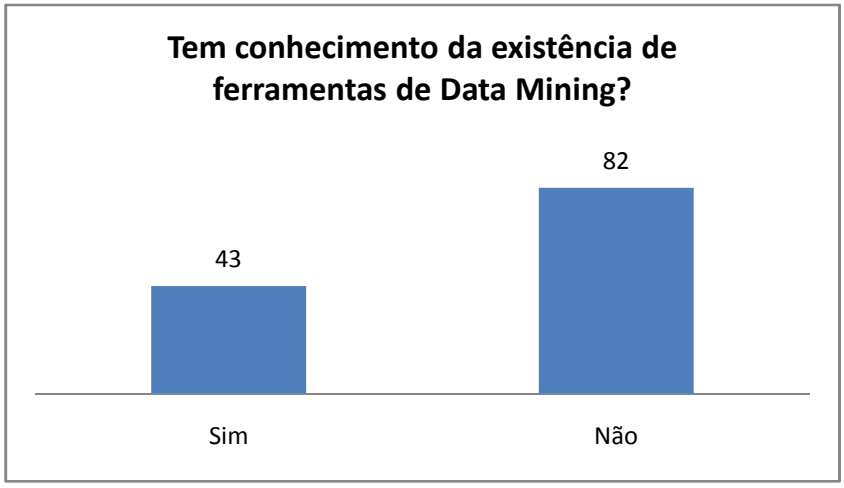


GRÁFICO 2

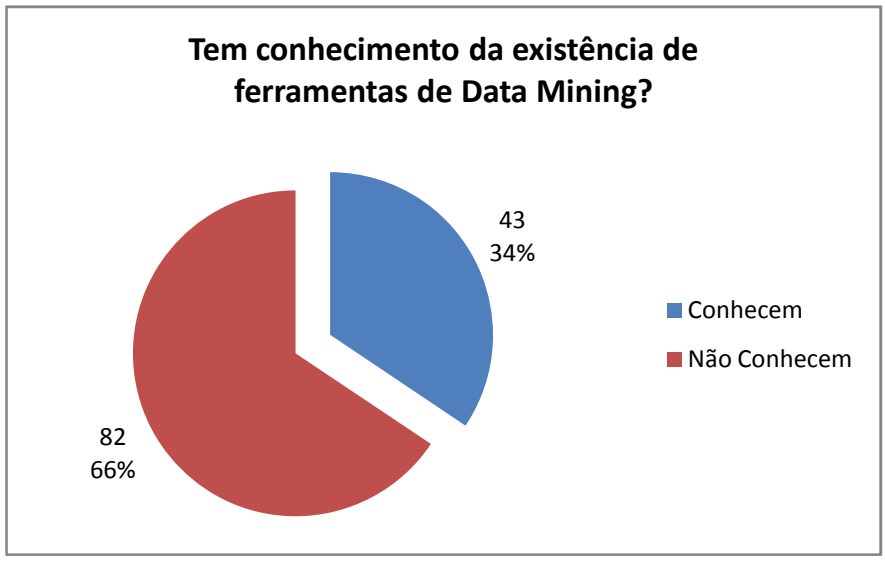


TABELA 2

Comando * Numa escala de 1 a 5, domina esta tecnologia do ponto de vista conceptual? Crosstabulation

		Numa escala de 1 a 5, domina esta tecnologia do ponto de vista conceptual?					
		Não domina nada	Domina pouco	Neutro	Domina	Domina Muito	Total
Comando	Coimbra	3	1	1	0	0	5
	COMETLIS	4	11	18	5	0	38
Total		7	12	19	5	0	43

GRÁFICO 3

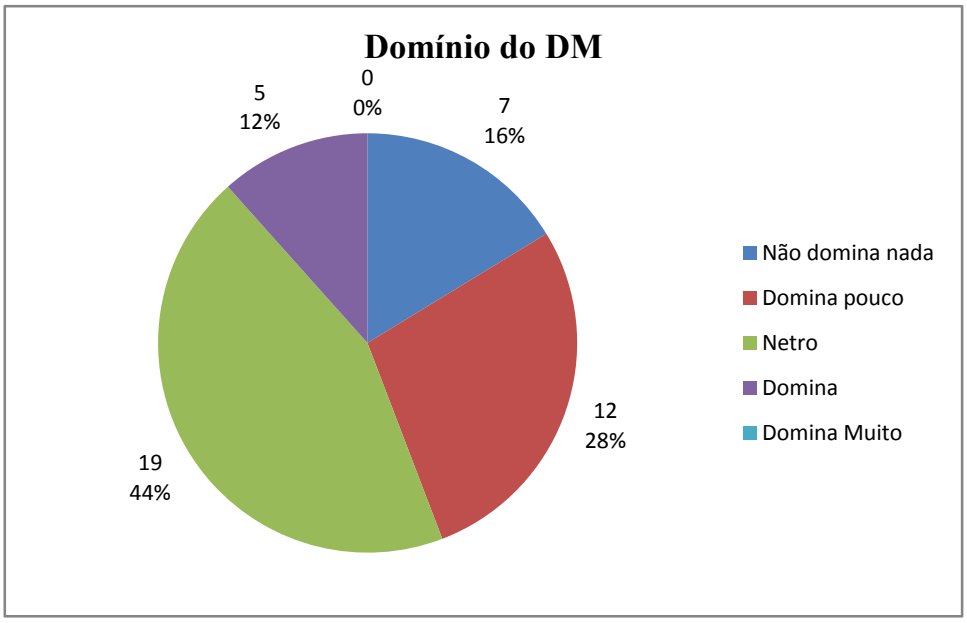


TABELA 3

			Tem conhecimento da existência de ferramentas de Data Mining?		Total
			Sim	Não	
Qual a natureza das funções que desempenha?	Funções Operacionais	Count	37	60	97
			38,1%	61,9%	100,0%
			86,0%	73,2%	77,6%
	Funções de Apoio Operacional	Count	6	22	28
			21,4%	78,6%	100,0%
			14,0%	26,8%	22,4%
Total		Count	43	82	125
			34,4%	65,6%	100,0%
			100,0%	100,0%	100,0%

GRÁFICO 4

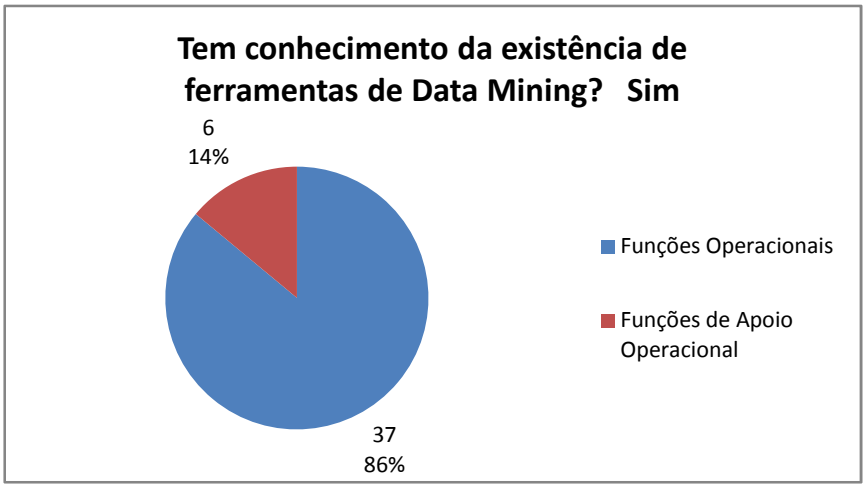


GRÁFICO 5

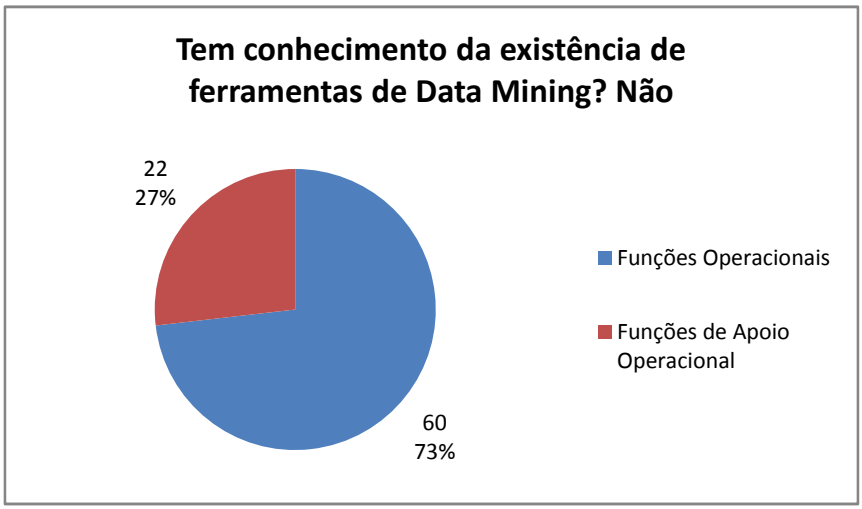
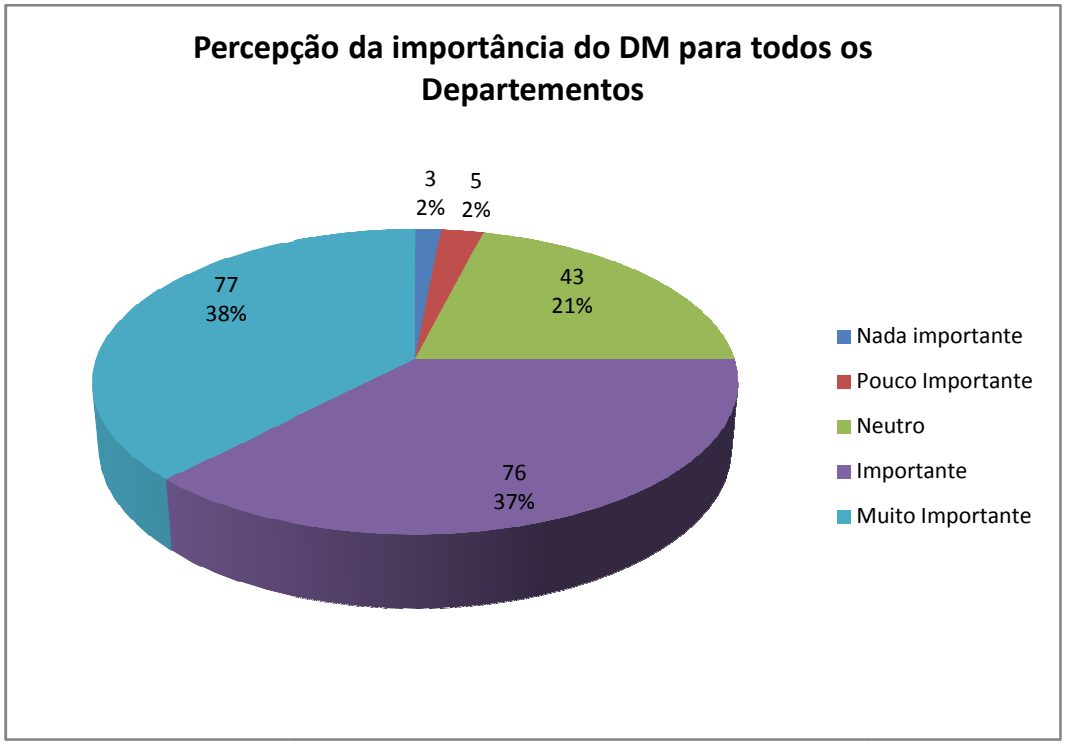


TABELA 4[177]

Percepção da importância do DM para todos os departamentos	
Respostas	N.º Total de Respostas
Nada importante	3
Pouco Importante	5
Neutro	43
Importante	76
Muito Importante	77

GRÁFICO 6



[177] Os valores desta tabela foram atingidos através da soma dos respectivos valores constantes das Tabelas 5, 6, 7, 8 e 9.

TABELA 5

Direcção Nacional					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nada Importante	0	0	0	0
	Pouco Importante	0	0	0	0
	Neutro	11	8,8	25,6	25,6
	Importante	9	7,2	20,9	46,5
	Muito importante	23	18,4	53,5	100,0
	Total	43	34,4	100,0	
Missing	System	82	65,6		
Total		125	100,0		

GRÁFICO 7

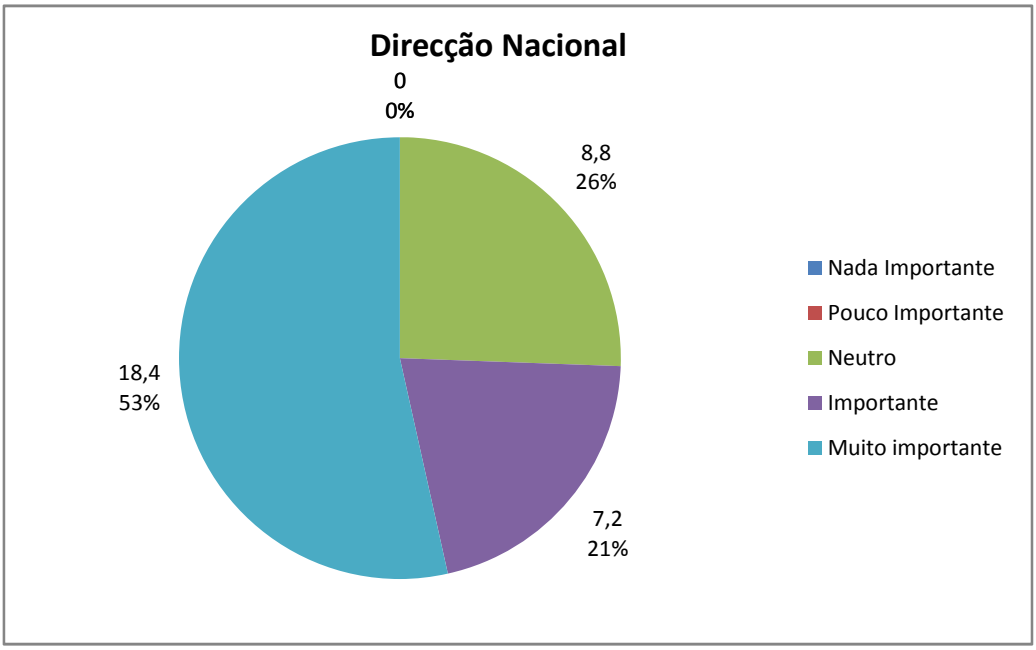


TABELA 6

Comando Metropolitano					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nada Importante	0	0	0	0
	Pouco Importante	0	0	0	0
	Neutro	8	6,4	18,6	18,6
	Importante	11	8,8	25,6	44,2
	Muito importante	24	19,2	55,8	100,0
	Total	43	34,4	100,0	
Missing	System	82	65,6		
Total		125	100,0		

GRÁFICO 8

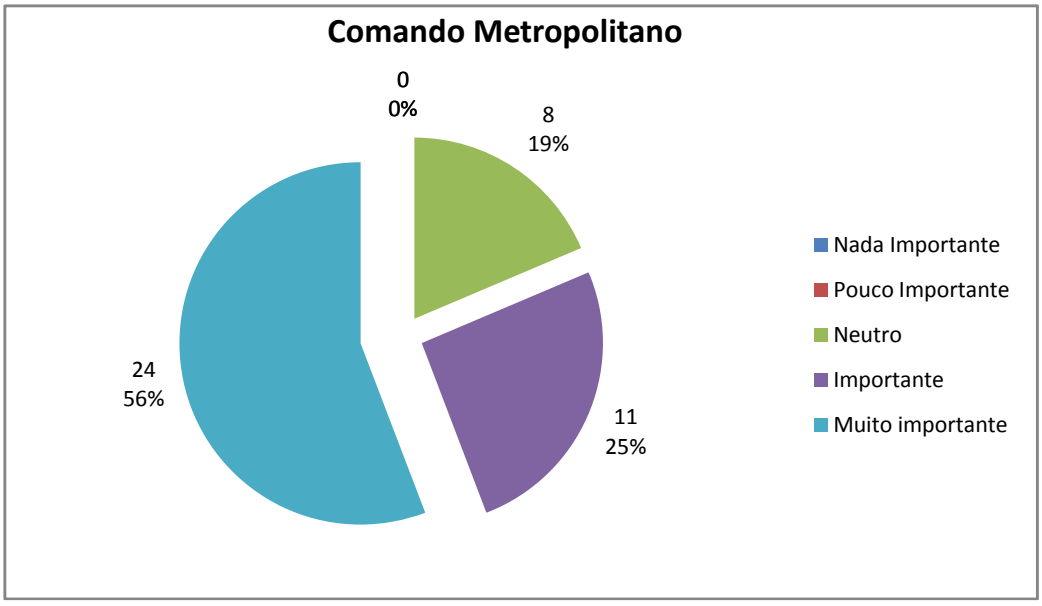


TABELA 7

Comando Distrital					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nada Importante	0	0	0	0
	Pouco Importante	0	0	0	0
	Neutro	7	5,6	16,3	16,3
	Importante	19	15,2	44,2	60,5
	Muito importante	17	13,6	39,5	100,0
	Total	43	34,4	100,0	
Missing	System	82	65,6		
Total		125	100,0		

GRÁFICO 9

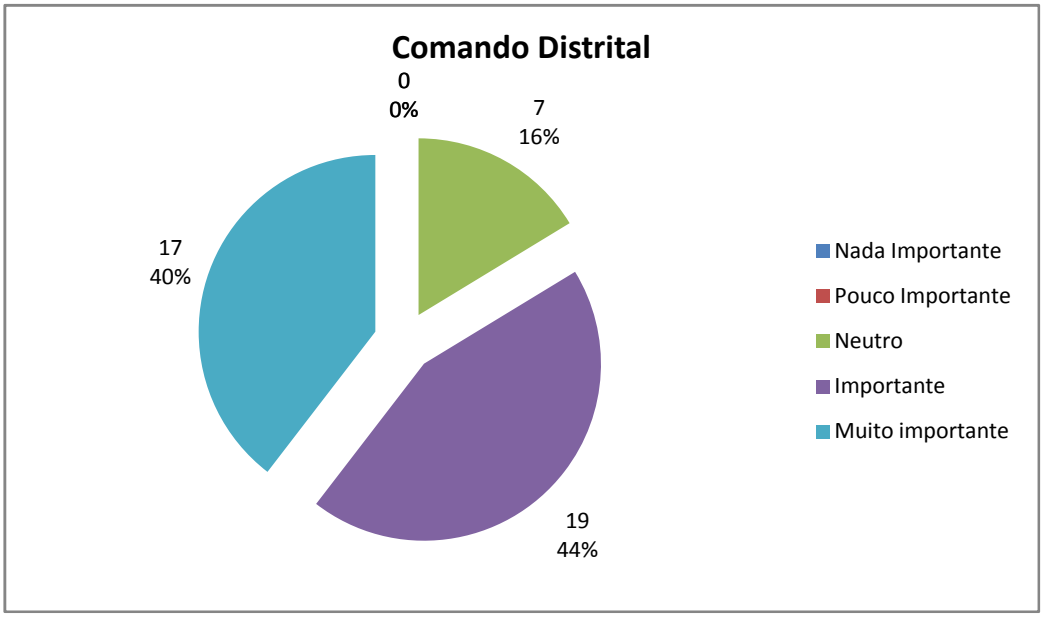


TABELA 8

Divisão					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nada Importante	0	0	0	0
	Pouco importante	2	1,6	4,7	4,7
	Neutro	12	9,6	27,9	32,6
	Importante	21	16,8	48,8	81,4
	Muito importante	8	6,4	18,6	100,0
	Total	43	34,4	100,0	
Missing	System	82	65,6		
Total		125	100,0		

GRÁFICO 10

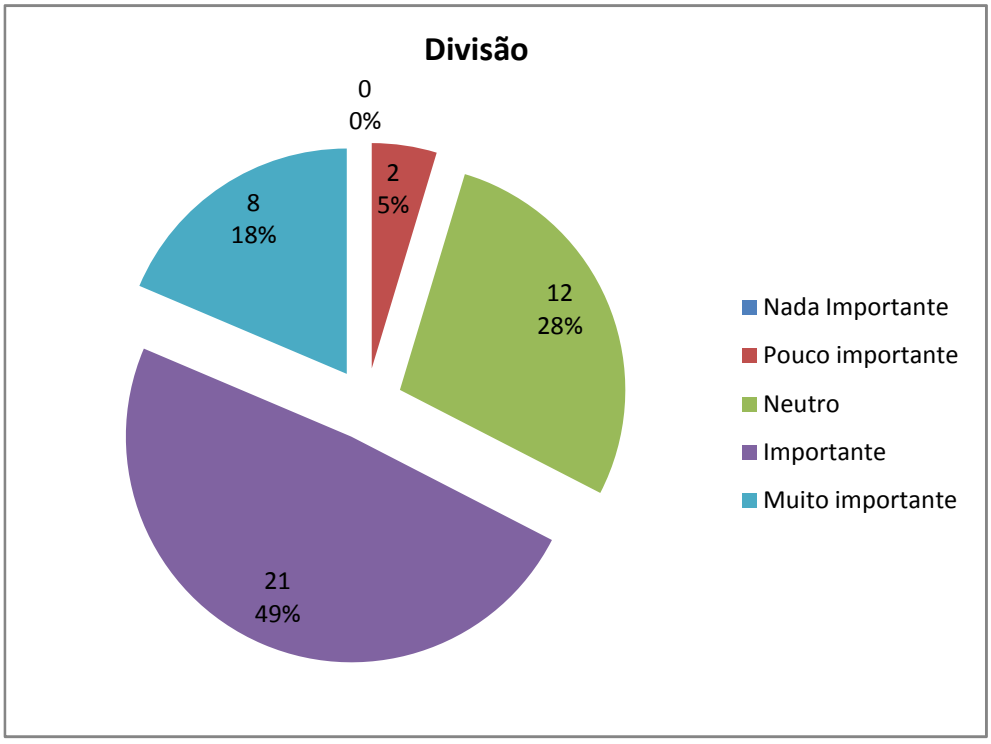


TABELA 9

Esquadra					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nada importante	3	2,4	7,0	7,0
	Pouco importante	3	2,4	7,0	14,0
	Neutro	16	12,8	37,2	51,2
	Importante	16	12,8	37,2	88,4
	Muito importante	5	4,0	11,6	100,0
	Total	43	34,4	100,0	
Missing	System	82	65,6		
Total		125	100,0		

GRÁFICO 11

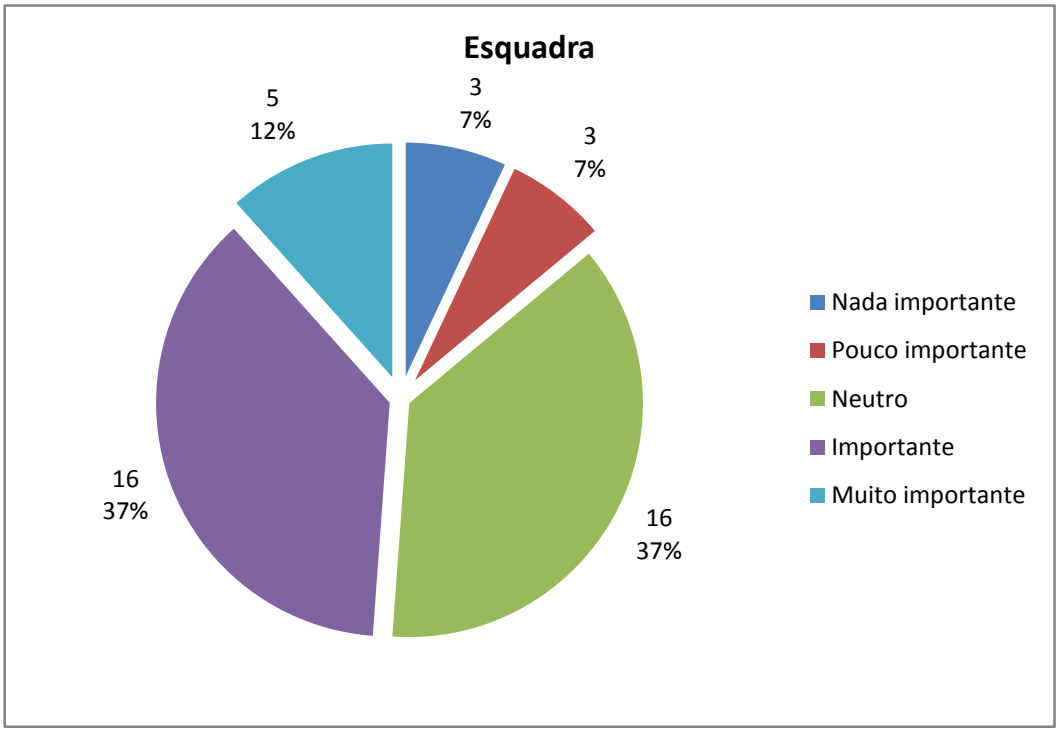


TABELA 10

			Tem conhecimento da existência de ferramentas de Data Mining?		
			Sim	Não	Total
Comando	C. D. Coimbra	Count	5	7	12
			41,7%	58,3%	100,0%
	COMETLIS	Count	38	75	113
			33,6%	66,4%	100,0%
Total		Count	43	82	125
			34,4%	65,6%	100,0%

GRÁFICO 12

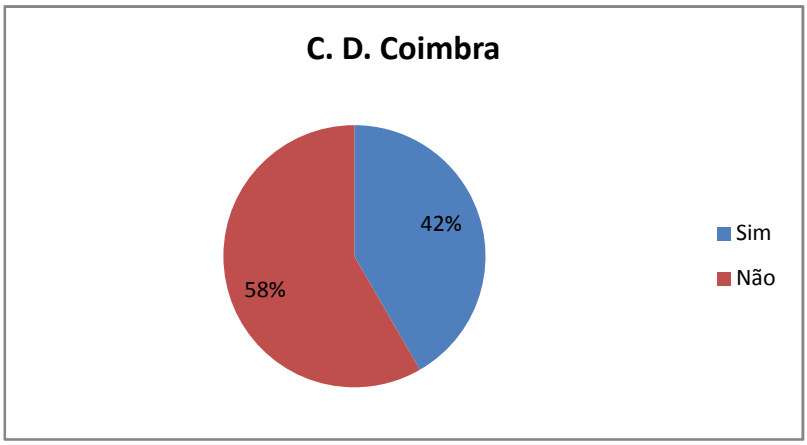
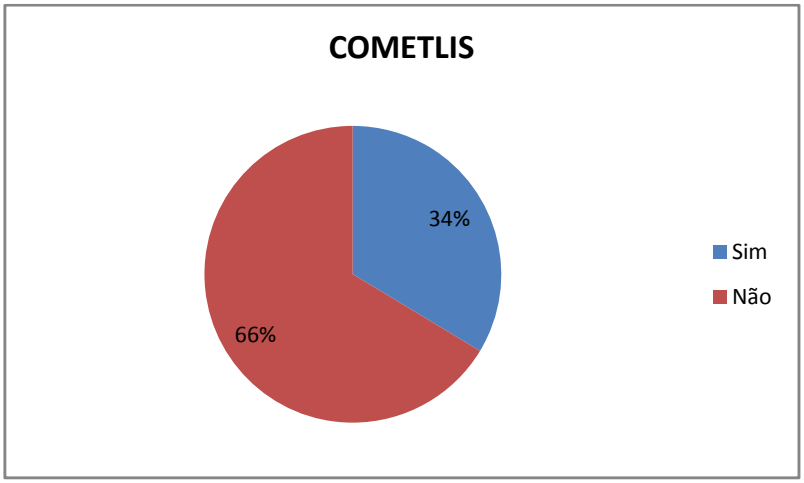


GRÁFICO 13



ANEXO VI

Metodologia CRISP-DM (Santos & Azevedo, 2005: 25-34; McCue, 2007: 49-52)

Concebida em 1996, a metodologia CRISP-DM surgiu do crescente interesse do mercado nas tecnologias de DM e na necessidade de se padronizar o processo (DM). Os conhecimentos académicos e teóricos do DM são a base da metodologia CRISP-DM, aplicados no âmbito das necessidades e problemas dos negócios.

A metodologia em apreço, é relatada como sendo um processo hierarquizado, com um ciclo de vida que se desenvolvido em seis fazes:

1. Estudo do Negócio;
2. Estudo dos Dados;
3. Preparação dos Dados;
4. Modelação;
5. Avaliação;
6. Implementação.

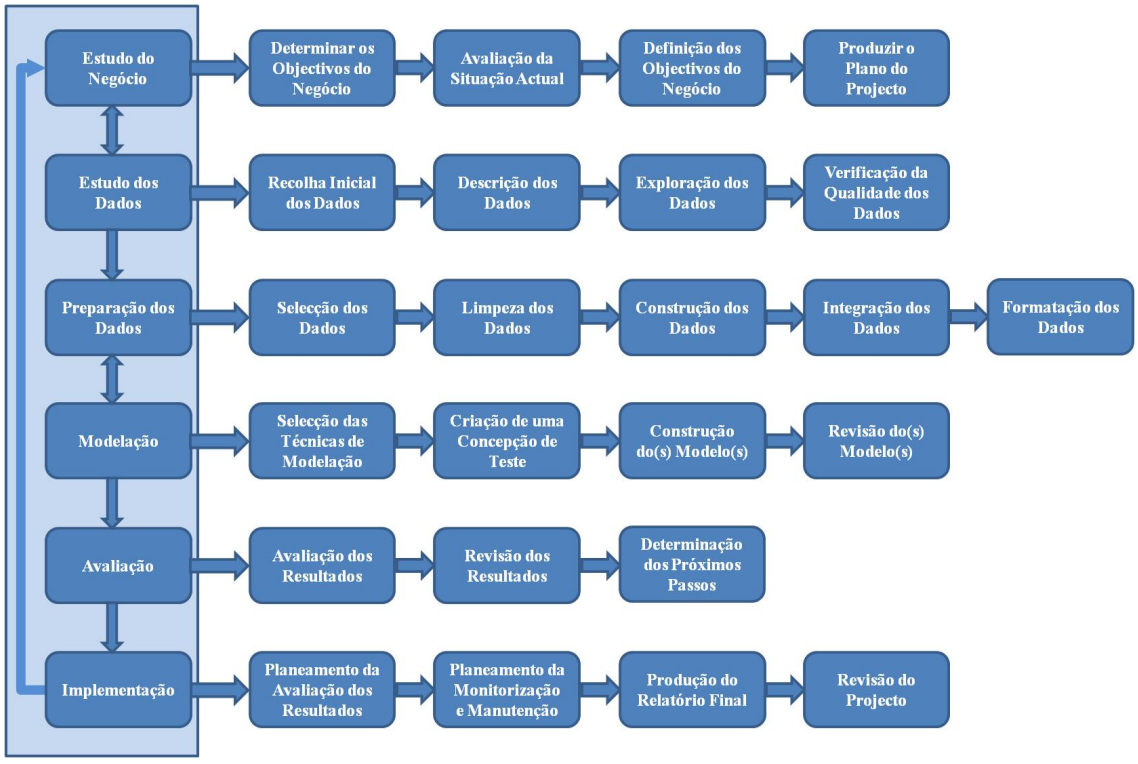
Estas fases não obedecem obrigatoriamente a uma sequência cíclica e continua, uma vez que, o resultado das fases ou da tarefa de uma determinada fase, pode não ser satisfatório ou até válido, o que implica o retrocesso a uma das fases anteriores (se tal for admissível para a fase em questão). Assim, a figura seguinte é a representação das relações entre as diversas fases bem como das relações entre estas (Figura 3).

A fase de estudo do negócio, foca-se na análise dos objectivos principais do projecto, bem como, nos requisitos (funcionais técnicos e temporais) do negócio. O primeiro grande passo é estudar a necessidade da realização de um projecto de DM, compreendendo o problema a resolver, equacionando os objectivos a atingir e descobrindo os factores mais importantes que influenciem os resultados do processo. Esta fase contempla as seguintes tarefas

1. Determinação dos objectivos do negócio: determinação de todos os detalhes acerca da situação actual do negócio e descrição dos objectivos primários dos clientes. Enumeração dos objectivos do negócio, bem como, dos critérios de sucesso deste.
2. Avaliação da situação actual: procede-se à elaboração de uma listagem dos recursos humanos, dos dados, do hardware e do software disponíveis para o projecto. Elaboração de um programa de realização, compreensibilidade, qualidade dos resultados, segurança, aspectos legais, ameaças ou eventos que possam

comprometer o projecto, bem como, dos respectivos planos de contingência. Elaborar ainda uma análise de custos e benefícios.

3. Definição dos objectivos do DM: Descrever os objectivos do DM e os critérios de sucesso deste (EX.: Classificação, Previsão, Segmentação, Associação e Visualização).



O ciclo do CRISP-DM

Fonte: Adaptado Santos, M. Filipe; Azevedo, Carla (2005), *Data Mining – Descoberta de Conhecimento em Bases de Dados*. FCA – Editora de Informática, Lda.

4. Elaborar um plano para o projecto: elaboração de um plano para o projecto, que inclua, a duração, os recursos, as fases, as fases intermédias, as interacções entre os processos, entradas, saídas e dependências. Definição dos pressupostos das ferramentas e técnicas a utilizar.

A fase de estudo dos dados, envolve com a recolha destes e a sua análise em termos da qualidade e fiabilidade dos dados. Antes da aplicação das técnicas de DM é importante a observância dos seguintes itens:

1. Recolha inicial dos dados: aquisição dos dados e a sua compreensão. Listagem dos dados adquiridos, bem como, da sua localização, métodos de aquisição, problemas e soluções encontradas;
2. Descrição dos dados: depois da recolha, é necessário descrevê-los, reconhecendo o formato dos dados, número de registos nas tabelas e identificação dos registos;

3. Exploração dos dados: resulta uma lista inicial de hipóteses, e do seu impacto no restante projecto. São utilizados gráficos e histogramas, que indicam as características dos dados, facilitando a sua exploração;
4. Verificação da qualidade dos dados: Listagem de problemas de qualidade, bem como, das possíveis soluções para estes.

A fase de preparação dos dados, resume-se à construção de um conjunto final de dados, na qual serão realizadas as várias tarefa do processo de DM. É imperativo realizar-se um conjunto de optimizações, através do desempenho de tarefas de, selecção de tabelas, registos e atributos, bem como de limpeza e transformação dos dados, usando para este último, algumas ferramentas de modelação (fase seguinte). Esta fase engloba os seguintes procedimentos:

1. Selecção dos dados: escolha dos dados a utilizar na análise, tendo em conta, os objectivos de DM, as restrições técnicas, a qualidade, os limites e os tipologia dos dados;
2. Limpeza dos dados: normalização dos dados e tratamento dos dados omissos;
3. Construção de dados: derivação de novos dados, criação de novos registos e transformação de dados;
4. Integração de dados: criação de novos registos e valores através da combinação de várias tabelas ou registos;
5. Formatação de dados: modificações sintácticas dos dados, que sem mudar o seu significado, mantêm a integridade, sendo um requisito da ferramenta de modelação (próxima fase).

A fase de modelação engloba a selecção das técnicas de modelação (ex: Árvores de Decisão, RNA, Algoritmo Genético [AG], entre outros), sendo que os seus parâmetros são ajustados de forma a optimizar os resultados. Estas técnicas carecem de uma preparação específica dos dados, logo, pode ser necessário o retorno à fase anterior para optimizar os dados em consonância com a técnica de DM/Modelagem a utilizar. É apenas neste estágio, que os dados previamente seleccionados e preparados são submetidos para modelação, logo, as técnicas de DM, têm que satisfazer os objectivos iniciais do processo (da fase de estudo do negócio), uma vez que todas as fases anteriores foram direccionadas para estes. Assim, esta fase contempla as seguintes tarefas:

1. Selecção de técnicas de modelação: escolha da técnica mais apropriada, tendo atenção ao tipo de problema, às ferramentas e aos objectivos do DM;

2. Criação de uma concepção de teste: Definição de um procedimento ou de um mecanismo para testar o desempenho do modelo;
3. Construção do modelo: depois de escolhida a ferramenta de modelação, procede-se à sua aplicação ao conjunto de dados preparados anteriormente, de forma a criar um ou mais modelos. Os modelos resultantes devem ser convenientemente interpretados e o seu desempenho explicado.
4. Revisão do modelo: os modelos devem ser interpretados de acordo com o conhecimento, os critérios de sucesso do projecto de DM e com o mecanismo de teste definido. É a avaliação do sucesso da aplicação e a discussão dos resultados do DM, no contexto do negócio.

A fase de avaliação tem o principal objectivo de avaliação da utilidade dos modelos gerados. São revistos os passos executados e verificados os modelos de forma aperceber se estes possibilitam atingir os objectivos inicialmente traçados. Nesta fase, incluem-se as tarefas subsequentes:

1. Avaliação dos resultados: avalia-se se o modelo atingiu os objectivos do negócio (e do DM), procurando determinar se o modelo é deficiente em algum ponto;
2. Revisão do processo: revisão de todas as fases, realçando as actividades que foram esquecidas ou precisam de ser repetidas;
3. Determinação dos próximos passos: se os passos anteriores são satisfatórios e os resultados cumpriram os objectivos, então passa-se à fase de implementação. Caso contrário, deve proceder-se a nova iteração das fases de preparação dos dados, modelação e avaliação, utilizando novos parâmetros.

A fase de implementação engloba a organização do conhecimento extraído, bem como, a sua apresentação para que o utilizador o possa utilizar. Nesta fase, distinguem-se dois caminhos distintos: geração de um simples relatório ou a nova implementação de todo o processo de DM (dependendo da satisfação ou não dos requisitos/objectivos iniciais). Nesta fase, realizam-se as seguintes tarefas:

1. Planeamento da avaliação dos resultados: definição das estratégias de implementação dos resultados do DM;
2. Planeamento da monitorização e manutenção: elaboração de estratégias de monitorização e manutenção, caso a implementação dos modelos ocorra diariamente. Bem como da necessária monitorização da sua correcta aplicação.

3. Produção do relatório final: elaboração de um relatório final, resumindo os pontos mais importantes do projecto, experiência adquirida e explicação dos resultados produzidos;
4. Revisão do projecto: avaliação dos pontos correctos e errados, mencionando o que correu bem ou o que necessita de ser melhorado. Resumindo as experiências, as armadilhas, as aproximações erradas e a forma de selecção das técnicas de DM.

A aplicação da metodologia CRISP-DM fornece um conjunto de resultados, na forma de documento (relatório), ao longo de todo o processo, nomeadamente:

1. Estudo do Negócio;
2. Relatório do conjunto inicial de dados;
3. Relatório da descrição de dados;
4. Relatório da qualidade dos dados;
5. Relatório da descrição da amostra de dados;
6. Relatório de modelação;
7. Relatório de avaliação;
8. Plano de monitorização;
9. Manutenção;
10. Relatório final.

A aplicação desta metodologia em projectos de DM, permite garantir uma maior celeridade, menores custos de execução, maior segurança, assim como, a sua maior exequibilidade e viabilidade. (Santos & Azevedo**,** 2005: 34).

Metodologia SEMMA (Santos & Azevedo**,** 2005: 35-38; Giudici, 2003: 11/12)

A metodologia SEMMA, como já foi referido anteriormente, foi desenvolvida pelo Instituto SAS[178]. Esta metodologia é considerada como uma ferramenta auxiliar para conduzir um projecto, em todas as suas etapas, desde a especificação do problema do negócio, até à sua implementação. O Instituto define Data Mining como o processo de extracção de informação valorosa e de relações complexas, de um grande volume de dados (SANTOS & AZEVEDO**,** 2005: 35). Esta metodologia estabelece alguns elementos básicos de DM, sem impor um caminho rígido e predeterminado para o projecto. Este fornece um

[178] O acrónimo SAS inicialmente significava "*statistical analysis system*", ou seja, sistema de análise estatística. Mas devido à crescente procura deste software, em 1976 foi fundada em empresa SAS. http://www.sas.com/company/about/history.html#s1=1 (15/03/11)

processo lógico, que permite aos analistas comerciais e aos especialistas estatísticos, alcançar os objectivos dos projectos de DM, através da possibilidade de escolha dos elementos gráficos de interface do utilizador (*Graphical User Interface* - GUI), que estes necessitem. A representação visual para esta estrutura é um fluxograma do processo (*Process Flow Diagram* – PFD), que graficamente ilustra os passos dados para completar o projecto de DM. O método SEMMA é definido pelo Instituto SAS como uma referência estrutural que pode ser usada para organizar as fases do projecto de DM. Esquematicamente, o método consiste numa série de passaos que têm que ser seguidos para completar a análise dos dados (Giudici, 2003: 11) Assim, o processo está dividido em cinco fases, que perfazem o acrónimo SEMMA, (*Sample* [Amostragem], *Explore* [Exploração], *Modify* [Modificação], *Model* [Modelação] e *Assessment* [Avaliação]), os quais são melhor definidos seguidamente (Santos & Azevedo**,** 2005: 36-38; Giudici, 2003: 11/12):

1) Amostragem: o processo inicia-se com a extracção uma parte dos dados do universo existente. A amostra[179] deve ser significativa onde cada elemento dos dados tem a mesma hipótese de ser incluído nesta. Por outro lado, a amostra tem que abranger elementos em quantidade suficiente para conter informação importante e ao mesmo tempo ser pequena o suficiente, para ser analisada rapidamente. Manipular uma amostra é mais fácil e mais rápido do que manipular todo o universo dos dados. Basear o processo de DM numa amostra representativa, reduz drasticamente o volume de tempo de processamento necessário para tirar informação crucial para o negócio
2) Exploração: os dados são examinados para se encontrar à partida, alguma relação ou anormalidades e para entender que dados podem ser os mais interessantes. Após a amostragem, as tendências ou agrupamentos inerentes aos dados, são explorados visualmente ou numericamente (ex: gráficos de distribuição e dispersão, histogramas, tabelas de frequência, mapas de associações e segmentação). A exploração ajuda definir o processo de descoberta. Esta etapa é marcada pela procura de tendências imprevistas e por anomalias, de forma a conhecer os dados e as suas relações.
3) Modificação: esta fase centra-se na realização de transformações, com base nos resultados da exploração. Estas transformações vão desde a inclusão de informação

[179] Amostra é a *selecção de um grupo restrito de indivíduos representativos da população* (Santos & Azevedo**,** 2005: 173).

à selecção ou introdução de novas variáveis. Ou seja, aumento da significância das variáveis, e consequentemente da amostra.

4) Modelação: este passo resume-se à procura de importantes variáveis e modelos, contidos nos dados, que possam fornecer informação. Onde são definidas as técnicas de construção de modelos de DM, no qual se incluem: técnicas de Aprendizagem Automática (AA) (como por exemplo, RNA, Indução de Regras e Árvores de Decisão) e modelos estatísticos (como por exemplo, Regressão Linear e Indução de Probabilidades). É necessário ter em conta que, os dados influenciam os modelos, que por ser singulares, têm determinadas propriedades e características, logo, é necessário haver uma adequação entre estes e as técnicas específicas de DM. (ex: as RNA alcançam melhores resultados através de dados com relacionamentos complexos e não lineares).
5) Avaliação: aqui é avaliada a utilidade e a fiabilidade da informação descoberta pelo processo de DM. As regras dos modelos, são aplicados ao ambiente real da análise. O grande objectivo é aferir o desempenho do modelo. Ou seja, este é aplicado a uma amostra de dados seleccionada para este fim (conjunto de teste). Se o modelo for válido, este deve funcionar tão bem como na amostra que serviu de base para a sua construção. Por vezes, pode ser necessário proceder a alguns ajustes.

Metodologia PMML (Santos & Azevedo**,** 2005: 38-39; Maimon & Rokach, 2010: 1112; Hornick et. al., 2007: 452-454; Ye, 2003: 454-455).

A metodologia PMML, mais conhecida como Especificação Predictive Model Markup Language, é uma linguagem de *markup*[180] XML[181] para descrever os modelos estatísticos e os modelos de DM. O seu principal objectivo é permitir o intercâmbio de modelos de DM entre sistemas, bem como entre fornecedores de implementações. O PMML suporta: a descrição do output do modelo de DM (em campos de dados necessários), a transformação necessária para preparar os dados, assim como os parâmetros que definem o modelo de DM.

[180] Na computação, markup, é o processo de atribuição de tags a vários elementos de um texto, indicando a natureza de cada um, em relação à estrutura do texto (Oxford English Dictionary, 2009).
[181] XML, o acrónimo para eXtensible Markup Language, é um subtipo da Linguagem Padronizada de Marcação Genérica (Standard Generalized Markup Language – SGML), capaz de descrever diversos tipos de dados. O seu propósito principal é a facilidade de partilha de informações através Internet. É um formato para a criação de documentos com dados organizados de forma hierárquica. http://pt.wikipedia.org/wiki/XML

ANEXO VII

DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS[182]

A Descoberta de Conhecimento em Bases de Dados (DCBD) *é um processo que envolve a identificação e reconhecimento de padrões numa BD, de uma forma automática* (Santos & Azevedo**,** 2005: 18). É o *processo de descoberta de conhecimento útil a partir de dados* (Fayyad et al., 1996: 28), onde descobrir conhecimento significa, extrair informações genéricas, relevantes e previamente desconhecidas a partir de grandes conjuntos de dados, sem nenhuma formulação prévia de hipóteses. Tendo a finalidade, de ajudar no processo da tomada de decisão (Santos & Azevedo**,** 2005: 18). Assim, o processo DCBD pode ser definido como, *o processo não trivial, de identificação de padrões válidos, novos, potencialmente úteis, e fundamentalmente compreensíveis, nos dados*[183] (Fayyad et al., 1996: 30).

Segundo Fayyad (et al., 1996: 30) e dissecando esta definição, o processo assume-se não trivial, no sentido em que, vai para além da quantificação computacional, abrangendo a procura de estruturas, modelos, padrões e parâmetros. Os padrões descobertos devem ser válidos, especialmente, para novos dados, ou seja, demonstrando algum grau de certeza, quer para os dados que os fundamentam, quer para os dados ainda por adquirir. Estes padrões devem ser novos (pelo menos para o sistema e preferencialmente para o utilizador) e potencialmente úteis para o utilizador ou para a realização de uma determinada tarefa e/ou decisão. Por último, os padrões devem ser compreensíveis, tanto no imediato como depois de algum pós-processamento[184] (Fayyad et al., 1996: 30).

O processo de descoberta de conhecimento é iterativo[185] e interactivo[186], sendo que, se inicia com a determinação dos objectivos da DCBD, e termina com a implementação do conhecimento descoberto (Maimon & Rokach, 2010: 2). Primeiramente é necessário compreender e desenvolver um estudo, acerca do domínio da aplicação[187], bem como, dos objectivos finais a ser atingidos. Seguidamente, um determinado conjunto de dados é

[182] Também denominado de *Knowledge Discovery Databases*.
[183] *The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* (Fayyad et aL., 1996: 30).
[184] Pós-processamento significa tratar os padrões para que estes sejam entendidos pelo utilizador.
[185] *Iterativo em cada passo, ou seja, pode ser necessário recuar para ajustar passos anteriores* (Maimon & Rokach, 2010: 2).
[186] Pelo facto de permitir *que os analistas, revejam o resultado, formem um novo conjunto de questões para refinar a procura e realimentem o sistema com novos parâmetros* (Santos & Azevedo**,** 2005: 19).
[187] É necessário saber exactamente qual o campo de aplicação da DCBD.

agrupado e organizado, para ser alvo da prospecção/análise. Passa-se então, à fase da limpeza dos dados (*Data Cleaning*[188]). Aqui, através do pré-processamento dos dados, procede-se à sua adequação aos algoritmos a utilizar. Aqui incluem-se operações como, integração de dados heterogéneos, eliminação de dados incompletos e/ou previsão destes (dependendo da metodologia adoptada [a desenvolver no capítulo seguinte]). Os dados pré-processados então sujeitos a uma transformação, com vista ao armazenamento dos mesmos num formato adequado, de modo a facilitar o uso das técnicas de *Data Mining*. Prosseguindo-se, chega-se à fase do *Data Mining* propriamente dita, que se inicia com a escolha dos métodos e técnicas a aplicar, dependendo, primordialmente, do objectivo/tarefa do processo de *Data Mining* (ou seja: Segmentação, Classificação, Previsão, Associação, Sumariação, Visualização e Estimativa). Nesta fase são aplicadas diversas técnicas, como as Redes Neurais Artificiais[189] (RNA), a indução de regras, as Árvores de Decisão[190] e técnicas estatísticas, tanto isoladamente como combinadas. No final do processo, o sistema de *Data Mining* gera um relatório de descobertas, onde os padrões são validados e interpretados, permitindo obter conhecimento (Santos & Azevedo**,** 2005: 19).

Assim, e citando Fayyad (et aL., 1996: 30-31) e Maimon & Rokach (2010: 2-5), o processo de DCBD encontra-se organizado em nove passos (conforme a Figura 2):

1º. Desenvolver um entendimento do domínio da aplicação: este é o passo inicial e preparatório. Aqui procede-se à preparação e representação do que é necessário fazer-se (nomeadamente sobre: transformação de dados, algoritmos a aplicar, representação dos padrões gerados, entre outros), não descorando o entendimento, necessário e imprescindível dos dados e do objectivo do processo de DCBD[191]. Neste momento, é admissível a introdução de algum conhecimento anterior, mas relevante para o sucesso do processo. À medida que o processo de DCBD avança, pode ser necessário fazer-se uma revisão e/ou mudança deste passo. Depois de

[188] As técnicas de *Data Cleaning*, também conhecidas como *Data Cleansing*, focam-se em resolver problemas como: valores de variáveis inconsistentes em diferentes conjuntos de dados, erros nos dados, valores inexistentes e registos desaparecidos (Kamel, 2009: 541).

[189] *Técnica de modelação não-linear complexa, baseada no modelo de neurónio humano. Podem ser utilizadas para prever a saída (valor das variáveis dependentes) a partir de um conjunto de entradas (variáveis independentes) através de combinações lineares das entradas e efectuando transformações não-lineares utilizando uma função de activação* (Santos & Azevedo**,** 2005: 177).

[190] *Estrutura arborescente que permite representar um conjunto de regras hierárquicas que terminam numa ou mais classes (as folhas)* (Santos & Azevedo**,** 2005: 174).

[191] As pessoas que são responsáveis pela implementação do projecto de DCBD, precisam de entender e definir os objectivos do utilizador final, bem como, do ambiente em que o processo de descoberta de conhecimento irá ser realizado (Maimon & Rokach, 2010: 2).

entendidos os objectivos da DCBD, o pré-processamento dos dados começa, como é definido nos próximos três passos[192];

2º. Seleccionar e criar um conjunto de dados onde a descoberta irá incidir (*Criar um conjunto de dados alvo*): ao ter os objectivos definidos, é necessário determinar e seleccionar os dados que irão ser usados na descoberta de conhecimento[193]. Desta selecção pode resultar uma amostra de dados (relativamente pequena) ou grandes conjuntos de dados. Este estádio é de relevante importância, uma vez que, a fase de *Data Mining* aprende e descobre a partir dos dados disponíveis, logo, estes são a base da construção dos modelos e/ou padrões. No caso de faltarem alguns atributos importantes, todo o estudo pode ser inviabilizado. Assim, para o sucesso do processo, é importante considerar, nesta fase, o maior número de atributos possível.

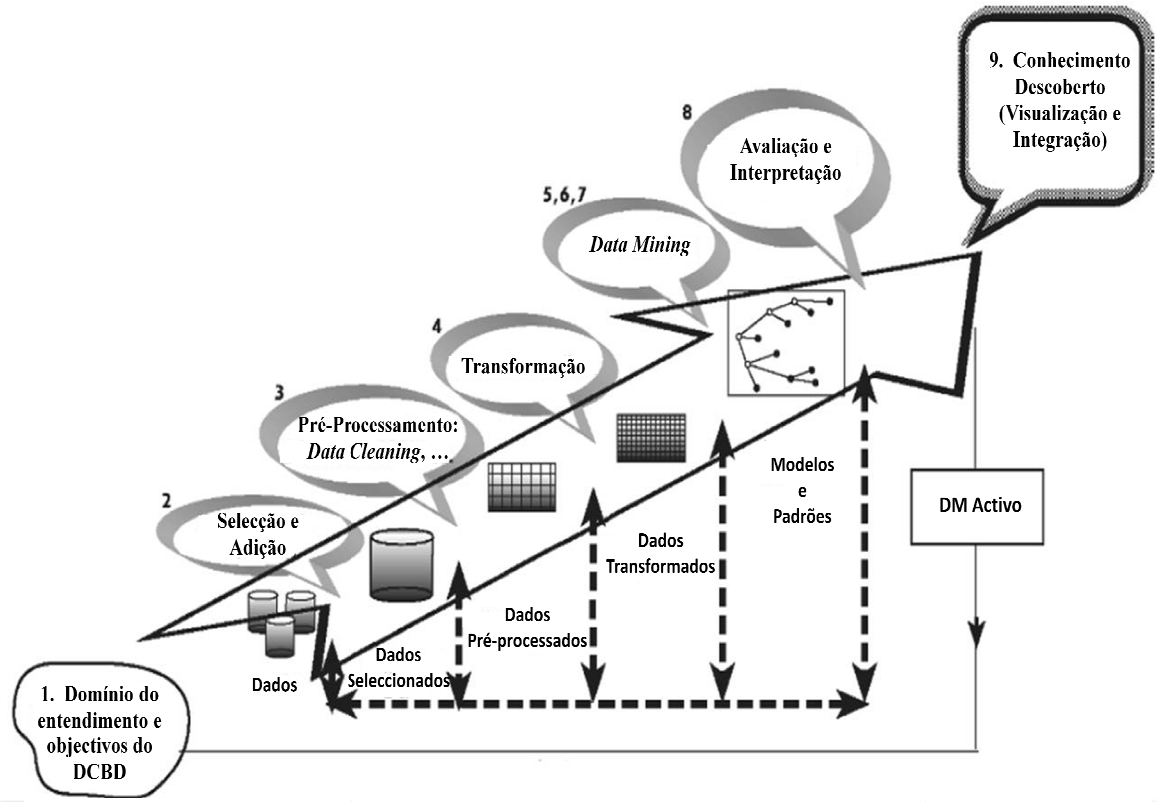


Figura 2: O processo de DCBD

Fonte: Adaptado de Maimon, Oded; Rokach, Lior (2010), *Data Mining and Knowledge Discovery Handbook Second Edition*. Springer.

[192] Denote-se que alguns dos métodos presentes neste primeiro passo, são similares aos algoritmos de *Data Mining*, mas aqui, são usados no contexto de pré-processamento (Maimon & Rokach, 2010: 2).

[193] No qual se incluem: descobrir que dados estão disponíveis, obter dados adicionais necessários e aí integrar todos os dados disponíveis num único conjunto de dados (incluindo os atributos a ser considerados no processo) (Maimon & Rokach, 2010: 2).

Pretendendo-se, *seleccionar ou segmentar os dados de acordo com alguns critérios (ex: pessoas que possuem carro) determinando os* seus *subconjuntos*. (Santos & Azevedo**,** 2005: 20).

3º. Pré-processamento e limpeza: é nesta fase que a fiabilidade dos dados é aperfeiçoada[194]. Inclui a clarificação de dados, como por exemplo, o tratamento de valores desconhecidos, remoção de ruídos[195] e incongruências. Pode ainda envolver, métodos estatísticos complexos, ou a utilização de um algoritmo de *Data Mining,* específico para este efeito. Por exemplo, se há suspeita de que, certo atributo não é suficientemente fiável, ou não existem dados suficientes, esse mesmo atributo pode tornar-se no objectivo de um algoritmo supervisionado[196] de *Data Mining*. É ainda nesta fase, que a definição da tipologia de dados a utilizar, deve ser obrigatoriamente escolhida. Em todo o caso, é um estudo necessário, importante e muitas vezes até clarificante, no que diz respeito, aos sistemas de informação das organizações.

4º. Transformação dos dados: neste estádio, processa-se a preparação e desenvolvimento de dados com melhor qualidade de forma a facilitar o processo de *Data Mining* [197]. Os métodos aqui incluem redução de dimensões (tais como, selecção e extracção de características, escolha e registo de amostras), e transformação de atributos (tais como discretização[198] de atributos numéricos e transformação funcional). Este passo é crucial para o sucesso de todo o projecto de DCBD, no entanto, esta transformação é específica, variando de projecto para projecto. É necessário realçar que os processos de DCBD reflectem-se em si próprios, e mesmo quando errados, orientam o utilizador, para um entendimento acerca da transformação necessária (como um especialista, detentor de um conhecimento exacto sobre uma determinada matéria, que realça e indica o rumo a seguir).

[194] Os dados são reconfigurados para assegurar a construção de uma BD consistente, uma vez que os dados provêm de várias fontes (exemplo: sexo pode ser guardado como "m" e "f" ou "0" e "1") (Santos & Azevedo**,** 2005: 20).

[195] *Os dados possuem ruído quando contêm erros (e.g., dados omissos, valores incorrectos)* (Santos & Azevedo**,** 2005: 177).

[196] Algoritmo que desenvolve um modelo de previsão para este atributo e assim os dados em falta podem ser previstos. (Maimon & Rokach, 2010: 3).

[197] Nos DW, os dados não são voláteis, são classificados por assunto e por natureza histórica, tendendo portanto, a tornarem-se grandes repositórios de dados extremamente organizados (Santos & Azevedo**,** 2005: 21).

[198] Discretização refere-se ao processo de converter atributos ou variáveis contínuas em atributos discretos ou nominais. http://en.wikipedia.org/wiki/Discretization (22/04/2011)

Ao serem completados os passos anteriores, os quatro passos seguintes estão directamente relacionados com a parte de *Data Mining*, onde o foco está nos aspectos algorítmicos de cada projecto.

5º. Escolher a tarefa de *Data Mining* apropriada: neste estádio, decide-se a tarefa de *Data Mining* a utilizar (por exemplo: Sumarização, Classificação, Regressão ou Segmentação). A escolha depende maioritariamente dos objectivos do *Data Mining*, e também dos passos anteriores.

6º. Escolher o algoritmo de *Data Mining*: havendo uma estratégia, agora decide-se a táctica. Esta fase inclui a selecção do(s) método(s) específico(s) a usar na procura de padrões[199]. Alguns exemplos destes métodos são: as redes neurais e as árvores de decisão. É aqui, decorre um processo de *Meta-Learning*[200], onde o principal objectivo é avaliar probabilidade de um algoritmo de *Data Mining*, ter sucesso ou não, aquando da sua aplicação a um determinado problema. Assim, esta abordagem tenta entender as condições em que, um algoritmo de *Data Mining* é mais apropriado, sendo que, cada algoritmo tem diferentes parâmetros e tácticas de aprendizagem[201].

7º. Empregar o algoritmo de *Data Mining*: é nesta fase que é realizada a implementação do algoritmo de *Data Mining*. Por vezes, pode ser necessário empregar o algoritmo várias vezes, até se obter um resultado satisfatório (por exemplo, entre cada aplicação, alterar os parâmetros de controlo do algoritmo [por exemplo, alterar número de instâncias de uma árvore de decisão, para que seja apresentado apenas um único ramo]).

8º. Avaliação/Interpretação: neste estádio, procede-se à avaliação e interpretação dos padrões gerados (regras, fiabilidade, etc), respeitando os objectivos definidos no 1º passo. É nesta fase, que é considera a possibilidade/necessidade de retorno, a qualquer um dos passos anteriores (exemplo: adicionar características no quarto

[199] A maioria dos métodos de DM, são baseados em conceitos de Aprendizagem Automática, Reconhecimento de Padrões, Estatística, Inteligência Artificial e modelos gráficos (Santos & Azevedo**,** 2005: 21).

[200] O principal objectivo do *Meta-Learning* é entender a interacção entre os mecanismos de aprendizagem, e os contextos concretos onde os mecanismos (de DM) serão aplicados. Assim, o *Meta-Learning* fornece um mecanismo automático para criar *Meta-Knowledge*. *Meta-Knowledge* é o conhecimento acerca do desempenho do mecanismo de DM. Ou seja, o *Meta-Knowledge* induzido pelo *Meta-Learning* fornece os meios para informar as decisões acerca das condições precisas em que um determinado algoritmo, ou sequência de algoritmos, é melhor do que outro para uma determinada tarefa (Brazdil, et. al., 2009:1207).

[201] *Efeito de aprender; aquisição, mediante a actividade de ensinar, dos conhecimentos necessários sobre determinado assunto* (Santos & Azevedo**,** 2005: 173).

passo, e repetir novamente o processo)[202]. Este passo, foca-se no grau de compreensão[203] e na utilidade do modelo gerado, onde é possível proceder-se: à visualização dos padrões extraídos, à remoção de padrões redundantes e/ou irrelevantes e ainda, à tradução dos padrões úteis, de modo a que sejam facilmente entendidos pelo utilizador. Nesta fase, o conhecimento descoberto, é armazenado (por exemplo, num *Data Warehouse*[204] [DW]) para uso posterior.

9º. Usar o conhecimento descoberto: este é o estádio final, onde o conhecimento gerado pode ter os seguintes fins: incorporação noutro sistema para futuras utilizações (num DW, conforme referido anteriormente), utilização deste para ajudar à tomada de decisão e comparação com conhecimento anteriormente gerado ou presumido. O conhecimento torna-se activo, no sentido em que, é possível fazer-se alterações ao sistema e medir os efeitos nos resultados/padrões gerados pelo processo de DCBD.

Tendo em conta as diversas fases que constituem o processo de DCBD, é possível agrupa-las, essencialmente, em três etapas fundamentais: o pré-processamento, o *Data Mining* e pós-processamento, sendo cada uma destas constituída por várias subtarefas (Santos & Azevedo**,** 2005: 19).

Importa ainda referir, que qualquer processo de DCBD pode deparar-se com alguns problemas que, em geral, dependem essencialmente, *dos objectivos do processo, da BD a utilizar e da aplicação do conhecimento descoberto*, e em particular, da *representação do conhecimento extraído, do controlo da operação de descoberta, da selecção do objectivo de Data Mining mais apropriado, e da escolha dos métodos e técnicas adequados* (Santos & Azevedo**,** 2005: 22). Fayyad (et. al., 1996: 33-34) e Santos & Azevedo (2005: 22), reforçam este facto, e enumeram os seguintes desafios inerentes a todo o processo de DCBD:

a) Volume da BD: Bases de dados de dimensões gigantescas, com milhões de registos, (tais como, tabelas que ocupam muito espaço de armazenamento e possuem grande número de registos), podem originar uma grande variedade de

[202] *Os resultados do processo de descoberta de conhecimento podem ser visualizados de diversas formas. Porém, estas devem possibilitar uma análise criteriosa para identificar a necessidade de retornar a uma qualquer das fases anteriores do processo DCBD* (Santos & Azevedo**,** 2005: 21).
[203] *A informação encontrada deve estar numa forma perceptível ao utilizador do sistema, e deve-se fazer a verificação da qualidade dessa mesma informação.* (Santos & Azevedo**,** 2005: 21).
[204] *Os data warehouses, ou armazéns de dados, são na sua essência mais básica Sistemas de Gestão de Informação e, como tal, a sua função básica é o processamento de dados em informação, que sirva de input aos mecanismos associados aos processos de tomada de decisão nas organizações* (Caldeira**,** 2008: 28).

padrões, combinações e hipóteses. Algumas possíveis soluções, passariam pela: aplicação de algoritmos que enumerem todas as regras de associação, inclusão da amostragens (*Sampling*)[205], aplicação de métodos de aproximação e massificação de processamentos paralelos.

b) Alta dimensionalidade da BD: a alta dimensionalidade deve-se à quantidade elevada de campos (variáveis ou atributos) numa BD. Este facto, provoca um aumento exponencial, quer em termos da extensão da procura, quer na probabilidade do algoritmo encontrar padrões falsos. Algumas soluções possíveis passariam pela: utilização de métodos de prioridades (de forma a, identificar variáveis irrelevantes), incorporação de técnicas redutoras de dimensionalidade e a incorporação de conhecimento prévio/anterior.

c) Dados inconsistentes: são atributos com valores inválidos[206], que pela sua natureza, poderiam revelar-se importantes para o processo. A validação cruzada, a regulamentação e a aplicação de estratégias sofisticadas de estatística, podem contribuir para a atenuação/colmatação deste problema.

d) Ruído na BD: Alguns atributos importantes podem não constar na BD, ou podem conter valores errados. Este facto pode resultar de erros do operador (erros de introdução), do sistema ou mesmo do processo de colecta de dados. Este fenómeno é intitulado como ruído. A solução seria, por exemplo, a utilização de métodos estatísticos para identificar variáveis ocultas e as suas dependências, ou então, utilização amostras muito grandes dos dados, tornando o ruído menos significativo.

e) Dados irregulares: diferentes BD podem ser utilizadas no processo (ex: MS-SQL, MS-Access, Oracle, Informix, DB2), no entanto, os dados operacionais de diferentes BD podem ter diferentes domínios, para definir uma mesma informação, provocando uma variação no nível de qualidade dos mesmos. A solução para este problema, seria uma análise efectiva de qual a melhor BD para seleccionar os dados, ou então, construir e alimentar um DW[207], o qual apresenta um ambiente estável para a integração dos dados.

---

[205] A acção de testar a qualidade de algo, através de amostras (Oxford English Dictionary, 2009). Amostragem é a *criação de um subconjunto a partir de um universo. Existem várias técnicas possíveis para efectuar a amostragem* (Santos & Azevedo**,** 2005: 173).

[206] Vide Subcapítulo 2.1.2.

[207] *O Data Wharehouse é uma metodologia que se propõe adequar* um *enorme manancial de dados às necessidades dos decisores, transformando os dados amontoados, armazenados em sistemas incompatíveis entre si, em informação fiável que pode ajudar as organizações a atingirem níveis superiores de compreensão. Um DW, é um repositório central de factos sobre múltiplos temas, desenvolvido com o objectivo de facilitar os mecanismos de pesquisa de informação.* (Caldeira**,** 2008: 28).

f) Dados constantemente alterados (mutabilidade): a rápida mudança dos dados (não estacionários), pode significar a inviabilização de padrões previamente gerados. Provocando, a um determinado ponto, a inferência de conclusões imponderadas e erradas, pois as variáveis medidas, que geraram o conhecimento, podem ter sido removidas e/ou modificadas. A solução passa pelo, incremento da utilização de métodos de actualização dos padrões (através da repetição do processo, após a integração dos novos dados) ou pela utilização exclusiva dos dados mais recentes, para a busca de padrões, podendo assim, avaliar a mudança e/ou evolução dos fenómenos.

g) Interacção com o utilizador: os ambientes interactivos entre o Homem e o computador, garantidos pelo processo DCBD, permitem dois tipos de descoberta (a desenvolver no subcapítulo seguinte): a descoberta humana assistida pela computação e a descoberta computacional assistida pelo Homem. A ênfase está na interacção Homem-Computador e não na completa automação do processo. O principal objectivo é desenvolver um apoio quer para especialistas, quer para utilizadores inexperientes. Os sistemas devem ainda ser autónomos e extrair apenas hipóteses úteis. Por outro lado, os sistemas devem apenas ser configurados, para a aplicação e para as BD de cada utilizador (aqui é reiterada a singularidade de cada processo), de acordo com as suas necessidades e o seu conhecimento pessoal. Algumas soluções passariam pelo desenvolvimento de técnicas de visualização, interpretação e análise dos padrões descobertos.

h) Conhecimento prévio: muitos métodos e ferramentas de DCBD não são interactivos, uma vez que, não é possível (de uma forma simples) a incorporação de conhecimento prévio acerca de determinado problema. A utilização do conhecimento probabilístico previamente retirado dos dados e de BD dedutivas torna-se importante para todas as etapas do processo DCBD.

i) Representação da informação: a exposição da informação descoberta deve ser clara, compreensível e facilmente acessível ao utilizador. Caso contrário, o conhecimento pode ser interpretado erroneamente. Algumas soluções possíveis seriam: a inclusão de representações gráficas, a aplicação de regras de estruturação e utilização de técnicas de visualização de dados e de conhecimento.

Assim, é possível dizer que, o processo de DCBD, tem como fase principal, e como núcleo de todo o seu processo, o *Data Mining*.