INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Área Departamental de Engenharia Electrónica Telecomunicações e Computadores

Sistema de Apoio à Decisão para a Classificação de Risco em Projectos de Tecnologias de Informação

Tiago João Mendonça Freire Ramalho

(Licenciado)

Tese de Mestrado para obtenção do grau de Mestre em Engenharia Informática

Orientadores:

Professor Luís Assunção

Professor Doutor Hélder Pita

Junho de 2011

Resumo

O trabalho que a seguir se apresenta tem como objectivo descrever a criação de um modelo que sirva de suporte a um sistema de apoio à decisão sobre o risco inerente à execução de projectos na área das Tecnologias de Informação (TI) recorrendo a técnicas de mineração de dados.

Durante o ciclo de vida de um projecto, existem inúmeros factores que contribuem para o seu sucesso ou insucesso. A responsabilidade de monitorizar, antever e mitigar esses factores recai sobre o Gestor de Projecto. A gestão de projectos é uma tarefa difícil e dispendiosa, consome muitos recursos, depende de numerosas variáveis e, muitas vezes, até da própria experiência do Gestor de Projecto. Ao ser confrontado com as previsões de duração e de esforço para a execução de uma determinada tarefa, o Gestor de Projecto, exceptuando a sua percepção e intuição pessoal, não tem um modo objectivo de medir a plausibilidade dos valores que lhe são apresentados pelo eventual executor da tarefa. As referidas previsões são fundamentais para a organização, pois sobre elas são tomadas as decisões de planeamento global estratégico corporativo, de execução, de adiamento, de cancelamento, de adjudicação, de renegociação de âmbito, de adjudicação externa, entre outros. Esta propensão para o desvio, quando detectada numa fase inicial, pode ajudar a gerir melhor o risco associado à Gestão de Projectos.

O sucesso de cada projecto terminado foi qualificado tendo em conta a ponderação de três factores: o desvio ao orçamentado, o desvio ao planeado e o desvio ao especificado. Analisando os projectos decorridos, e correlacionando alguns dos seus atributos com o seu grau de sucesso o modelo classifica, qualitativamente, um novo projecto quanto ao seu risco. Neste contexto o risco representa o grau de afastamento do projecto ao sucesso.

Recorrendo a algoritmos de mineração de dados, tais como, árvores de classificação e redes neuronais, descreve-se o desenvolvimento de um modelo que suporta um sistema de apoio à decisão baseado na classificação de novos projectos. Os modelos são o resultado de um extensivo conjunto de testes de validação onde se procuram e refinam os indicadores que melhor caracterizam os atributos de um projecto e que mais influenciam o risco. Como suporte tecnológico para o desenvolvimento e teste foi utilizada a ferramenta Weka 3.

Uma boa utilização do modelo proposto possibilitará a criação de planos de contingência mais detalhados e uma gestão mais próxima para projectos que apresentem uma maior propensão para o risco. Assim, o resultado final pretende constituir mais uma ferramenta à disposição do Gestor de Projecto

Palavras-chave: Mineração de dados, Sistemas de Suporte à Decisão, Gestão de Projectos, Análise de Risco, *Key Process Indicators* (KPI).

Abstract

The aim of this work is to create a model that sustains a decision support system which determines the inherent risk of the execution of projects in the Information Technologies context.

During the life cycle of a project, there are a number of factors that contribute to its success or failure. The responsibility of monitoring, predicting and mitigating these factors belongs to the Project Manager. The management of projects is a hard and expensive task, it consumes many resources and depends on many variables and, frequently, even on the experience of its own Project Manager. When confronted with the estimate of duration and effort for the execution of a specific task, the Project Manager doesn´t have, except for his own perception and personal intuition, an objective way to measure the plausibility of the values which are presented by the task executor. These predictions are paramount for the organization, because many strategic decisions, concerning projects, are base on them. Decisions such as: global planning, execution, postponement, cancellation, outsourcing, scope renegotiation, etc. This tendency for deviation, when detected at an initial stage, may help to better manage the risk of Project Management.

The success of each completed project was calculated taking into account the weighting of three factors: was it on time, on budget and within client specification. After analyzing all closed projects, and correlating some of its attributes with the degree of success, the model classifies, qualitatively, a new project for risk. In this context the risk is the degree of deviation from the project to success.

Using data mining algorithms, such as classification trees and neural networks, this report describes the development of a model that supports a Decision Support System for the classification of new projects for risk. The models are the result of an extensive set of tests to seek validation and refine the indicators that best characterize a project and that most influence the risk. As technological support for developing and testing the tool was used Weka 3.

A good use of the proposed model will enable the creation of more detailed contingency plans and closer management for projects with a greater propensity for risk. The final result of this project aims to be one more tool which the Project Manager can have at his disposal. A good use of the proposed model will imply the creation of more detailed contingency plans and a closer management for projects that have a larger tendency for risk.

***Keywords**: Data Mining, Decision Support Systems, Project Management, Risk Analysis, Key Process Indicators.*

Agradecimentos

Ao Hélder Pita e ao Luís Assunção quero agradecer a orientação prestada neste trabalho. As indicações, os conselhos, a sensatez e, acima de tudo, a paciência.

À Cláudia quero agradecer o seu apoio e incentivo, o seu companheirismo, a sua compreensão e o seu incondicional amor em todos os momentos.

A minha mãe o seu apoio e compreensão, ao meu pai as muitas orientações, correcções e revisões.

Ao João Zorro e à Ana Telma obrigado pelas revisões, orientação, apoio e amizade.

A todos os meus colegas tenho que agradecer pelas inúmeras explicações, compreensão e apoio.

Minima maxima sunt…

Índice

Índice de gráficos

Índice de imagens

Índice de tabelas

Índice de esquemas

1. Introdução

Neste capítulo, o leitor irá encontrar uma breve contextualização do problema, uma sucinta explicação sobre a motivação e uma definição do âmbito e dos objectivos deste trabalho.

1.1 Contexto e âmbitos de estudo

As organizações empresariais apresentam necessidades de negócio (pedidos) constantes no âmbito das Tecnologias de Informação (TI). À medida que os negócios mudam face ao mercado, torna-se indispensável um acompanhamento contínuo dos sistemas de informação que os suportam. Uma resposta rápida e eficiente a essas necessidades constitui um factor de peso no que diz respeito à competitividade. No entanto, as respostas aos pedidos solicitados devem ser feitas a partir de uma perspectiva global ao contexto empresarial**.** Nesse sentido, as necessidades são analisadas quanto aos impactos, ao custo de oportunidade, às vantagens competitivas, aos benefícios, ao retorno de investimento, ao esforço e à duração.

Segundo o *Chaos Report* [1] um projecto para ser *bem sucedido* tem de reunir a seguintes condições: "*The project is completed on-time and on-budget, with all features and functions as initially specified"*. Em 1994, o mesmo relatório estimava que 16% dos projectos TI eram bem-sucedidos. Em 2009, o seu valor era apenas de 32% [2]. Considera-se que o grau de risco para esta indústria é elevado, existindo um grande potencial nos esforços desenvolvidos para mitigar o insucesso. Embora um Gestor de Projecto seja obrigado a manter-se ao corrente das condicionantes da organização em que o projecto está inserido, a tomada de decisões continua a ser da sua responsabilidade.

Nos cenários em que a capacidade de execução de novas actividades é limitada face ao número de pedidos, estes não podem ser simplesmente atendidos por "ordem de chegada". Deverão ser agrupados por período de tempo (ciclos), por sistema/tecnologia e por importância estratégica. O resultado desta metodologia resulta numa prioridade de pedidos. Inevitavelmente alguns serão adiados, outros fundidos ou cancelados. Neste contexto, é fundamental medir a capacidade que um projecto tem de cumprir um plano, em termos de duração e esforço.

Pode imaginar-se a informação relevante sobre a execução de um projecto representada num gráfico a duas dimensões (benefício futuro *versus* custo previsto). Este trabalho propõe uma terceira dimensão: *o risco*. Esta dimensão constitui mais uma variável no apoio à decisão ilustrado no *Gráfico 1*.

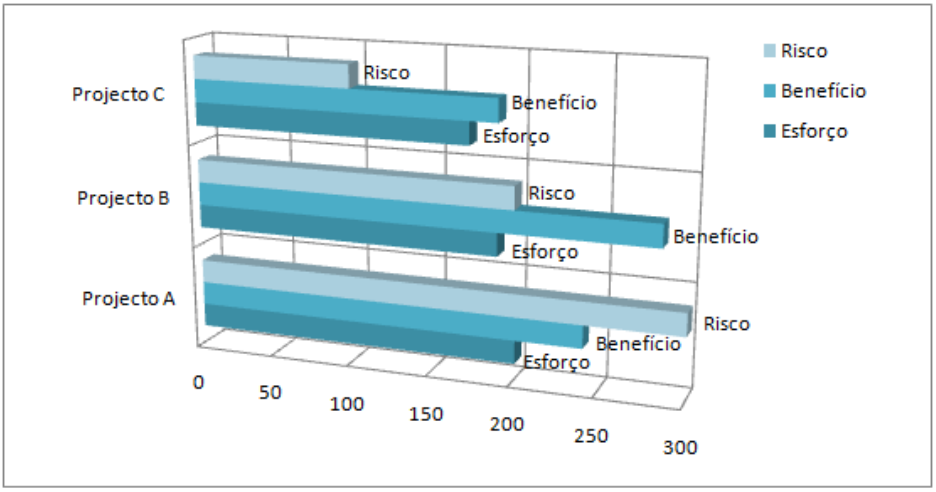


Gráfico 1 - Esforço, benefício e risco para exemplificado em três projectos (A, B e C)

Num contexto onde existem necessidades constantes e a capacidade é limitada, a combinação custo, benefício e risco permitem à gestão determinar, dentro do plano global estratégico, a importância de cada projecto, a ordem pela qual deve ser executado e o tipo de acompanhamento, ou seja, a quem deve ser entregue a sua gestão.

No âmbito deste trabalho foi necessário realizar um levantamento que permita compreender melhor a metodologia utilizada. Importa descrever como é feito o planeamento, o modo como é realizada a orçamentação e como são levantados os requisitos de negócio. Em suma, é indispensável analisar todos os procedimentos nesta fase do projecto. O contexto incidirá sobre a Direcção de Sistemas de Informação (DSI) do Grupo Caixa Seguros e Saúde ® [3].

1.2 Objectivos

A partir do contexto real de uma organização, onde existe um repositório de informação sobre a execução de múltiplos projectos de TI, tenta-se extrair padrões bem caracterizados sobre factores de risco na execução de futuros projectos, usando técnicas de mineração de dados.

Assim, pretende-se criar um mecanismo que antecipe a qualificação do risco que a eventual execução de um projecto possa apresentar. Para se atingir o objectivo foram seguidas as seguintes fases/etapas/tarefas:

i. A criação de um *Data Mart* [4], com as informações das bases de dados já existentes na DSI, cumpre parcialmente, o objectivo secundário de reunir a informação;
ii. O estudo da entropia dos atributos presentes no *Data Mart* criado, examinado o seu grau de correlação, e qual a sua contribuição para a quantificação final do risco. Se a má qualidade da informação o justificar, será necessário voltar ao passo anterior e refazer o *Data Mart*;
iii. A análise das diferentes técnicas, de modo a obter o melhor modelo possível de quantificação do risco para futuros projectos. Com este modelo pretende-se gerar uma ferramenta para que a Gestão de Projectos consiga, durante as diferentes fases do

projecto, determinar o risco que ele representa. Esta capacidade tem como objectivo melhorar a gestão de carteira de projectos e servir de alerta para projectos "problemáticos".

1.3 Organização do documento

No primeiro capítulo define-se o âmbito do trabalho, é feita a sua contextualização, uma breve introdução ao problema e a enumeração dos objectivos finais. Descreve-se ainda organização do documento.

O segundo capítulo incide sobre o domínio do problema. Nele são apresentados os conceitos necessários para a compreensão do mesmo. É neste capítulo que se faz uma introdução aos temas da Gestão de Projectos, define-se o que é o risco e quais são as técnicas usadas para a realizar sua identificação, avaliação e gestão. Explica-se o processo de obtenção de *Key Process Indicators* (KPI) [5] e o modelo de classificação de projectos. Discute-se o estado da arte.

No terceiro capítulo é apresentada a metodologia *Cross Industry Standard for Data Minning* (CRISP-DM) [6], expõe-se o processo de mineração de dados, justificam-se as opções tomadas e descrevem-se as principais técnicas e algoritmos utilizados. Este capítulo descreve, de um ponto de vista genérico, o trabalho efectuado.

O capítulo quatro dedica-se ao enquadramento do problema. Descreve o processo operacional para a realização de um novo projecto, introduzindo e explicando os conceitos necessários para se compreender a sua realização. Discute-se os diferentes tipos de organização e os impactos da mudança. Fala-se sobre o controlo de tempos e faz-se um ponto de situação quanto a eventuais soluções actuais.

No quinto é descrito todo o processo de implementação. Descreve-se o processo de selecção, recolha e transformarão de atributos em indicadores. Relatam-se os ensaios realizados na tentativa de refinar o modelo. Por fim é apresentada uma validação do modelo que melhor desempenho teve em todos os ensaios realizados.

O quinto capítulo apresenta as conclusões finais.

2. Gestão de projectos

É neste capítulo que se introduzem alguns dos conceitos importantes para a compreensão do problema do ponto de vista do negócio. O leitor irá contextualizar-se sobre alguns dos aspectos específicos da área de Gestão de Projectos, com especial ênfase na Gestão de Risco. Serão também discutidos os *Key Process Indicators* e o estado da arte.

2.1 Gestão de Risco

A Gestão de Riscos é a tarefa de identificação, de avaliação e da priorização de riscos[1] (definido na norma ISO 31000 como *"(…)the effect of uncertainty on objectives, whether positive or negative*"), seguido da aplicação coordenada e mais eficiente possível dos recursos para minimizar, monitorar e controlar a probabilidade e/ou impacto de eventos inesperados, ou para maximizar a realização de oportunidades [7] [8] [9].

Segundo o *Project Management Institute* (PMI) a Gestão de Risco num projecto [10], e cito: "(…) *includes the process of conducting risk management planning, identification analysis, response planning, monitoring and control on a project*". Esta gestão é composta por seis processos distintos, com diferentes intervenções ao longo de várias fases. A capacidade de fazer uma gestão apropriada de risco depende, entre outros, dos seguintes factores: qualidade do planeamento, capacidade de identificação de riscos, análise qualitativa e quantitativa de riscos, planeamento de resposta ao risco, monitorização e controlo de riscos. A Gestão de Risco tem as suas origens na incerteza presente em todos os projectos. Os riscos conhecidos são aqueles que foram identificados e analisados, tornando possível planear a sua resolução e prever a probabilidade de ocorrerem. Os riscos desconhecidos não podem ser geridos proactivamente, o que salienta a necessidade de criar planos de contingência [11].

Ambas as fontes, PMI e ISO, têm uma *framework*[2] muito bem definida descrevendo todo o processo operacional para a Gestão de Risco.

2.2 Identificação do risco

Assim, a identificação de riscos pode começar com a origem dos problemas, ou com o problema em si [10].

- **Análise da origem**: as fontes do risco podem ser internas ou externas ao sistema.

[1] A norma ISO 31000 (2009) /ISO *Guide* 73 define risco como, e cito: "*(…) risk is the 'effect of uncertainty on objectives'. In this definition, uncertainties include events (which may or not happen) and uncertainties caused by a lack of information or ambiguity. This definition also includes both negative and positive impacts on objectives.*" [7]

[2] *Framework: "A structure for supporting or enclosing something else, especially a skeletal support used as the basis for something being constructed; A set of assumptions, concepts, values, and practices that constitutes a way of viewing reality."* [31]

- **Análise ao problema**: os riscos estão relacionados com ameaças previamente identificadas. Detalhar o problema.
- **Objectivos em riscos**: as organizações e equipas no projecto têm objectivos a cumprir. Quando um evento põe em perigo um objectivo, em parte ou totalmente, esse evento é identificado como um risco.
- **Cenários de identificação de riscos**: na fase da análise ao risco são criadas diferentes situações hipotéticas alternativas, baseadas nos objectivos existentes. Os cenários são criados como parte do processo de Gestão de Risco, e são as vias alternativas para alcançar os objectivos. Qualquer evento que desencadeie um cenário alternativo é identificado como um risco.
- **Taxonomia baseada em identificação de riscos**: taxonomia baseada em identificação de riscos é uma discriminação das possíveis fontes de risco. Com base na taxonomia e no conhecimento das melhores práticas, é compilado um questionário. As respostas revelam riscos [12].
- **Riscos comuns**: para várias indústrias, estão disponíveis listas com os riscos conhecidos.
- **Matriz de mapeamento de risco**: este método combina as abordagens mencionadas anteriormente com os factores que podem exponenciar ou mitigar os riscos. A criação de uma matriz contendo estas rubricas, permite uma variedade de abordagens. Podem analisar-se os recursos e considerar as ameaças a que estão expostos, assim como as consequências que afectam cada um. Alternativamente, podem analisar-se as ameaças e verificar quais os recursos que afectariam, ou analisar as consequências e determinar a combinação de ameaças para os recursos afectados. O resultado final funciona como um mapa de resposta ao risco [8].

2.3 Avaliação do risco

Depois de identificados os riscos, estes devem ser avaliados quanto à sua gravidade potencial de impacto (geralmente com impacto negativo) e quanto à probabilidade de ocorrência. Estas quantificações podem ser simples ou impossíveis de determinar. O processo de avaliação é fundamental para tomar as melhores decisões de modo a adequadamente dar prioridade à implementação do plano de gestão de risco.

A dificuldade fundamental na avaliação de risco reside na determinação da taxa de ocorrência. A informação estatística só existe para incidentes passados. Além disso, qualificar a gravidade das consequências (o impacto) é bastante difícil. Assim, existem várias teorias na tentativa de quantificar o risco. Existem fórmulas diferentes, mas talvez a fórmula mais aceite para a quantificação do risco seja [13]:

$$M_e = P_e . C_e \rightarrow R_t = \sum_{i=1}^{n} M_i$$

Onde M_e representa a magnitude do risco de um evento, P_e a probabilidade de ocorrência de um evento, C_e o grau de severidade da consequência de um evento; R_t representa o risco total de um projecto.

A quantificação de risco é uma ferramenta eficaz para lidar com o imprevisto no futuro que deverá funcionar com a experiência adquirida de erros passados.

2.4 Key Process Indicators (KPI)

É possível definir *Key Process Indicators* (KPI) como [14]: "*(…)a set of values used to measure against. These raw sets of values, which are fed to systems in charge of summarizing the information, are called indicators. (…) Key Performance Indicators, in practical terms and for strategic development, are objectives to be targeted that will add the most value to the business*". Neste contexto, os KPI seriam o conjunto de indicadores utilizados para inferir, sumariar e caracterizar informação sobre a execução dos projectos.

Como fonte de KPI foram consultados: KPI Library[3], Deloitte e Accenture Consulting Services e os especialistas no negócio.

2.5 Modelos de classificação de sucesso

Existem inúmeros modelos de classificação de projectos [15] [16]. Alguns dos modelos avaliam indicadores mais subjectivos, como a "percepção de produto", e outros medem factores mais concretos, como o lucro absoluto. Para este projecto, o modelo escolhido teria de ser simples e mensurável. Foi adoptado um dos modelos mais simples e aceites, denominado "*efficiency of project execution*", ou modelo clássico. Este modelo é baseado em três indicadores do projecto, ou dimensões:

- Está dentro do planeado?
- Está dentro do orçamento?
- Está dentro da especificação?

Convém sublinhar que "dentro do planeado", "dentro do orçamentado" e "dentro da especificação" significa cumprir, o mais rigorosamente possível, o que foi determinado no princípio do projecto.

2.6 Estado da arte

2.6.1 Risk Management Information Systems

A classe de software que faz Gestão de Risco é designada na indústria por *Risk Management Information Systems* (RMIS).

[3] http://kpilibrary.com/, organização que se dedica a catalogar KPIs

Todas as ferramentas analisadas seguem as recomendações da especificação emitida pelo *National Institute of Standards and Technology* (NIST) [17]. Todas têm sistemas para a identificação de risco e para o registo de ocorrência, que podem, ou não, realimentar a matriz de mapeamento de risco. Há a tendência no sentido de existirem modelos específicos para determinadas indústrias e que têm matrizes de mapeamento de risco específicos bem conhecidos. Algumas das ferramentas oferecem sistemas de *Bussiness Inteligence* permitindo algum tipo de prospecção nos dados acumulados.

A empresa AON eSolutions® [4] oferece uma série de produtos na área do risco mas estão particularmente vocacionados para a gestão da carteira de apólices de seguros (particularmente de saúde e vida). Esta empresa especializou-se em determinadas áreas de negócio, nomeadamente: indústria da aviação, construção civil, *Fast Moving Consumer Goods* (FMCG)[5], seguros, logística, vendas a retalho e gestão de recursos naturais[6]. Estes sectores estão abrangidos por matrizes de atribuição de riscos apropriadas.

O registo de incidentes é feito no módulo *RiskRegister* ® e a matriz de mapeamento de risco é feita pelo módulo *SafetyLogic*®.

Como ponto forte de venda, a empresa OrigamiRisk®[7], afirma não se especializar em nenhuma área em particular, mas na "gestão de risco pura". Tem os módulos habituais de matriz de mapeamento de risco e registo de incidentes. Garante permitir à organização, através de alguma mineração de informação, calcular o risco total do portfólio de projectos a decorrer.

Este produto tem uma grande parte do seu sistema focado na gestão da carteira de apólices e gestão de incidentes.

Existe algum esforço de providenciar ferramentas para a importação de dados de repositórios existentes.

Ao contrário dos sistemas expostos até agora, a empresa MetricStream® Risk Management System ® [8], oferece produtos de gestão de risco especificamente para a área das TI. Este produto é uma *Framework* completa com a documentação (segundo as normas ISO) e ferramentas para o registo de incidentes e a configuração de uma matriz de atribuição de risco.

Segundo o fornecedor, o ponto forte do sistema é a "protecção" no caso de auditorias. Esta necessidade de protecção contra auditorias surge porque o pagamento de muitos dos serviços prestados na área de TI encontram-se indexados à qualidade. Uma das maneiras de medir a qualidade é medir o número de incidentes (riscos).

[4] Informação retirada do sítio corporativo: http://www.aon-esolutions.com/aonES
[5] Produtos baratos, de venda massificada e com pouco envolvimento do consumidor (e sem fidelização à marca)
[6] À data desta análise, 1º trimestre de 2011
[7] Informação retirada do sítio corporativo: http://www.origamirisk.com
[8] Informação retirada do sítio corporativo: http://www.metricstream.com

3. Mineração de Dados

Neste capítulo são expostos os diferentes processos que compõem a metodologia adoptada. Pretende-se, assim, familiarizar o leitor com os alguns termos e definições e também com a particularização de alguns aspectos da implementação.

3.1 Metodologia

Uma vez que a resolução do problema se insere na área da mineração de dados importa, em primeiro lugar, dar uma definição, citando o professor Jason Frand [18]:

"Data mining (sometimes called Data or Knowledge Discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among inumerous fields in databases."

Dentro deste contexto, é importante redefinir alguns conceitos comuns mas, que assumem significados diferentes. De acordo com Russel Ackoff, teórico de sistemas, o conhecimento humano pode ser classificado em cinco categorias [19]:

1. *Data: symbols*
2. *Information: data that are processed to be useful; provides answers to "who", "what","where", and "when" questions*
3. *Knowledge: application of data and information; answers "how" questions*
4. *Understanding: appreciation of "why"*
5. *Wisdom: evaluated understanding.*

Neste trabalho utilizar-se-ão os seguintes termos: dados, informação, conhecimento, compreensão e sabedoria para representar os conceitos enumerados em inglês.

A metodologia utilizada na resolução deste trabalho é a *CRoss Industry Standard Process for Data Mining* (CRISP-DM) [6]. O CRISP-DM foi desenhado por um consórcio de empresas[9] e investigadores sendo um projecto co-financiado pela União Europeia[10]. É o resultado de uma

[9] Lideradas por: ISL, NCR Corporation, Daimler-Benz e OHRA.
[10] Ao abrigo do programa *European Strategic Program on Research in Information Technology* (ESPRIT)

pesquisa para uma solução genérica a fim de ser aplicada aos projectos de mineração de dados. Esta metodologia tem como objectivo desenhar um processo para a definição e validação de projectos de mineração de dados aplicáveis ao domínio de qualquer indústria, independentemente da tecnologia utilizada no seu desenvolvimento.

O recurso a uma metodologia normalizada, como esta, permite ajudar o desenvolvimento de grandes projectos, ambicionando acelerar os processos necessários à sua realização, reduzir custos e aumentar a fiabilidade. Pretende também promover uma gestão mais eficiente.

A *Imagem 1* ilustra de forma esquematizada os diferentes procedimentos descritos pela metodologia.

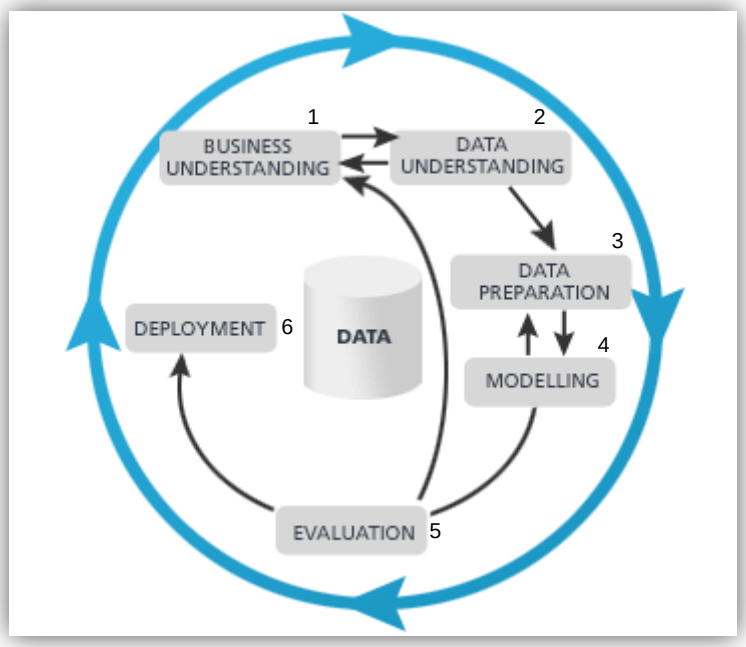


Imagem 1 - CRISP-DM

Enumerando ponto a ponto:

1. **Compreensão do problema** (*Business Understanding*) - esta fase inicial consiste numa análise ao problema em questão, cujos resultados deverão ser a definição do âmbito do problema, o levantamento de requisitos funcionais do ponto de vista do negócio, o levantamento dos procedimentos operacionais e um plano pouco detalhado sobre como alcançar o objectivo final.

2. **Compreensão dos dados** (*Data Understanding*) **-** esta fase consiste num estudo sobre os dados disponíveis, a sua aplicação no domínio do problema, a sua qualidade, a sua quantidade e as suas possíveis interdependências. Implica interceptar o conhecimento sobre o negócio e correlacioná-lo com os dados analisados. Este cruzamento, por sua vez, gera mais conhecimento sobre o negócio e sobre os próprios dados, que poderá gerar iterações subsequentes entre os pontos 1&2.

3. **Preparação dos dados** (*Data Preparation*) - esta fase cobre todas a tarefas necessárias para a obtenção de um conjunto de dados, que contenham informação com a qualidade necessária e em quantidade suficiente que permitam a geração de um modelo. Estas tarefas incluem, entre outras, as seguintes operações sobre os dados: filtrar, seleccionar, ordenar, agrupar, calcular e avaliar. Na construção desta fase devem ser tomados em consideração os objectivos pretendidos do passo seguinte, e os dados devem estar perfeitamente adaptados ao modelo.

4. **Criação do modelo** (*Modelling*) - nesta fase são seleccionadas as deferentes técnicas de mineração de dados, sendo o modelo continuamente refinado para a obtenção de melhores resultados. O aperfeiçoamento do modelo pode ser feito através da configuração e escolha dos algoritmos, ou melhorando o ponto anterior. A mudança de algoritmo implica, quase sempre, uma alteração, por vezes profunda, aos dados. Essas alterações, quando necessárias, reflectem-se através de iterações subsequentes entre os pontos 3&4

5. **Avaliação** (*Evaluation*) - nesta fase, o modelo é testado através de várias técnicas que medem a sua qualidade. Uma destas técnicas consiste na comparação do modelo com situações reais. Se os modelos não forem capazes de atingir os objectivos pretendidos teremos de voltar ao ponto 1.

6. **Disponibilização** (*Deployment*) - nesta fase do processo, o modelo é disponibilizado ao cliente em ambiente de produção[11]. O resultado final pode ser um relatório, um conjunto de regras para uma árvore de decisão ou uma aplicação complexa.

3.2 Compreensão do problema

3.2.1 Determinar os objectivos de negócio

O primeiro objectivo do analista é entender completamente o que se pretende, de facto, realizar na perspectiva do negócio. O negócio pode ter muitos objectivos, algumas vezes complexos, e restrições que devem ser devidamente equilibrados. O objectivo do analista é, numa fase inicial, descobrir factores importantes ao projecto sobre os quais recaem suspeitas de vir a influenciar o resultado final. A provável consequência de negligenciar esta etapa seria despender uma grande quantidade de esforço produzindo as respostas correctas para as perguntas erradas.

Nesta fase, é necessário recolher e agrupar as informações sobre o negócio e sobre a organização. Estes detalhes não servem apenas para ajudar a identificar os objectivos do negócio (o problema), mas também para identificar os recursos, tanto humanos como materiais, que podem ser utilizados ou que serão necessários à execução do projecto.

[11] Na indústria das TI, este é o termo que designa a entrega, para utilização, de um sistema ao seu cliente.

O guia operacional da metodologia [20] recomenda que, nesta fase, sejam ser executadas as seguintes tarefas:

3.2.1.1 Na organização

- Desenvolver organogramas que identifiquem as divisões de responsabilidade, departamentos, grupos de projecto, áreas funcionais, entre outros. Identificando os nomes dos gestores e responsabilidades;
- Identificar as pessoas-chave no negócio e os seus papéis;
- Identificar um patrocinador interno (patrocinador financeiro, utilizadores primários e especialista do domínio);
- Existe um comité de direcção? Quem são os membros?
- Identificar as unidades de negócios que são impactados pela mineração de dados, por exemplo, Marketing, Vendas, Finanças.

3.2.1.2 Quanto ao problema

- Identificar as áreas envolvidas no problema; por exemplo, Marketing, Atendimento ao Cliente, Desenvolvimento de Negócios, etc.;
- Descrever o problema em termos gerais;
- Verificar o estado actual do projecto. (Verificar se ele está claro dentro da unidade de negócios para a qual se está a realizar o estudo da mineração de dados e quais os objectivos a atingir. Questionar se é necessário publicitar?);
- Clarificar os pré-requisitos do projecto, por exemplo, qual é a motivação? O negócio já usa mineração de dados? Existe uma consciência corporativa acerca desta técnica?);
- Se necessário, preparar apresentações sobre mineração de dados para o negócio;
- Identificar grupos-alvo para o resultado do projecto, por exemplo, os utilizadores finais;
- Identificar as necessidades e expectativas dos utilizadores.

3.2.1.3 Solução actual

- Descrever qualquer solução actualmente em uso para o problema;
- Descrever as vantagens e desvantagens da solução actual;
- Descrever o nível de aceitação pelos seus utilizadores.

3.2.1.4 Objectivos

Descrever o objectivo principal do projecto, a partir de uma perspectiva de negócio. Além do objectivo de negócio principal (o problema), normalmente há um grande número de questões

relacionadas que se podem abordar. Por exemplo: o negócio conseguirá manter a actual taxa de crescimento? Continuará a reduzir custos? Que questões secundárias podem ser respondidas?

- Especificar os critérios de sucesso do negócio;
- Identificar quem avalia os critérios de sucesso;

3.2.2 Avaliar a situação

Esta tarefa consiste numa averiguação mais detalhada sobre todos os recursos, restrições, suposições e outros factores que devem ser considerados na análise de dados e no plano global do projecto. O resultado produzido será uma lista com todos os recursos disponíveis, incluindo colaboradores (de negócios e dados de especialistas, suporte técnico, pessoal de mineração de dados), dados (o acesso a dados dinâmicos armazenados ou operacionais), recursos de computação (plataformas de hardware) e software (ferramentas de mineração de dados, software relevante e outros).

3.2.2.1 Colaboradores

- Identificar patrocinador do projecto (se for diferente do patrocinador interno definido na secção 3.2.1);
- Identificar o administrador do sistema, os administradores dos diferentes SGBD[12] e outros colaboradores relevantes para o apoio técnico;
- Identificar os analistas de mercado, os especialistas em mineração de dados e verificar a sua disponibilidade;
- Verificar a disponibilidade de especialistas de domínio para as fases posteriores.

3.2.2.2 Dados

- Identificar as fontes de dados;
- Identificar as fontes de conhecimento;
- Identificar o tipo de fontes de conhecimento: fontes *online*, especialistas, documentação escrita e outras;
- Verificar as ferramentas e técnicas disponíveis;
- Descrever, formal ou informalmente, o conhecimento prévio relevante.

3.2.2.3 Computação

- Identificar o hardware base;

[12] SGBD – Sistemas de Gestão de Bases de Dados

- Estabelecer a disponibilidade do hardware base para a mineração de dados;
- Identificar o hardware disponível para as ferramentas de mineração de dados a serem usadas (se a ferramenta for conhecida nesta fase).

Será ainda necessário listar todos os requisitos do projecto, incluindo cronograma de conclusão, compreensibilidade e qualidade dos resultados, segurança e qualquer questão legal. Nesta fase devem existir garantias de que existe permissão para usar os dados.

Devem ser listados os pressupostos sobre os dados que serão verificados durante a mineração de dados. Nesta lista podem ser incluídos pressupostos não verificáveis, particularmente se o projecto lhes causar impacto ou o seu sucesso for dependente deles.

3.2.2.4 Requisitos

- Identificar o perfil do grupo-alvo;
- Identificar todos os requisitos da programação;
- Identificar os requisitos de precisão, compreensibilidade, capacidade, manutenção, repetibilidade do projecto e os seus modelos resultantes;
- Identificar requisitos em matéria de segurança, restrições legais, privacidade, documentação, *reporting* e o cronograma do projecto.

3.2.2.5 Pressupostos

- Esclarecer todos os pressupostos, incluindo os implícitos, e torná-los explícitos;
- Listar as suposições sobre a qualidade dos dados; por exemplo, a precisão e a disponibilidade;
- Listar todos os pressupostos sobre factores externos, por exemplo, questões económicas, produtos da concorrência, avanços técnicos ou outros;
- Esclarecer os pressupostos que influenciam as estimativas, por exemplo, o preço de uma ferramenta específica é assumido como sendo inferior a 1000€.

3.2.2.6 Restrições

- Listar as restrições de verificação geral, tais como, questões jurídicas, orçamento, prazos e recursos;
- Verificar os acessos às fontes de dados e perfis de segurança necessários;
- Verificar a acessibilidade técnica de dados, tais como: sistemas operativos, SGBD, formatos, ferramentas;
- Verificar se o conhecimento relevante está acessível;

- Verificar as restrições orçamentárias: custos fixos, custos de implementação, etc.

A percepção dos riscos, ou seja, os eventos que possam ocorrer e que afectem o plano, o custo ou o resultado devem ser identificados e para cada um deve ser criada uma lista de planos de contingência correspondente. É, assim, necessário verificar que acção poderá ser tomada para evitar ou minimizar o impacto, ou recuperar a partir da ocorrência do risco previsto.

3.2.2.7 Identificar os riscos

- Identificação dos riscos do negócio (por exemplo, o concorrente aparece primeiro com melhores resultados);
- Identificação dos riscos organizacionais (por exemplo, o departamento de projecto solicitante não ter financiamento para o projecto);
- Identificação dos riscos financeiros (por exemplo, o financiamento depende dos resultados iniciais da mineração de dados);
- Identificação dos riscos técnicos;
- Identificação dos riscos que dependem de dados e das suas fontes (por exemplo, má qualidade e pouca fiabilidade).

3.2.2.8 Planos de contingência

- Determinar as condições em que cada risco pode ocorrer;
- Desenvolver planos de contingência.

3.2.3 Determinar os objectivos da mineração de dados

Enquanto um objectivo de negócio estabelece os objectivos na terminologia do negócio, os objectivos da mineração são estabelecidos em termos técnicos. Por exemplo, um objectivo de negócio pode consistir em - aumentar as vendas de catálogo para os clientes existentes. Por contraste, um objectivo de mineração de dados pode consistir em - prever quantos produtos um cliente vai adquirir do catálogo no próximo ano, baseado nos dados relativos às suas compras dos últimos três anos, nas informações demográficas e no preço dos produtos:

- Traduzir as questões comerciais para as metas de mineração de dados (por exemplo, a campanha de comercialização requer segmentação de clientes, a fim de decidir a quem se dirige. O tamanho dos segmentos deve ser especificado);
- Especificar o tipo de dados no problema de mineração (por exemplo: classificação, descrição, previsão e *clustering*).

3.2.3.1 Critérios de sucesso para a mineração de dados

A definição dos critérios para um bom resultado do projecto deve ser feita em termos técnicos. Por exemplo: um certo nível de precisão da previsão. Estes critérios podem ser subjectivos; mas, se for esse o caso, devem ser identificadas as pessoas que irão fazer a apreciação subjectiva. Por isso, é preciso:

- Especificar os critérios da avaliação do modelo; (por exemplo: a precisão do modelo, o desempenho e a complexidade);
- Definir os parâmetros para os critérios de avaliação;
- Especificar os critérios que abordam critérios de avaliação subjectiva.

3.2.4 Produzir o plano de projecto

Descrever o plano destinado a alcançar os objectivos de mineração de dados e, assim, alcançar os objectivos de negócio.

Listar as etapas a ser executadas no projecto, juntamente com a duração, os recursos necessários, *inputs*, *outputs* e dependências. Tornar explícitas as iterações do processo de mineração de dados, como por exemplo, as repetições na fase de modelação e de avaliação. Analisar as dependências entre o cronograma e os riscos, indicar os resultados dessa análise explicitamente no plano do projecto, de preferência com acções e recomendações para os diferentes riscos:

- Definir o plano de projecto inicial e discutir a disponibilidade de todos os envolvidos;
- Colocar todos os objectivos identificados e técnicas seleccionadas para formar um procedimento coerente que resolva as questões colocadas pelo negócio;
- Estimar o esforço e os recursos necessários para alcançar e implementar a solução: (considerar a experiência de outros envolvidos na estimativa);
- Identificar os passos críticos;
- Identificar as iterações principais.

No final da primeira fase, deve ser realizada uma avaliação inicial de ferramentas e técnicas. Nessa fase devem ser seleccionadas ferramentas de mineração de dados que suportem os vários métodos do processo. É importante avaliar ferramentas e técnicas no início do processo, porque a selecção influencia todas as fases subsequentes.

3.3 Compreensão dos dados

3.3.1 Recolher os dados iniciais

Adquirir os dados (ou acesso a eles) listados nos recursos. Esta tarefa inclui o carregamento da colecção inicial para uma ferramenta específica (se estiver a ser utilizada).

- Planificar qual a informação necessária e quando;
- Verificar se todas as informações necessárias estão realmente disponíveis;
- Especificar os critérios de selecção. Por exemplo: saber os atributos necessários para as metas de mineração de dados? E quais os atributos que foram identificados como sendo irrelevantes e os atributos que se encontram correlacionados;
- Seleccionar as tabelas/arquivos/ficheiros de interesse;
- Seleccionar dados em uma tabela/arquivo/ficheiro;
- Determinar qual o histórico de registos a utilizar (1 ano? 2 anos?).

3.3.1.1 Inserção de dados

Alguns indicadores, que não estão presentes nos dados, terão, nesta fase, de ser obtidos.

3.3.1.2 Descrever os dados

Descrever os dados que foram “adquiridos”, incluindo: o formato, a quantidade (por exemplo, o número de registos e campos dentro de cada tabela), as identidades dos campos e qualquer outra característica dos dados que foram descobertos.

3.3.1.3 Análise volumétrica de dados

- Identificar os dados e o método de captura;
- Descrever as fontes de acesso;
- Descrever as tabelas (ou outras estruturas) e as suas relações;
- Verificar o volume de dados, multiplicidade e a complexidade;

3.3.1.4 Tipos de atributos e valores

- Verificar a acessibilidade e disponibilidade de atributos;
- Verificar os tipos de atributo (se é numérico, a taxonomia, se é simbólico, etc.);
- Verificar os intervalos de valor do atributo;
- Analisar as correlações do atributo;
- Analisar o significado de cada atributo em termos do negócio;

- Para cada atributo, calcular estatísticas básicas (por exemplo, calcular média, distribuição, máximo, mínimo, desvio padrão, variação, etc.);
- O significado do atributo é usado de forma consistente?
- O especialista de domínio deu a sua opinião sobre a relevância do atributo;
- É necessário equilibrar os dados? (dependendo da técnica de modelagem utilizada).

3.3.2 Explorar os dados

Esta tarefa aborda as questões de mineração de dados, que podem ser resolvidas usando consultas, visualização e relatórios. Estas análises podem dirigir directamente os objectivos de mineração de dados. No entanto podem também contribuir para refinar a qualidade dos dados e dos relatórios.

3.3.2.1 Exploração de Dados

- Analisar as propriedades dos atributos interessantes em detalhe.
- Analisar as características das subpopulações.

3.3.2.2 Formular suposições

- Considerar e avaliar as informações e conclusões nos dados descritos.
- Formular hipótese e identificar acções.
- Transformar hipóteses em objectivos de mineração de dados, se possível.
- Esclarecer os objectivos da mineração de dados ou torná-los mais precisos.
- A “busca cega” não é necessariamente inútil, mas uma pesquisa mais voltada para os objectivos negócios é preferível.
- Realizar análise básica para verificar a hipótese.

3.3.3 Verificar a qualidade dos dados

Examinar a qualidade dos dados, abordando questões como: os dados são completos (cobrem todos os casos); se correctos ou contêm erros; se houver erros, de que tipo são; há valores em falta nos dados; em caso afirmativo, como são representados, onde ocorrem e quão comuns são.

- Verificar os atributos identificadores;
- Verificar se os significados de atributos e valores neles contidos fazem sentido?
- Identificar os atributos ausentes e os campos nulos;
- Determinar o significado de dados em falta;

- Verificar se há atributos com valores diferentes que têm significados semelhantes;
- Verificar a ortografia de valores;
- Verificar se há desvios. Analisar se o desvio é um fenómeno relevante ou apenas ruído;
- Verificar se há plausibilidade dos valores.

3.3.3.1 Ruído e inconsistências entre as fontes.

- Verificar as consistências e redundâncias entre diferentes fontes;
- Planear como lidar com o ruído;
- Detectar o tipo de ruído e quais os atributos afectados.

3.4 Preparação dos dados

O resultado desta fase é um conjunto de dados (*dataset*), que será o objecto de trabalho da fase subsequente.

3.4.1 Selecção de dados

Decidir sobre os dados a serem utilizados para análise. Os critérios incluem relevância, qualidade, técnicas utilizadas e restrições, tais como limites de volume de dados ou tipos de dados.

- Recolher dados adicionais apropriados, a partir de diferentes fontes;
- Realizar testes de significância e correlação para decidir se os campos devem, ou não, ser incluídos;
- Reconsiderar os critérios de selecção de dados (consultar 3.3.1) à luz das experiências sobre a qualidade de dados (ou seja, poder-se incluir /excluir outros conjuntos de dados);
- Reconsiderar os critérios de selecção de dados (consultar 3.3.1) à luz das experiências sobre o modelo (ou seja, o modelo de avaliação pode mostrar que outros conjuntos de dados são necessários);
- Seleccionar subconjuntos de dados diferentes, por exemplo: diferentes atributos, apenas os dados que satisfaçam determinadas condições;
- Considerar o uso de técnicas de amostragem, por exemplo: uma solução rápida pode envolver a redução do tamanho do conjunto de dados ou a divisão em conjuntos de dados e de treino. Pode também ser útil utilizar amostras para dar importância diferente a atributos ou valores diferentes;
- Verificar as técnicas disponíveis para a amostragem de dados.

3.4.2 Limpar dados

Elevar a qualidade de dados ao nível exigido pela análise técnica. Isso pode envolver a selecção de um subconjunto de "dados limpos", a inserção de padrões adequados ou técnicas mais ambiciosas como a estimativa de dados não contabilizados na modelagem.

- Reconsiderar como lidar com o tipo de ruído observado: corrigir, remover ou ignorar o ruído;
- Como lidar com valores especiais e seu significado. Os valores especiais podem dar origem a muitos resultados estranhos e devem ser cuidadosamente examinados. Um exemplo desta codificação seria atribuir o valor '99' às perguntas de um questionário que não foram respondidas. Supondo que o resultado é de 1 a 10, este valor adulteraria as conclusões da modelação se não fosse tido em conta.

3.4.3 Construir dados

Esta tarefa inclui as operações de construção de novos dados, tais como, a produção de atributos novos, construídos a partir da combinação de atributos existentes, ou a transformação de atributos existentes.

- Verificar os mecanismos de construção disponíveis na lista de ferramentas sugeridas para o projecto.
- Decidir se é melhor executar a construção dentro da ferramenta ou fora (optar pelo que for mais eficiente, exacto e repetível).
- Reconsiderar os critérios de selecção de dados (consultar 3.3.1), com base nas experiências de construção de dados (ou seja, incluir/excluir outros conjuntos de dados).

3.4.3.1 Atributos derivados

Os atributos derivados são novos atributos construídos a partir de um ou mais atributos existentes. Um exemplo de atributo derivado é *área = largura * comprimento*. O que motiva a construção de atributos derivados?

- O conhecimento sobre o negócio indica que há algum facto importante que não está a ser representado pelos dados.
- O algoritmo de modelação processa apenas determinados tipos de dados.
- Os resultados da fase de modelação sugerem que certos factos não estão a ser representados.

3.4.3.2 Calcular atributos derivados

- Decidir se algum atributo deve ser normalizado.

- Considerar a relevância de atributos e analisar a utilização da técnica de ponderação de valores.
- Decidir como podem ser reconstruídos atributos ausentes.

3.4.3.3 Transformar atributos

- Especificar as etapas de transformação necessária
- Executar as etapas de transformação.

3.4.3.4 Gerar registos

Os registos devem acrescentar novos conhecimentos, representam novos dados que não estariam representados anteriormente, ou a ser utilizados para "alimentar" a fase da modelação.

3.4.4 Integrar dados

Descrever os métodos pelos quais a informação é combinada das várias tabelas, ou outras fontes de informação, para criar novos registos ou valores.

A fusão de dados refere-se exclusivamente à junção de dados, que possuem informações diferentes sobre os mesmos objectos de negócio. Nesta fase, também pode ser aconselhável documentar a geração de novos registos ou de valores agregados.

- Verificar se os mecanismos de integração disponíveis conseguem integrar os dados conforme necessário;
- Integrar as fontes e armazenar os resultados.

3.4.5 Formatar dados

Esta fase refere-se às modificações, principalmente sintácticas, para que os dados respeitem os requisitos das ferramentas de modelação. Estas modificações não alteram o significado original dos dados.

3.5 Modelação

3.5.1 Seleccionar uma técnica para a modelação

Com o primeiro passo selecciona-se a técnica de modelação inicial. Se tiverem sido especificadas várias técnicas, deve executar-se esta tarefa para cada uma separadamente.

Nem todas as ferramentas e técnicas são aplicáveis a cada tarefa. Para determinados problemas, apenas algumas técnicas são apropriadas (consultar *Apêndice 1 – Tipificação de problemas na mineração de dados*). Além destas ferramentas e técnicas existem ainda "requisitos políticos" e outras restrições que limitam ainda mais as escolhas da mineração. Poder-se-á ter disponível apenas uma ferramenta ou uma técnica para resolver o problema. A ferramenta pode não ser a tecnicamente melhor.

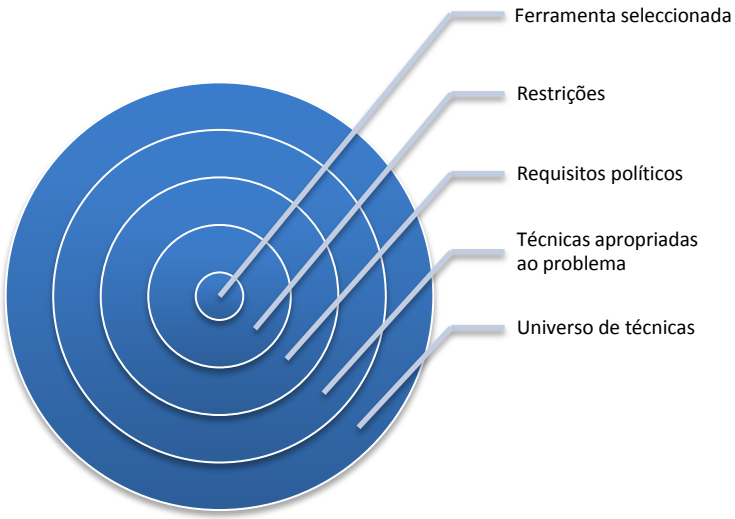


Ilustração 1 - Selecção de técnicas

A selecção da ferramenta final resulta das intercepções entre os pontos descritos na *Ilustração 1 - Selecção de técnicas*.

3.5.1.1 Árvores de decisão, o algoritmo J48

O algoritmo J48 é uma implementação do algoritmo de árvores de classificação C4.5, que, por sua vez, é uma extensão do algoritmo ID3 [21]. O J48 permite classificar uma nova instância baseando-se nos vários atributos de uma amostra populacional, denominada de conjunto de treino. Deste modo, é necessário que o conjunto de treino seja constituído pelos atributos que mais contribuem para a classificação e pela própria classificação. Pelo facto de precisar de uma classificação prévia, este algoritmo recai na classe de algoritmos de aprendizagem supervisionada [22].

O seu funcionamento utiliza a medição do ganho de informação que cada atributo representa para a classificação. Ao encontrar o atributo com o maior ganho de informação é criado um nó a partir desse atributo. O número de ramos representa os valores possíveis para cada atributo. Para cada ramo procura-se o atributo com maior ganho de informação e repete-se o processo. As folhas representam a classe onde aquele exemplo se qualifica.

Vejamos um exemplo em pseudocódigo que descreve o algoritmo. Sendo T um conjunto de exemplos (casos) de treino e {c1,c2, …, cn} o conjunto de classes:

1. Caso T só contenha casos pertencentes à classe c_i ou tenha um número mínimo (predefinido) de casos não pertencentes à classe c_i, então é retornado um nó folha etiquetado com essa classe.

2. Caso contrário, é seleccionado, segundo um critério, o atributo que melhor divide o conjunto de exemplos aqui contido, sendo retornado um nó de decisão com esse atributo

e as subárvores que resultam da aplicação do algoritmo aos subconjuntos de exemplos onde o atributo seleccionado toma um determinado valor.

Restrições funcionais do algoritmo:

- O número de atributos terá de ser fixo
- As classes que classificam os exemplos têm de ser expressas num conjunto fixo, e o seu valor bem delimitado (todos os valores têm de ser possíveis de serem definidos);
- O conjunto de dados tem de ter exemplos suficientes para caracterizar o problema.

Algumas das vantagens operacionais, que levaram este algoritmo a ser considerado como solução para este problema, são:

- Os dados podem ser categóricos ou contínuos[13]
- Possam existir atributos com valores indefinidos
- Realização da "poda" da árvore resultante, ou seja, definir à custa da introdução de algum erro, quais os ramos que permitimos que o algoritmo "corte". Esta poda faz com que alguns ramos contenham uma percentagem de exemplos mal classificados. Esta tarefa pretende impedir que a árvore fique sobre-adaptada, situação que quando ocorre faz com que o modelo passe a memorizar em vez de generalizar[14].

3.5.1.2 Redes de perceptrões multicamada (MLP)[15]

Uma rede neuronal artificial pode ser vista como uma aproximação às redes neuronais biológicas, como por exemplo, as do nosso cérebro. Como as redes neuronais biológicas, também as artificiais realizam um processamento massivamente paralelo e distribuído com propensão natural para armazenar conhecimento experimental.

Podem, assim, destacar-se as seguintes características de semelhança entre os dois tipos de redes:

- O conhecimento é adquirido através de um processo de aprendizagem
- Os pesos sinápticos presentes nos neurónios artificiais representam o acumular de potencial nas sinapses dos neurónios biológicos e têm o mesmo fim: armazenar o conhecimento aprendido.

[13] Isto é possível porque os mecanismos internos do algoritmo tornam os valores contínuos em intervalos de valores.
[14] Uma informação mais detalhada sobre a sobre-aprendizagem pode ser encontrado aqui: http://www.vcclab.org/articles/jcics-overtraining.pdf
[15] Do inglês *multilayer perceptron*

Uma rede neuronal é tipicamente composta por uma camada de entrada, por um conjunto de camadas escondida (que fazem o processamento), e por uma camada de saída.

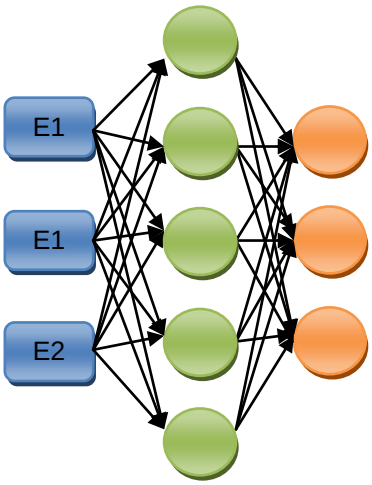


Imagem 2 - Uma rede neuronal

O processo de aprendizagem utilizado pelas redes neuronais multicamada consiste na capacidade de retropropagação (que se inspira no reforço das sinapses dos neurónios). O objectivo é permitir que o sistema aprenda com base no erro gerado à saída e que reajuste os parâmetros (sinapses ou pesos) dos neurónios para reduzir esse erro.

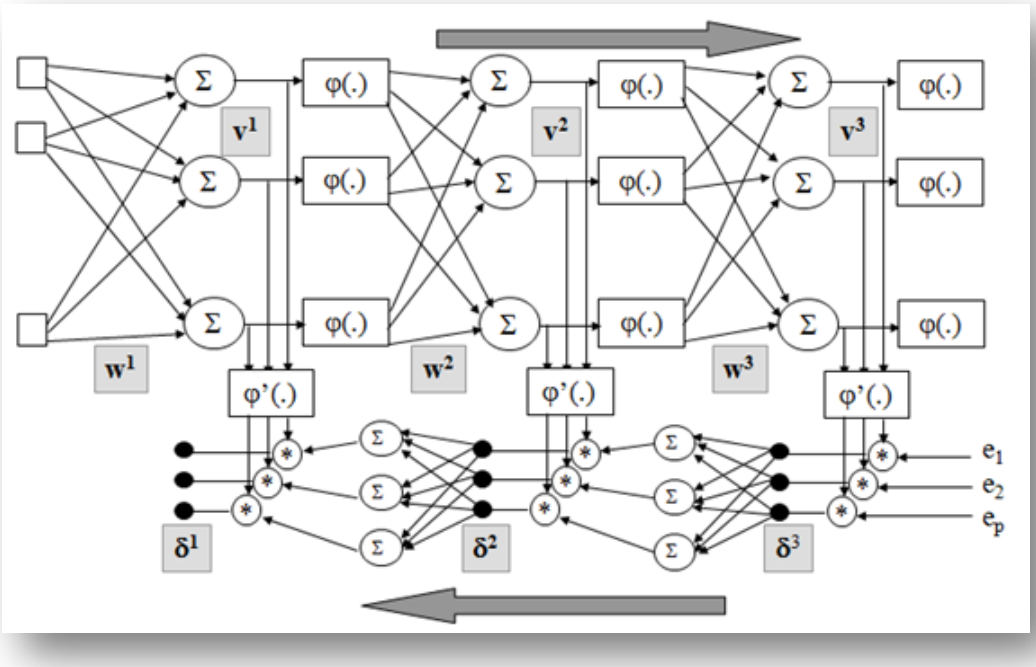


Imagem 3 - Retro-propagação de uma rede neuronal

A minimização do erro no processo de retro-propagação recorre ao método do gradiente descendente (os parâmetros δ).

$$\Delta w_{ij}(n) = \eta\delta_j(n)\gamma_i(n)$$

$$\begin{cases} \Delta_j(n) = \varphi'\left(v_j(n)\right) e_j(n)) \text{ , onde } j \text{ é um perceptrão da camada de saída} \\ \delta_j(n) = \varphi'\left(v_j(n)\right) \sum_k \delta_k(n)\, w_{jk}(n), \\ \text{onde } j \text{ é um percetrão de uma camada escondida e} \\ k \text{ é um perceptrão da camada seguinte} \end{cases}$$

Vantagens sobre as outras técnicas:

- São totalmente imparciais. Não fazem suposições sobre a natureza da distribuição dos dados.
- As capacidades para efectuar classificações não lineares tornam-nas úteis para a resolução de problemas desta natureza. A título de exemplo considere-se as séries temporais (consultar 3.5.1.3) cujos atributos, dinâmicos por natureza, necessitam de técnicas não lineares para que se possam detectar as suas relações.

Desvantagens:

- A rede neuronal pode tentar relacionar atributos sem relação real;
- O tempo de treino para grandes volumes de dados pode ser demasiado elevado;
- Depende fortemente dos coeficientes de aprendizagem e da topologia da rede.

3.5.1.3 Séries temporais

Uma série temporal é uma sequência de pontos de dados, medida normalmente em momentos sucessivos espaçadas em intervalos de tempo uniforme.

Os dados de séries temporais têm uma ordenação natural temporal. Isso faz com que a análise das séries temporais seja distinta de outros problemas comuns de análise de dados, em que não há nenhuma ordem natural das observações.

A técnica de análise a séries temporais compreende todos os métodos de análise a dados de séries temporais a fim de extrair características significativas dos dados ou fazer previsões sobre o seu comportamento no futuro baseado em eventos passados; existem vários métodos para o fazer, analisemos alguns.

Média Móvel (MM)

Este método considera como previsão para o futuro (x_t) a média das observações passadas (x_{t-n}).

$$x_t = \frac{x_{t-1} + x_{t-2} + \cdots + x_{t-n}}{n}$$

A *Tabela 1* exemplifica um caso prático que se encontra ilustrado no Gráfico 2 - Média Móvel*Gráfico 2*. No exemplo pretende-se prever o momento t_{10}

Tempo	1	2	3	4	5	6	7	8	9	10
x_t		700	750	800	800	780	750	728,5	725	733,3
x_{t-1}	700	800	900	800	700	600	600	700	800	

Tabela 1 - Previsão utilizando média móvel

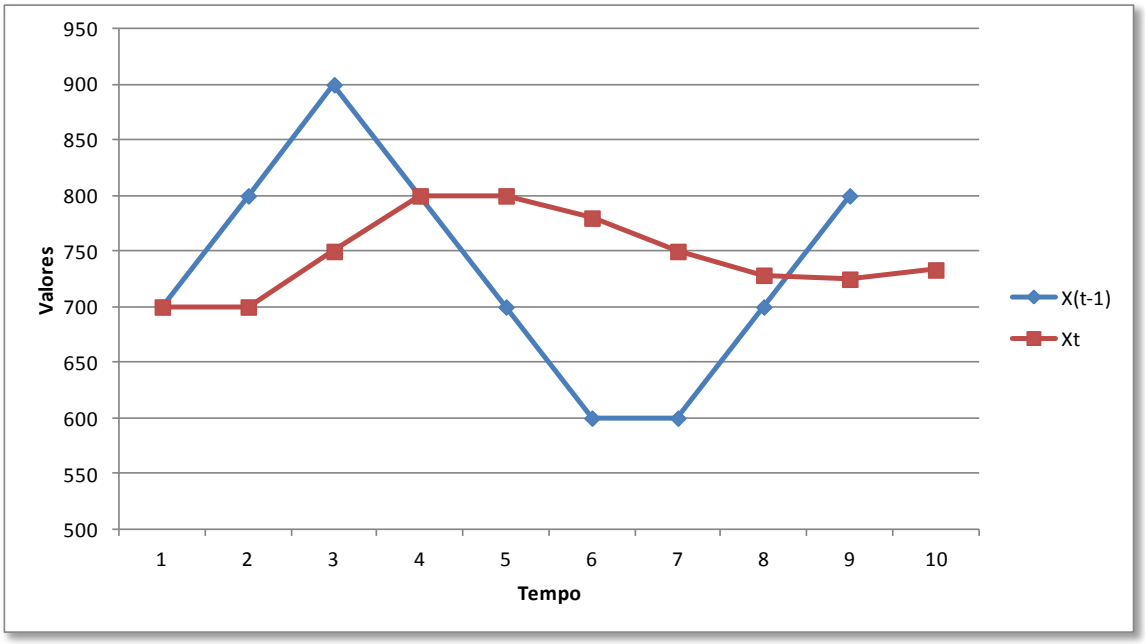


Gráfico 2 - Média Móvel

x_t = Previsão para o período t

x_{t-1} = Valor real para o período $t-1$

Alisamento Exponencial Simples (AES)

Semelhante à média móvel com a diferença que este método atribui um peso a cada observação. O tratamento diferenciado das observações da série temporal baseia-se na suposição de que as últimas observações contêm mais informações sobre o futuro e, portanto, são mais relevantes para a previsão.

$$F_t = \alpha A_{t-1} + (1-\alpha)F_{t-1}$$

F_t = Previsão para o período t

A_{t-1} = Valor real para o período $t-1$

F_{t-1} = Previsão para o período $t-1$

α = Constante de alisamento onde $0 < \alpha < 1$

A utilização desta técnica é apropriada apenas quando os dados não apresentam tendência ou sazonalidade e têm nível de ruído[16] desprezável.

Não existe uma metodologia definida para encontrar o valor de α, sendo encontrado por experimentação. Está disponível um pequeno exemplo na *Tabela 2* e ilustrado no *Gráfico 3.*

Tempo	1,0	2,0	3,0	4,0	5,0	6,0	7,0	8,0	9,0	10,0
Ft		760,0	844,0	817,6	747,0	658,8	623,5	669,4	747,8	839,1
Alpha	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,6	
A(t-1)	700,0	800,0	900,0	800,0	700,0	600,0	600,0	700,0	800,0	
F(t-1)	700,0	700,0	760,0	844,0	817,6	747,0	658,8	623,5	669,4	

Tabela 2 - Alisamento Exponencial Simples

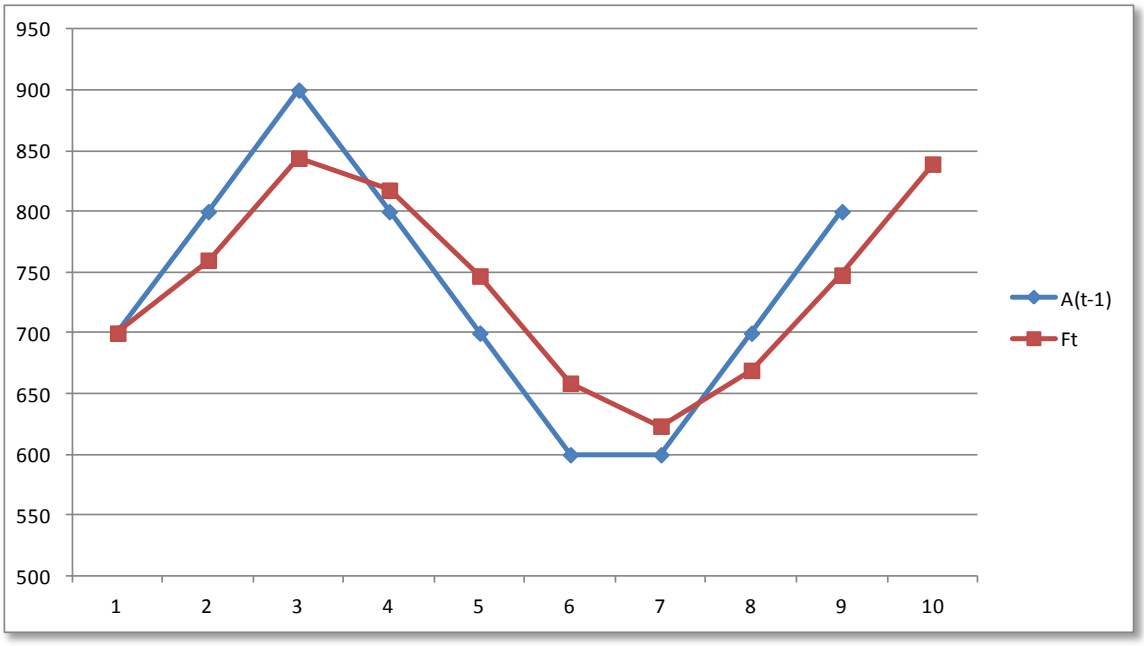


Gráfico 3 - Alisamento Exponencial Simples

Alisamento Exponencial Linear (AEL)

Quando o método Alisamento Exponencial Simples é aplicado na previsão de séries temporais que apresentam tendência entre as observações passadas, os valores previstos têm tendência para subestimar ou sobrestimar as previsões.

Para evitar esse erro sistemático foi desenvolvido o método Alisamento Exponencial Linear procurando reconhecer a presença de tendência na série de dados

$$F_{t+m} = S_t + T_t$$

[16] O ruído é a presença de valores aleatórios, sem informação, nos dados

$$S_t = \alpha x_t + (1-\alpha)(S_{t-1} + T_{t-1})$$

$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1}$$

O valor β é análogo ao α. Sendo uma constante de alisamento.

F_{t+m} = Previsão para o período t

α = Constante de alisamento onde $0 < \alpha < 1$

β = Constante de alisamento onde $0 < \alpha < 1$

Está disponível um pequeno exemplo na *Tabela 3* e ilustrado pelo *Gráfico 4.*

Tempo	1,0	2,0	3,0	4,0	5,0	6,0	7,0	8,0	9,0	10,0
St		1,6	2,7	5,8	12,1	33,2	57,4	101,1	193,5	357,4
Alpha	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,6
A(t-1)	1,0	2,0	3,4	7,8	16,4	47,2	73,6	130,1	255,0	
F(t-1)	1,0	1,0	1,6	2,7	5,8	12,1	33,2	57,4	101,1	
Tt		0,4	0,8	2,1	4,7	14,5	20,3	34,3	69,2	126,0
Beta	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,6
F(t+1)	1,0	2,0	3,5	7,9	16,8	47,7	77,8	135,4	262,6	483,4

Tabela 3 - Alisamento Exponencial Linear

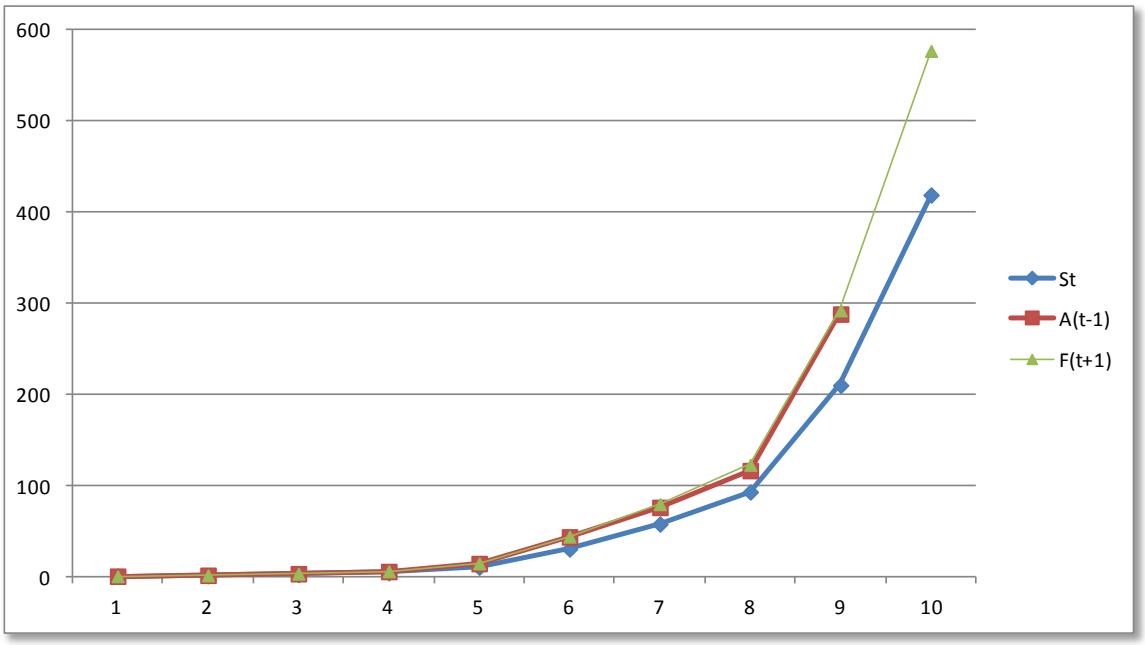


Gráfico 4 - Alisamento Exponencial Linear

3.5.2 Definir o procedimento de testes

Antes de se construir um modelo é necessário definir um procedimento para testar a qualidade e a validade do modelo. Em tarefas de mineração de dados supervisionados, tais como classificação, é comum o uso de taxas de erro como medidas de qualidade (percentagem de dados bem/mal classificados). A fase de teste específica que o conjunto de dados deve ser separado em conjunto de treino e conjunto de testes. O modelo é construído sobre o conjunto de treino e sua qualidade estimada com o conjunto de testes.

- Decidir sobre medidas necessárias (número de iterações, erro máximo, etc.);
- Preparar os dados necessários para o teste.

3.5.3 Construir modelo

Executar a ferramenta de modelação no conjunto de dados preparados para criar um ou mais modelos.

- Definir os parâmetros iniciais;
- Documentar as razões para a escolha desses valores;
- Executar a técnica seleccionada no conjunto de dados de entrada para produzir o modelo.

3.5.3.1 Descrever o modelo

Descrever o modelo resultante e avaliar a sua precisão, robustez e possíveis insuficiências. Documentar a interpretação dos resultados e das dificuldades encontradas.

- Descrever todas as características do modelo;
- Registar os parâmetros usados para produzir o modelo;
- Para modelos baseados em regras, é necessário, listar as regras e qualquer avaliação de precisão ou da cobertura;
- Para os modelos opacos, é preciso listar todas as informações sobre o modelo (tais como a topologia da rede neuronal) e todas as descrições comportamentais utilizadas pelo processo (tais como precisão ou sensibilidade);
- Descrever o comportamento do modelo e interpretação a respeito dos padrões nos dados (se houver).

3.5.4 Avaliar o modelo

O modelo deve ser avaliado para garantir que ele corresponde aos critérios de sucesso de mineração de dados. Esta é uma avaliação puramente técnica com, base no resultado das tarefas de modelação.

- Avaliar o resultado do teste de acordo com a estratégia (treinar e testar com validação cruzada, *bootstrapping*[17], etc.);

[17] Para aumentar o conjunto de dados disponíveis, inserem-se, aleatoriamente, registos duplicados [34].

- Comparar os resultados da avaliação e interpretação;
- Criar uma ordenação de resultados no que diz respeito ao sucesso versus critérios de avaliação;
- Seleccionar os melhores modelos;
- Interpretar os resultados em termos de negócios (tanto quanto possível, nesta fase);
- Obter comentários sobre os modelos dos especialistas;
- Verificar a plausibilidade do modelo;
- Verificar se o resultado é fidedigno;
- Analisar os potenciais de implementação de cada resultado;
- Se tiver sido gerada uma descrição compreensível do modelo (por exemplo, via regras), avaliar se são lógicas, se são viáveis e se fazem sentido;
- Ajustar os parâmetros até se obter um melhor modelo.

3.5.4.1 Validação cruzada (*Cross-validation*)

A validação cruzada consiste na divisão dos dados em N conjuntos. A partir dessa divisão treina-se o modelo com os dados de N-1 conjuntos. O modelo é testado no conjunto que ficou de fora. Mantendo a mesma divisão, repete-se a operação N vezes até todos os dados terem sido cobertos pelo treino (N-1 vezes) e pelos testes (1 vez).

3.5.4.2 Matriz de confusão

Uma matriz de confusão é um resultado possível da classificação de um modelo. Depois de treinado, o modelo classifica o conjunto de testes de modo a gerar uma matriz com o formato ilustrado na *Tabela 4* [23].

		Classe prevista	
		Sim	Não
Classe real	Sim	Casos bem classificados (TP)	Falsos negativos (FN)
	Não	Falsos positivos (FP)	Casos bem classificados (TN)

Tabela 4 - Matriz de confusão

No exemplo exposto o modelo classifica como “sim” ou “não”. Cada célula contém a contagem dos casos correspondentes. A partir desta matriz podem medir-se vários parâmetros relacionados com a qualidade da matriz.

Precisão ou a percentagem de exemplos correctamente classificados (*Accuracy*),

$$AC = \frac{TN + TP}{TP + FN + FP + TN}$$

Rácio de exemplos positivos correctamnte classificados (*True Positive Rate,* ou *Recall*),

$$TPR = \frac{TP}{TP + FN}$$

Rácio de exemplos negativos incorrectamente classificados como positivos (*False Positive Rate*),

$$FPR = \frac{FP}{TN + FP}$$

Rácio de exemplos positivos que foram incorrectamente classificados como positivos (*True Negative Rate),*

$$TNR = \frac{TN}{TN + FP}$$

Rácio de exemplos previstos como positivos que foi correctamente classificados (*Precision*),

$$P = \frac{TP}{FP + TP}$$

F-Measure(FM) representa a media harmónica entre o indicador *Recall* e a Precisão. Ou seja $FM = \frac{2P.TPR}{P+TPR}$. Este indicador serve como uma medida da precisão do teste, o melhor valor possível é 1 e o pior é 0.

3.6 Avaliação

Até agora lidou-se com factores técnicos como a precisão, a generalidade do modelo e a sua qualidade. Esta etapa avalia se o modelo atende aos objectivos de negócios e procura determinar se há alguma razão para o considerar improfícuo. São comparados os resultados com os critérios de avaliação definidos no início do projecto.

3.6.1 Avaliar os resultados

A avaliação de resultados deve cobrir todos os dados gerados no processo de mineração. São avaliados os resultados que estão relacionados com os objectivos do negócio original e todos os outros que não estão mas que foram surgindo como subprodutos da metodologia.

- Compreender o resultado de mineração de dados;
- Interpretar os resultados em termos de aplicação;
- Verificar o resultado da mineração de dados contra a base de conhecimento inicial e verificar se as novas descobertas sobre as informações são úteis;
- Avaliar o resultado em relação aos critérios de sucesso empresarial, ou seja, o projecto atingiu os objectivos do negócio original?

- Comparar os resultados da avaliação e interpretação;
- Há objectivos de negócio novos para serem endereçados mais tarde no projecto ou em novos projectos?
- Foram identificadas conclusões para futuros projectos de mineração de dados.

Após a avaliação do modelo, no que diz respeito aos critérios de sucesso do negócio, obter-se-á, eventualmente, a aprovação dos modelos que os satisfizerem.

3.6.2 Processo de revisão

Neste ponto, o modelo resultante parece satisfazer as necessidades do negócio. Convém, agora, fazer uma análise mais aprofundada do trabalho de mineração de dados, a fim de se determinar se existe algum factor importante, ou tarefa, que foram de alguma forma negligenciados.

- Documentar a visão geral do processo de mineração de dados utilizado;
- Analisar dados do processo de mineração;
- Para cada etapa do processo:
    - Em retrospectiva, o que era necessário?
    - A execução foi ideal?
- De que forma poderia ser melhorado?
- Identificar falhas, quando existem;
- Identificar acções alternativas possíveis, caminhos inesperados no processo;
- Revisão dados dos resultados de mineração no que diz respeito aos critérios de sucesso do negócio.

3.6.3 Próximos passos

De acordo com os resultados da avaliação e revisão do processo, será determinado o próximo passo do projecto. É necessário decidir a passagem para a implementação, se são necessárias mais iterações, ou ainda, se é necessário partir para um novo projecto de mineração de dados.

- Analisar o potencial para a implementação de cada resultado;
- Estimar o potencial para a melhoria do processo actual;
- Verificar se ainda existe disponibilidade adicional para executar mais iterações;
- Recomendar continuações alternativas;

- Refinar o plano do processo;
- Dar prioridade às possíveis acções:
    - Seleccionar uma das acções possíveis;
- Documentar as razões da escolha;

3.7 Disponibilização

Esta tarefa toma em consideração os resultados da avaliação e conclui uma estratégia para a implementação do resultado da fase anterior.

- Resumir os resultados;
- Desenvolver e avaliar planos alternativos para implementação;
- Como é que o conhecimento, ou a informação, será propagado para o seu utilizador final?
- Como será o resultado monitorizado ou os seus benefícios medidos (quando aplicável)?
- Decidir para cada modelo o resultado em termos de software;
- Como será o resultado do modelo ou software implementado dentro da organização?
- Como será a sua utilização monitorizada e seus benefícios medidos (onde aplicável)?
- Identificar possíveis problemas durante a implementação dos resultados da mineração de dados (armadilhas de implementação).

3.7.1 Planear a manutenção e monitorização

A monitorização e manutenção serão questões importantes, se o resultado de mineração de dados se tornar parte das operações quotidianas.

A preparação cuidadosa de uma estratégia de manutenção ajuda a evitar períodos desnecessariamente longos sobre os resultados da mineração de dados.

A fim de monitorar a implementação do resultado da mineração de dados, o projecto precisa de um plano detalhado sobre o processo. Este plano leva em conta o tipo específico de implantação.

Resumir a estratégia, incluindo as medidas necessárias a e como realizá-las.

- Verifique se há aspectos dinâmicos (que as coisas podem mudar no ambiente?).
- Como será a precisão monitorizada?

- Quando é que o resultado de mineração de dados não deve ser mais usado? Identificar critérios (limiar de validade, de precisão, novos dados, mudança no domínio do problema, etc.). O que deve acontecer se o modelo não puder mais ser usado? (Actualização modelo, criar projecto de mineração de dados, etc.)

- Será que os objectivos de negócio exigem uma mudança do modelo ao longo do tempo?

- Desenvolver plano de monitorização e manutenção.

3.7.2 Relatório final

No final do projecto será elaborado (pelo menos) um relatório final, onde todos os tópicos estarão reunidos. Para além de identificar os resultados obtidos, o relatório deverá ainda descrever o processo; mostrando onde foram despendidos os esforços. Deverá definir quaisquer desvios ao plano original, descrever os planos de implementação e fazer as recomendações para trabalhos futuros. O conteúdo real do detalhe do relatório depende muito do público-alvo. Deverá identificar quais relatórios que são necessários produzir: apresentação de slides, resumo de gestão, conclusões detalhadas, a explicação de modelos, entre outros.

- Analisar quantos objectivos de mineração iniciais foram cumpridos.

- Identificar os grupos-alvo para o relatório.

- Esboçar estrutura e conteúdo do relatório.

- Seccionar as descobertas a serem incluídos nos relatórios.

- Escrever um relatório

3.7.3 Rever o projecto

Resumir as experiências mais importantes durante o decorrer do projecto. Por exemplo: as armadilhas, as abordagens enganosas ou os critérios para a selecção das melhores técnicas de mineração de dados. No projecto de ideal a documentação abrange a experiência global. Incluí o conhecimento de todos os relatórios individuais que foram escritos pelos participantes do projecto em todas as fases.

- Entrevistar todas as pessoas importantes envolvidas no projecto e questionar sobre suas experiências

- Os utilizadores finais estão satisfeitos com o resultado de mineração de dados? O que poderia ter sido feito melhor? Será que eles precisam de apoio adicional?

- Resumir comentários e descrever a experiência

- Analisar o processo (as coisas que funcionaram bem, erros cometidos, lições aprendidas, etc.)
- Documentar o processo de mineração de dados específicos. Como podem os resultados e a experiência serem realimentados de volta no processo?
- Resumo de detalhes para tornar a experiência útil para projectos futuros.

4. Enquadramento do caso de estudo

Neste capítulo são expostos alguns pormenores específicos do caso de estudo. É importante contextualizar o leitor porque os conceitos apresentados serão necessários à compreensão e justificação de algumas decisões na fase de implementação.

4.1 Direcção de Sistemas de Informação

Ao analisar este trabalho, temos de ter em conta o facto deste se inserir num contexto muito específico. Os dados em análise derivam de uma empresa que opera no sector segurador. Estão, portanto, intrinsecamente ligados às tecnologias e metodologias da organização estudada.

A Direcção de Sistemas de Informação (DSI) é um departamento da empresa Caixa Seguros e Saúde, SGPS® [3] (CxS), que por sua vez é uma *holding*[18] da Caixa Geral de Depósitos® (CGD)[19]. A DSI tem como missão prestar serviços na área das Tecnologias de Informação a todas os outras organizações do universo CxS.

Existe um paralelismo entre a DSI e uma *software house*[20] convencional. As excepções são o domínio do negócio e os clientes. A DSI funciona num regime de exclusividade, prestando serviços unicamente às empresas do universo corporativo CGD (ver *Esquema 1*).

[18] *Holding:* sociedade de investimento de capitais que tem, teoricamente, por objectivo a gestão de uma carteira de valores mobiliários industriais ou comerciais [30].
[19] Esta afirmação é verdadeira na data da fase de análise do relatório: 2º Trimestre de 2011.
[20] *Software house:* é uma organização que se dedica ao desenvolvimento e comercialização de software [32].

Esquema 1 - Organigrama da Caixa Seguros, SGPS, SA – 2009

4.1.1. Interacção no contexto da organização

A DSI pode participar em projectos promovidos pela CGD (hierarquicamente superior). Nesses casos, do ponto de vista da DSI, no projecto constará a equipa do topo da raiz hierárquica, "Caixa Seguros e Saúde, SGPS, SA", que caracteriza todas as entidades CGD no contexto CxS. Esta organização significa que, na perspectiva da DSI, quando há participação de entidades externas, os seus detalhes, número de equipas, recursos e outras informações, não estão disponíveis. Esta é uma informação importante a ter em conta quando a informação estiver a ser processada.

As equipas subcontratadas (*outsourcing*) são sempre dependentes de equipas DSI. Existem, todavia, duas formas de colaboração: integração de equipas externas em equipas internas, ou solução "chave na mão". Ao adquirir serviços do tipo "chave na mão", a equipa principal detalha um custo adicional ao projecto mas não são incluídas a análise e a planificação; apenas se conhece o custo. A integração de recursos implica anexar colaboradores à hierarquia da equipa e a respectiva planificação. Este processo significa que a equipa tem, momentaneamente, mais recursos (disponibilidades); no entanto, estes serviços são prestados num contexto de um projecto e não para o trabalho regular da equipa, facto que tem de ser levado em conta quando for analisada a informação.

4.1.2. Stakeholders

As descrições das responsabilidades dos *stakeholders*, a seguir enumeradas, estão limitadas ao contexto do processo, descrito no ponto: *4.1.3 Procedimento operacional para novos projectos*.

- As **Áreas de Negócio** representam a estrutura ilustrada no *Esquema 1*. A DSI também é considerada uma área de negócio.

- O **Cliente** é um colaborador que faz o elo de ligação entre as diferentes áreas de negócio e a DSI; o cliente representa os interesses das áreas de negócio.
- O **Gestor de Relação (GR)** é um colaborador que estabelece a ligação entre a DSI e as áreas de negócio; o GR representa os interesses da DSI.
- O **Desenvolvimento** representa uma equipa (ou o seu responsável).
- O **Comité de Sistemas** representa um conjunto de pessoas com a responsabilidade de determinar o destino de um projecto. Este Comité pode atribuir prioridade, adiar, unir, cancelar, ou renegociar projectos. A reunião é feita de semestralmente e denomina-se **Fórum de Projectos**
- O ***Portfólio Management Office*** **(PMO)** representa uma equipa com a responsabilidade de monitorar o portfólio de projectos. Uma das suas responsabilidades é promover a sincronização do desenvolvimento com o plano e monitorar o seu desempenho.
- O **Gestor de Projectos (GP)** representa um recurso com a responsabilidade de gerir um projecto.
- Um projecto **intercalar** é um projecto cuja avaliação de execução é feita entre o ciclo do fórum de projectos. São projectos urgentes, consistindo em imposições legais ou grandes oportunidades de negócio. Por definição, quando aprovados, estes projectos têm precedência sobre os projectos não intercalares.
- Uma **Ficha de Impacto Global** descreve todos os impactos de uma actividade sobre as outras actividades planeadas para aquele ciclo.

4.1.3. Documentação processual

Além dos *stakeholders* no processo é útil mencionar alguns dos documentos que são produzidos pela DSI que são relevantes ao enquadramento do problema.

- O **Estudo de Viabilidade (EV)** é um documento produzido pelas equipas e pelo Gestor de Projecto contendo a informação relevante necessária para que o cliente possa decidir se é vantajoso prosseguir com o projecto. Este documento inclui um parecer técnico acerca do impacto tecnológico, uma possível solução genérica, a indicação das equipas que participam, os requisitos de negócio e, talvez o ponto mais importante de todos, a estimativa do esforço necessário. Este documento não representa um vínculo com o cliente, apenas um orçamento.

- O **Dossier de Projecto (DP)** é o contracto final com o cliente. Nele estão descritos todos os requisitos de negócio assim como uma análise técnica, a descrição da implementação, a enumeração dos executantes, estão também explicitados os valores do esforço e as datas relativas à execução.

- O **Relatório de Alteração e Progresso (RAP)** é o documento que deve reflectir qualquer desvio ao DP, seja por uma alteração de âmbito do cliente ou seja por incumprimento da equipa.

4.1.4. Procedimento operacional para novos projectos

A *Ilustração 2* - Fluxo de um novo projecto esquematiza o levantamento feito ao procedimento operacional para a criação de uma nova actividade. Trata-se de uma generalização do procedimento e não uma descrição completa, que pode ser descrita pelos seguintes passos:

1. O processo começa com um pedido do negócio. Este reflecte uma necessidade de uma ou mais áreas de negócio.
2. O negócio através do seu representante[21] elabora um pedido de actividade em cooperação com um GR. Neste documento, são definidos o âmbito, os benefícios e os custos de modo a determinar se o pedido cumpre as especificações e se é considerado um novo projecto. Para ser considerado como um novo projecto, tem de ter uma determinada dimensão ou importância estratégica para a empresa.

2.1 Caso o pedido não seja classificado como projecto este processo termina, dando lugar a um outro processo, que não é relevante para este trabalho.

3. O GR faz uma análise ao pedido de modo a analisar os requisitos de negócio. O desenvolvimento estima o esforço necessário para a realização de cada requisito. O responsável da equipa de Gestores de Projecto determina se este projecto vai ter, ou não, intervenção de um membro da sua equipa. O resultado de toda a informação produzida até ao momento é compilado num documento denominado “Estudo de Viabilidade”. A partir daí, o pedido fica “em carteira”.
4. O Comité de Sistemas reúne-se ciclicamente para poder avaliar todos os pedidos em carteira. Será atribuída prioridade consoante a sua importância estratégica. Nesta fase, alguns pedidos podem ser adiados para o próximo ciclo, anulados, renegociados (com alteração do âmbito), suspensos, agrupados ou divididos. O número de total de novos projectos é determinado pela capacidade disponível da DSI para o próximo ciclo.

4.1 O Comité de Sistemas pode reunir-se fora do ciclo para decidir sobre a aprovação de projectos intercalares.

4.2 Se o projecto intercalar for aprovado deve ser preenchida uma Ficha de Impacto Global contendo a avaliação de impactos nos projectos planeados.

5. O Desenvolvimento, juntamente com o GR e com o GP (caso intervenha), cria um plano mais detalhado da execução da actividade. Este documento, chamado “Dossier de Projecto (DP)”, contém os detalhes do esforço, da duração, do planeamento, dos entregáveis, dos custos adicionais, etc. Quando aprovado, reflecte o trabalho contratualizado com o cliente.
6. O Cliente revê o DP e, se estiver de acordo, aprova o documento formalizando o contracto[22]. O cliente pode renegociar o DP caso não concorde com algum dos pontos mencionados.

[21] Na literatura, o responsável pelo pedido denomina-se por “*Project Sponsor*”, ou patrocinador do projecto

[22] É vulgar denominar esta acção por *sign off*

7. Baseado no planeamento global, o PMO prepara as actividades para serem iniciadas. Verifica a disponibilidade do desenvolvimento e promove a sua passagem para o estado “actividade em execução”.
8. Durante a execução de um projecto, são continuamente avaliados os desvios da especificação (desvios à aceitação). Se forem detectados, as equipas criam um Relatório de Alterações e Progresso (RAP).

8.1 Os desvios à aceitação são avaliados quanto à origem (DSI ou no cliente) e quanto aos custos. O cliente e a DSI devem chegar a um acordo, ou para aumentar o custo total do projecto, ou para reduzir o âmbito.

9. O projecto é concluído.

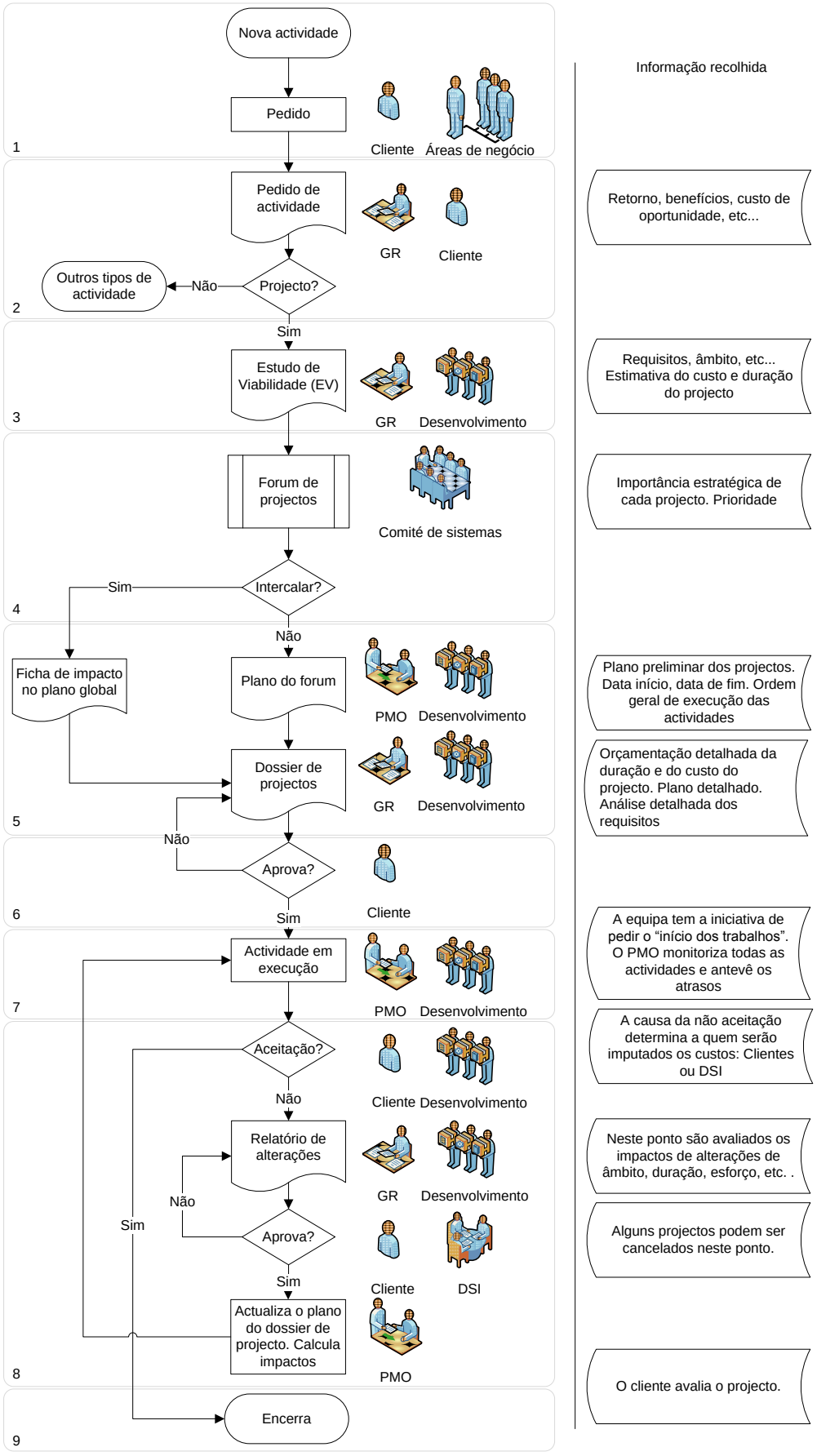


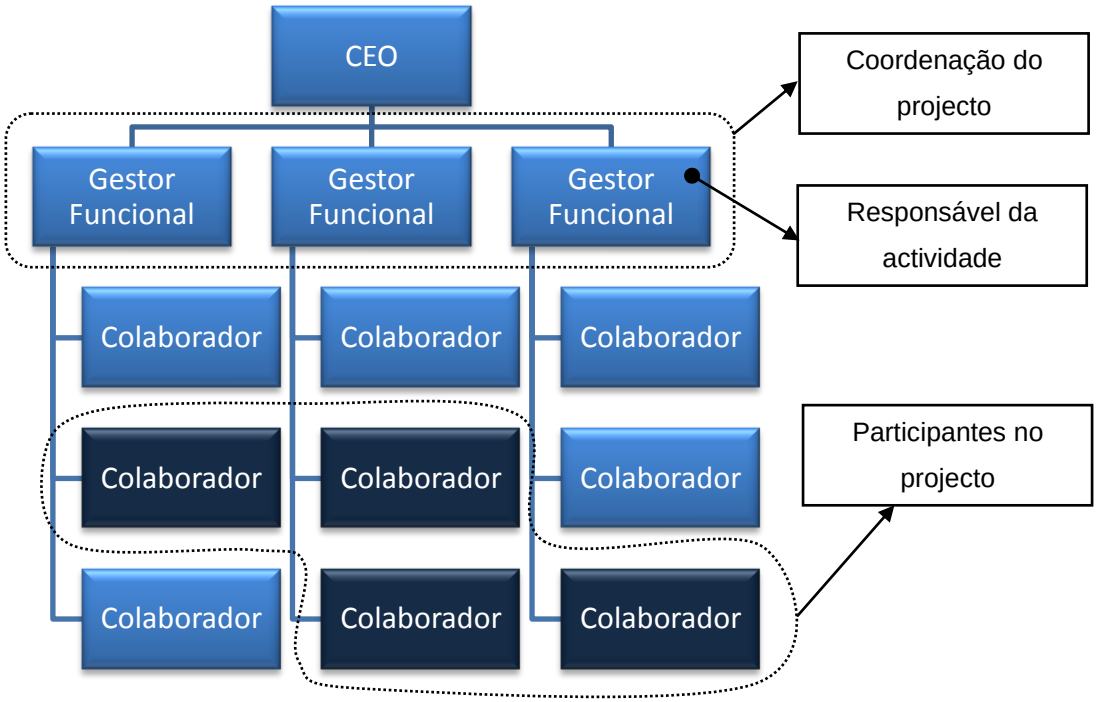
Ilustração 2 - Fluxo de um novo projecto

4.2 Tipo de organização

Sobre que entidade recai a responsabilidade da gestão de um projecto dentro da DSI? A resposta a esta pergunta pode ser obtida através da análise à estrutura da organização.

Segundo o PMBOK [24], as empresas podem estar organizadas nos seguintes modelos estruturais: funcional, matricial e orientado ao projecto[23].

As organizações funcionais caracterizam-se pela existência de uma hierarquia clara para cada colaborador. Os colaboradores são agrupados por especialização como, por exemplo, produção, marketing, área comercial, etc. Cada uma destas áreas funcionais pode ainda ser subdividida em outras áreas de maior especialização. No contexto do projecto, cada área executa o seu trabalho independentemente das outras. A sua coordenação encontra-se dividida por cada responsável da área (Gestor Funcional).

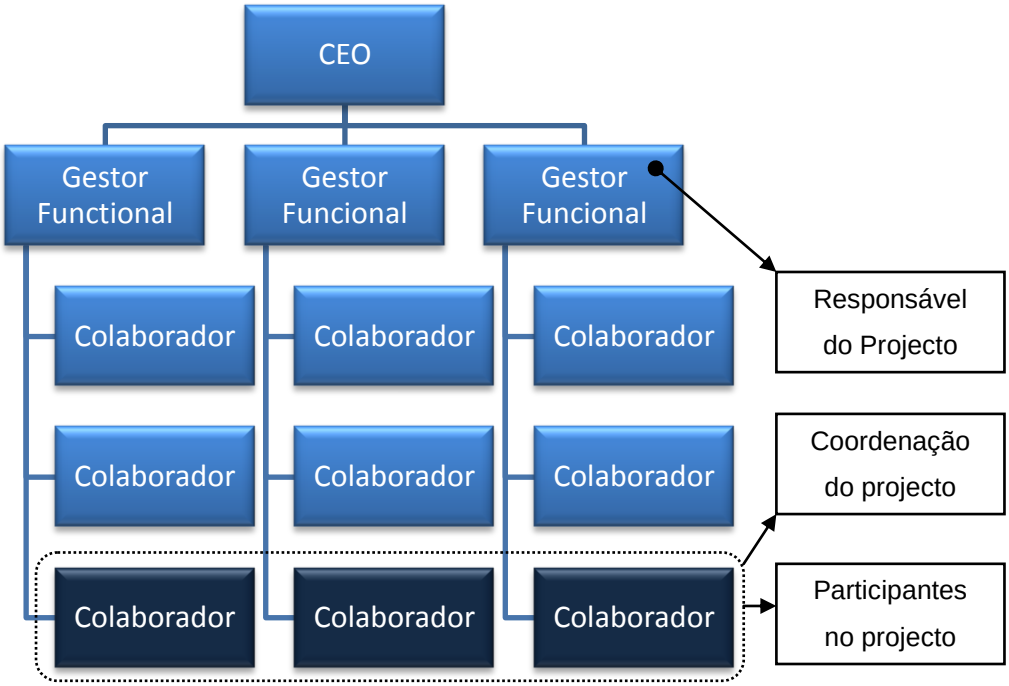


Esquema 2 - Organização funcional

No caso particular da DSI, embora a coordenação do projecto se encontre dividida entre todas as áreas funcionais, é seleccionado um responsável de área que será o responsável da actividade. Este tem como função promover a coordenação do projecto com os restantes responsáveis das áreas funcionais. O Gestor de Projecto, papel desempenhado pelo responsável da actividade, tem pouca ou nenhuma autoridade sobre o projecto [25].

[23] A expressão "organização orientada ao projecto" foi adaptada da expressão em inglês "*projectized organization*".

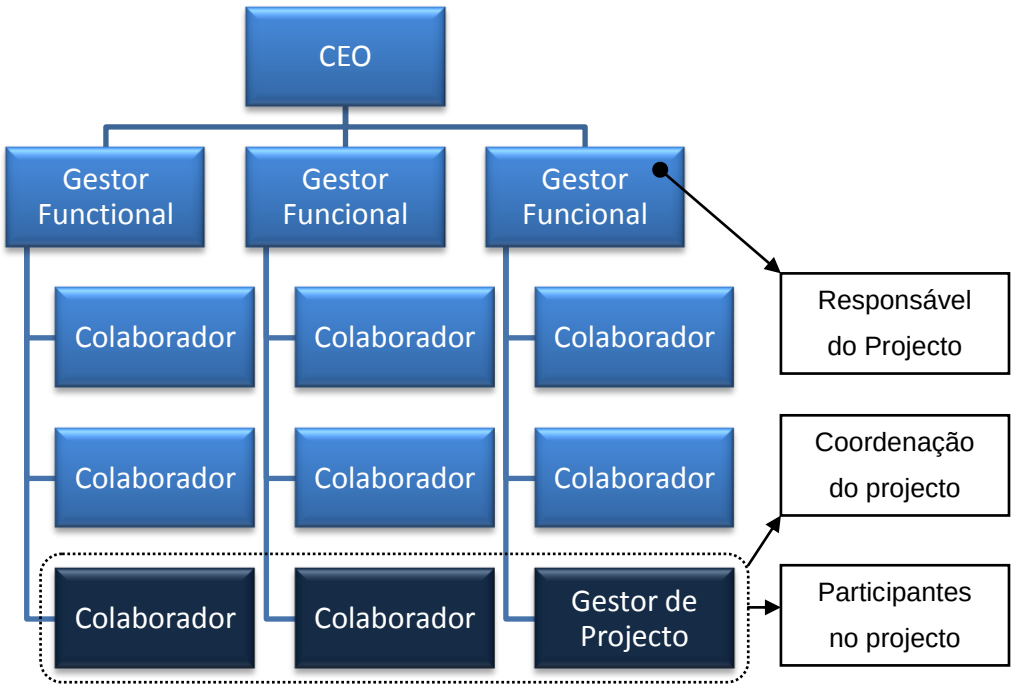
O modelo matricial é caracterizado por se situar entre o modelo orientado ao projecto e o modelo funcional. Existem 3 variações: fraco, balanceado e forte. A grande diferença entre eles reside no grau de responsabilidade, autoridade e autonomia do gestor de projectos.



Esquema 3 - Modelo matricial fraco

O modelo matricial fraco mantém muitas das características do modelo funcional. A coordenação do projecto é feita por um dos colaboradores participantes no projecto, embora as suas funções sejam mais no sentido de coordenar e orientar o trabalho do que fazer a gestão do projecto. A autoridade do Gestor de Projecto é limitada [25].

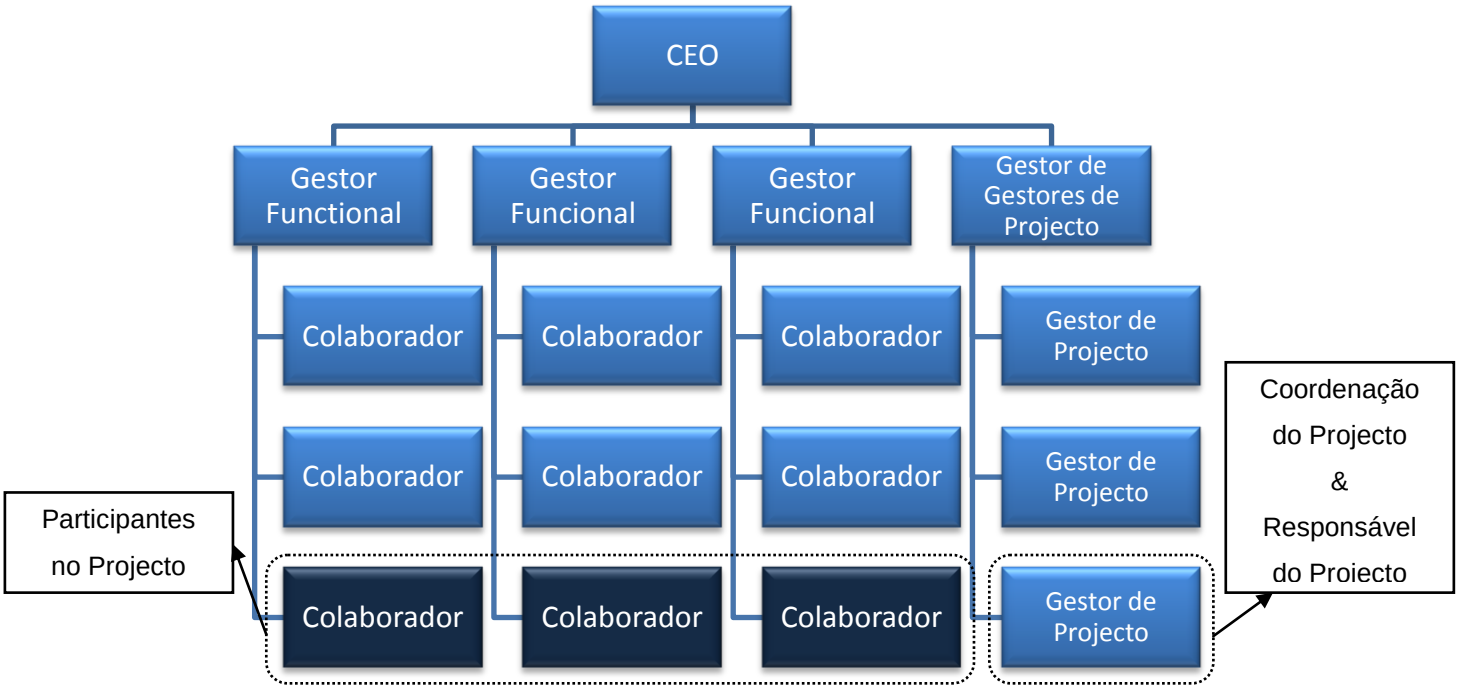
O modelo matricial balanceado caracteriza-se pelo reconhecimento, por parte da organização, da necessidade de existência de um Gestor de Projecto (a tempo inteiro) com mais autonomia, embora não lhe atribua total autoridade, pois os recursos ainda estão dependentes da área funcional.



Esquema 4 - Modelo matricial balanceado

Nos modelos matriciais começamos a observar uma tendência para alocar recursos com responsabilidades exclusivas na gestão de projectos.

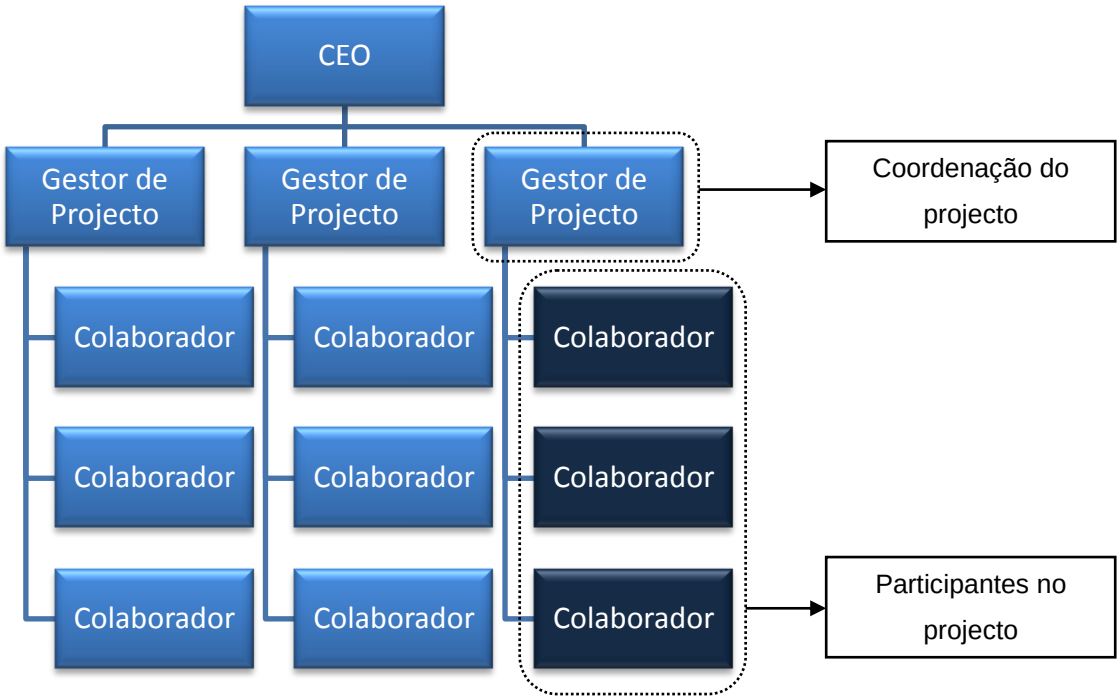
No modelo matricial forte, a organização reconhece que a Gestão de Projectos é uma área funcional específica, e que os gestores devem ter uma maior autonomia e autoridade.



Esquema 5 - Modelo matricial forte

No espectro oposto da organização funcional encontramos a organização orientada ao projecto. Neste tipo de organização, os colaboradores são alocados para um projecto e

respondem perante o seu responsável (Gestor de Projecto), que tem autonomia e controlo total sobre o mesmo. Este modelo não se encontra implementado na DSI.



Esquema 6 - Organização orientada ao projecto

4.3 Mudança organizacional

Na análise efectuada torna-se visível uma mudança na organização. No final de 2006, a DSI começou a mudar de uma estrutura com um modelo funcional rígido para uma organização de matriz forte.

A equipa de gestão de projectos começou a operar no 2º trimestre de 2009[24] e essa alteração marcou a passagem da organização para um modelo matricial forte. Os impactos da mudança organizacional tornam-se visíveis quando medimos o total dispendido em gestão de projectos e o total dispendido em projectos ao longo do tempo, consultar *Gráfico 5.* Constata-se uma acentuada descida sem, no entanto, haver uma diminuição equivalente no esforço total disponível (capacidade).

[24] Existiram alguns projectos, anteriores ao 2º trimestre de 2009, que funcionaram como actividades piloto/teste; estes projectos não foram contabilizados

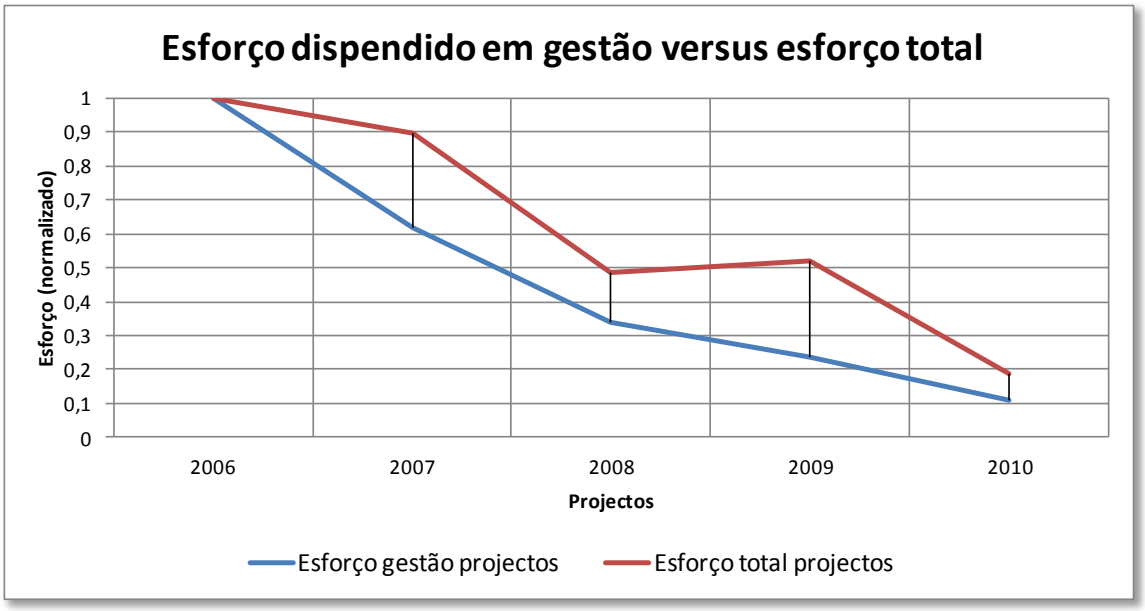


Gráfico 5 - Consumo normalizado do esforço dispendido (total e gestão) para projectos

Uma das consequências desta mudança organizacional foi a eliminação de alguma redundância na gestão. Dentro do contexto de um projecto apenas um recurso despende tempos em tarefas de gestão; o seu Gestor de Projecto. A prova desta afirmação é o *Gráfico 6* que apresenta o consumo global de esforço na organização.

Se existem menos tempos imputados em gestão e não há quebras significativas na disponibilidade[25] pode-se concluir que o esforço anteriormente utilizado em gestão está a ser aproveitado noutras tarefas.

[25] A disponibilidade é a quantidade de esforço disponível para ser consumido num determinado período de tempo. A disponibilidade só faz sentido quando falamos de tempo futuro.

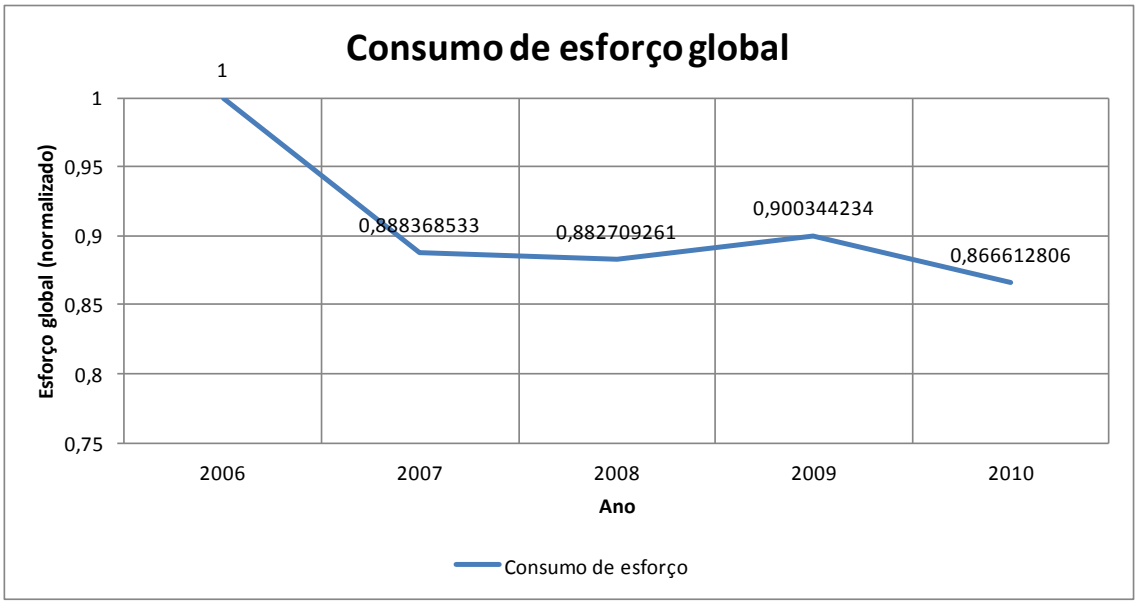
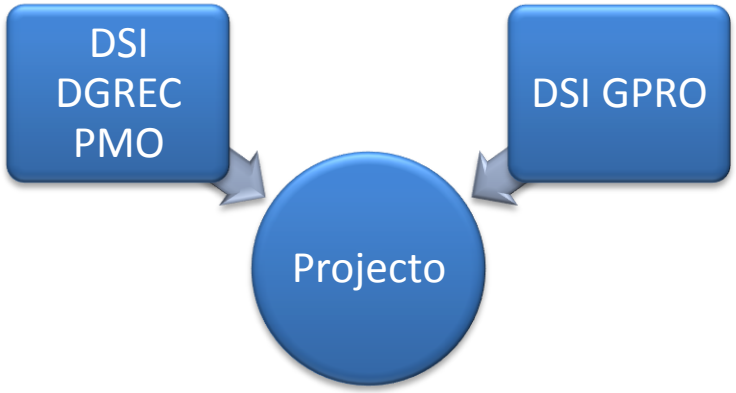


Gráfico 6 - Evolução do consumo do esforço global ao longo do tempo

4.4 Impacto organizacional

Organigrama de áreas funcionais envolvidas:



Esquema 7 - Áreas funcionais intervenientes no projecto

A DSI intervém através de duas áreas o *Portfolio Management Office* (PMO) e a área de Área de Gestão de Projectos (GPRO). O PMO é responsável pelas fontes de informação e está dependente da Direcção de Gestão de Recursos e Contractos (DGREC). A Área de Gestão de Projectos (GPRO) desempenha o papel de utilizador final.

Responsabilidades **DSI-DGREC-PMO**

- Fornecer informação acerca dos processos operacionais sobre os processos de gestão de projectos, *Project Portfolio Management (PPM*).
- Fornecer as fontes de dados contendo a informação acerca do PPM.
- Através dos seus especialistas de negócio, validar as conclusões iniciais do processo.
- Disponibilizar o acesso às aplicações e outros dados.
- Garantir que algumas informações estão devidamente ofuscadas

Responsabilidades **DSI-GPRO**

- Validar o modelo final
- Testar o modelo contra casos reais
- Aceitação final

4.5 Controlo de tempos

Para melhor se entender a análise apresentada, será necessário, futuramente, caracterizar a utilização dos recursos nos seus vários contextos. Trata-se de encontrar uma norma, valores de *benchmark,* que possam ser utilizados em futuras comparações.

Muito deste trabalho desenrola-se à volta de do conceito de controlo de tempos. Na prática, o que o controlo de tempos representa, é a rastreabilidade do tempo gasto pelos colaboradores. Existem essencialmente três escalas de controlo: o tipo de tarefa, a tarefa e a descrição da tarefa. Ao preencher a sua *timesheet*[26] o colaborador escolhe um projecto, uma tarefa e faz uma pequena descrição do trabalho executado, vejamos o exemplo:

Projecto	Tarefa	Data	Horas	Comentário
Projecto A	Análise Orgânica	01/01/2000	7	Análise de documentos
Projecto B	Desenvolvimento	02/01/2000	5	Criação da rotina X
Projecto C	Manutenção de aplicação. Erros	02/01/2000	2	Análise e correcção do erro Z reportado pelo utilizador Y.

Tabela 5 - *Timesheet*

As tarefas, por sua vez, estão agrupadas por tipo para que possam ser tratadas com menos detalhe, por exemplo:

Tarefa	Tipo de Tarefa
Gestão de equipa, planificação de recursos	**Gestão**
Criação do DP / Análise Funcional	
Análise Orgânica	
…	**…**

Tabela 6 - Tipos de tarefa (exemplo)

Este agrupamento por tipos permite-nos analisar o esforço empregue pela DSI de um modo muito mais especializado porque apenas queremos estudar os tempos relacionados com novos projectos. As observações apresentadas nos próximos pontos são baseadas em informação extraída do controlo de tempos e agrupamento da informação.

4.5.1 Projectos

Qual o comportamento "normal" de um projecto e qual a média no que diz respeito aos consumos de tempo? Para responder a esta questão temos de ter em conta a reestruturação que foi levada a cabo em 2006, e que afectou os projectos especialmente a partir de 2007.

[26] Uma *timesheet* é a descrição de todas as horas trabalhadas por um colaborador.

Esta reestruturação, como foi referido anteriormente, consistiu na evolução de uma organização com uma estrutura funcional (consultar *Esquema 2*) para uma estrutura de matriz forte (consultar *Esquema 5*).

A mudança organizacional, como seria de esperar, reduziu os tempos de gestão e da análise, focando as equipas no desenvolvimento e certificação.

Os valores do *Gráfico 7* espelham os valores antes da reestruturação.

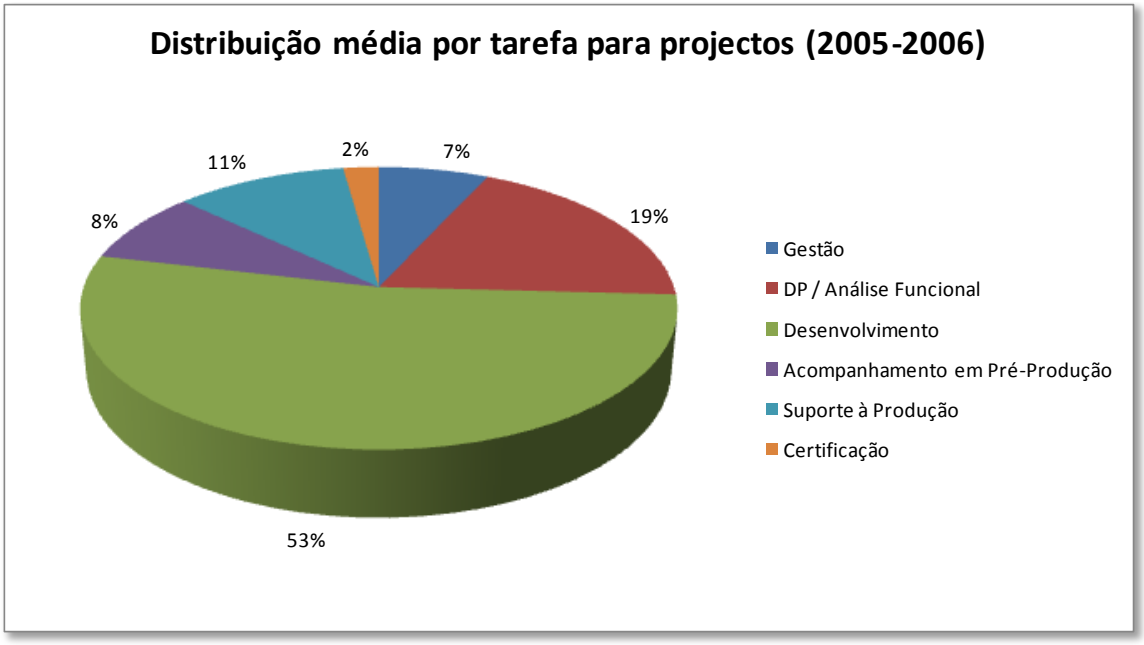


Gráfico 7 - Consumo médio por tarefa de projectos entre 2005 e 2006

Os valores do *Gráfico 8* espelham os valores depois da reestruturação

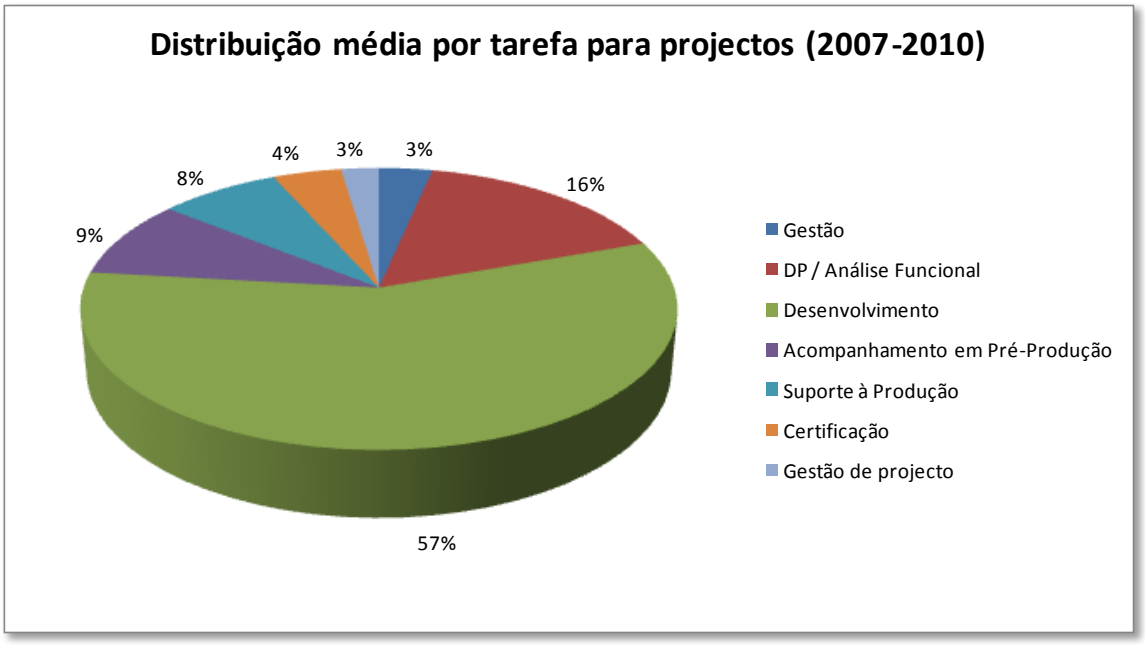


Gráfico 8 - Consumo médio por tarefa dos projectos entre 2007 e 2010

Pelas razões expostas, em futuras comparações e *benchmarking* serão utilizados sempre os valores entre 2007 e 2010 a não ser que seja explicitamente indicado o contrário.

Ao começar um projecto, a organização recolhe a duração esperada, através da análise funcional, que consta no DP[27]. No final de um projecto, é calculada a duração real (a diferença, em dias, entre o primeiro e ultimo registo). Através destas medições, calcula-se a duração real média e a duração esperada média. O *Gráfico 9* faz uma comparação entre os valores esperados e os valores reais obtidos através do controlo de tempos.

[27] DP – Dossier de Projecto, a mais recente contratualização do projecto, consultar ponto 4.1.3 *Procedimento operacional para novos projectos.*

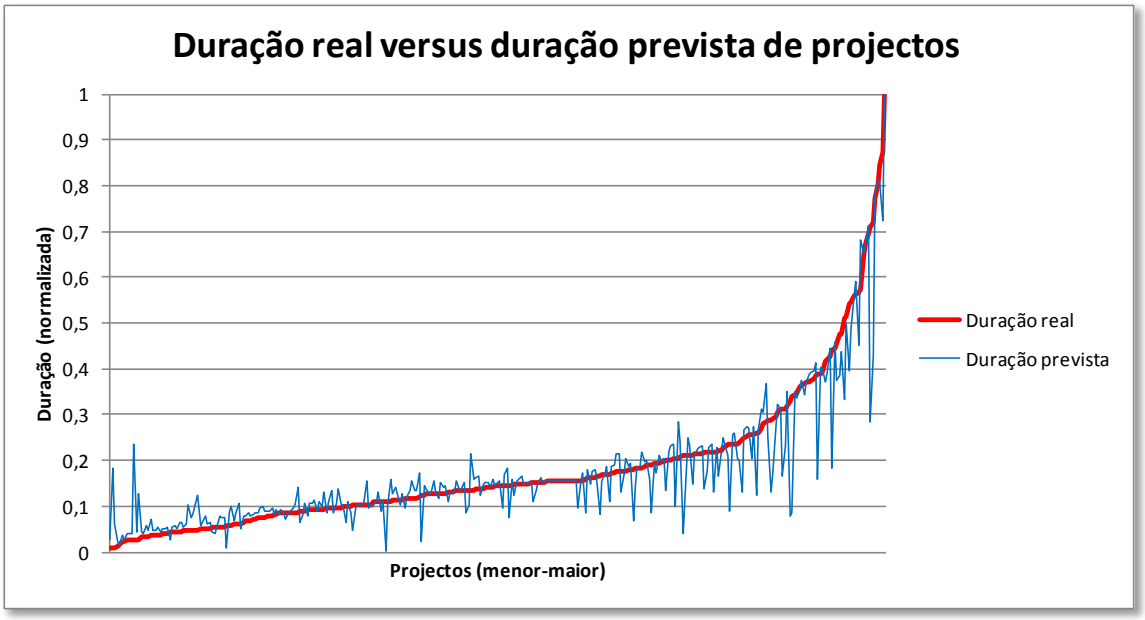


Gráfico 9 - Duração real e duração prevista para projectos terminados (2007-2010)

	Duração real	Duração prevista
Média	0,180483	0,174419
Variância	0,023986	0,020048

Tabela 7 - Média e variância da duração real e prevista

Uma análise aos valores apresentados leva-nos a concluir que parece haver uma tendência para subestimar a duração.

Neste estudo não estão a ser contemplados os projectos que foram anulados, onde ocorreu desistência do cliente, os projectos duplicados ou rejeitados. Se fossem contemplados a diferença iria ser significativamente maior visto que se um projecto ultrapassar em demasia o esforço e o seu benefício já não for compensador, o projecto pode ser anulado. Se a janela de oportunidade de negócio estiver fechada, este pode sofrer uma desistência do cliente. Se um projecto for integrado num outro projecto, por afinidade funcional ou janela de oportunidade, o projecto é considerado duplicado e os trabalhos prosseguem num outro projecto. No *Gráfico 10* podemos ter uma visão sobre a quantificação destes projectos e a razão genérica da sua desistência.

A razão pela qual estes projectos não são incluídos na análise reside no facto de, na maior parte das vezes, serem abortados por factores completamente imprevisíveis relacionados com o negócio. Não existindo informações sobre esse processo de decisão, já que muitas vezes são decisões de carácter financeiro ou estratégico não disponibilizado pela DSI, não estão incluídas na análise.

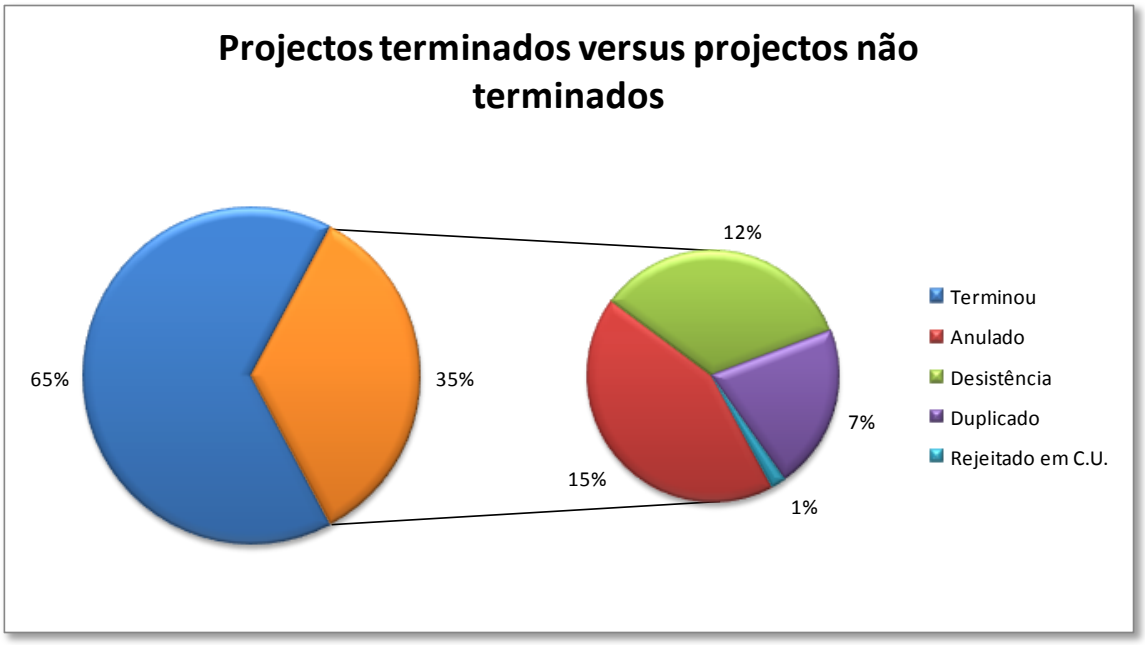


Gráfico 10 - Percentagem de projectos terminados e não terminados

Embora pareça haver uma tendência no decréscimo de projectos (consultar *Gráfico 11*) cancelados pelos clientes, é necessário ter em conta que ainda existem projectos de 2007, 2008, 2009 e 2010 a decorrer que podem ser cancelados até ao fim do seu ciclo de vida. No entanto, a proporção de projectos a serem alvo de cancelamento acompanha a curva de esforço de projectos (consultar *Gráfico 5*).

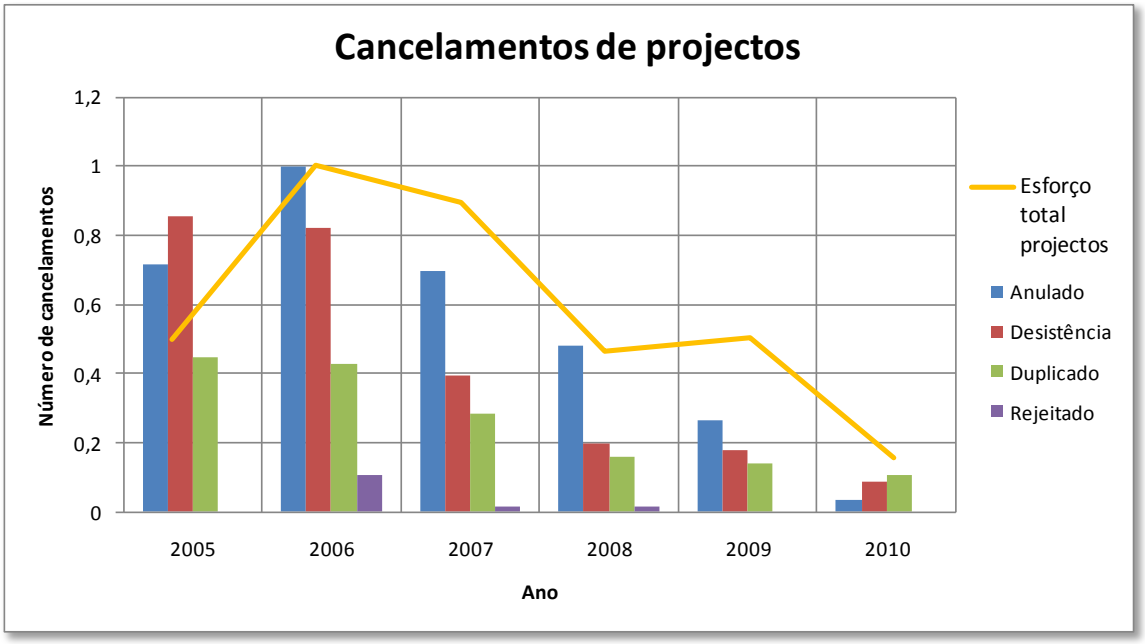


Gráfico 11 - Cancelamentos de projectos, tendência temporal e justaposição com o esforço total para projectos

4.5.2 Colaboradores

Cada colaborador, além das suas responsabilidades normais perante a sua área funcional e respectivo responsável (Gestor Funcional), pode participar em projectos.

Gráfico 12 - Percentagem de tempo alocado a projectos por colaborador

Não serão caracterizadas as diferentes tarefas agrupadas em "outros", mas fica aqui a nota de que estão a ser contabilizados as férias, os feriados, as licenças, as ausências, as formações e as baixas médicas que perfazem uma margem relativa de ≈30% (≈17% global).

No entanto, a caracterização do tempo dispendido nos projectos é interessante para o contexto deste relatório, e é equivalente à distribuição presente no *Gráfico 7*, e no *Gráfico 8* para os períodos de tempo equivalentes.

4.6 Solução actual

Não foram identificadas soluções prévias de mineração de dados para a resolução do problema proposto. Actualmente, os gestores de projecto baseiam-se numa *Framework de Gestão de Projectos*[28] específica à DSI e fortemente baseada no PMBOK [24]. Os Gestores de Projecto não têm nenhuma ferramenta automatizada e baseada em modelos de mineração de dados para prever o risco.

Foram identificados os seguintes pré-requisitos:

[28] Informação obtida através de um conjunto de entrevistas realizadas com vários especialistas de negócio.

- Os resultados teriam de ser submetidos a uma ofuscação, de modo a reduzir a sua legibilidade; mantendo no entanto a sua significância global;
- Os modelos teriam de ser aprovados pelo PMO, e, depois, pela GPRO, antes de se proceder à fase de disponibilização;
- A ferramenta de gestão de modelos teria de ser isenta de custos adicionais para a DSI, assim como de qualquer contracto de licenciamento;
- Teriam de ser respeitados os calendários propostos no *Regulamento geral dos ciclos de estudos conducentes ao grau de mestre* [26].

5. Implementação

Neste capítulo são descritos os diferentes passos, de acordo com a metodologia descrita, até ser atingido o objectivo final.

5.1 Compreensão do problema

Nesta fase foram analisados os processos (descrito no ponto *4.1.4. Procedimento operacional para novos projectos*) e todos os repositórios que os suportam ou que poderiam contribuir com informação relevante para a caracterização do modelo. Foram realizadas uma série de entrevistas com especialistas da metodologia, com Gestores de Projecto, com especialistas do negócio e com responsáveis do desenvolvimento. Foram realizadas um total de 12 entrevistas, ao longo de 3 meses, que resultaram no levantamento e compreensão do processo.

Para melhor definir o problema, do ponto de vista da mineração de dados, temos primeiro de compreender quais os objectivos que o negócio pretende atingir. Trata-se da classificação qualitativa, quanto ao risco de desvio, que um projecto novo apresenta para a DSI.

De um ponto de vista de mineração de dados, o mesmo problema pode ser definido como a classificação de novos projectos em intervalos de risco segundo a fórmula: $]0, x_n]\cup]x_{n+1}, x_{n+2}[\cup[x_{n+m}, \infty[$, cada intervalo é denominado de classe e o número de classes será determinado pelos especialistas e pelos resultados da mineração de dados. Para resolver o problema, seguindo a metodologia descrita anteriormente primeiro será necessário classificar os projectos terminados. Depois de classificados os projectos segundo um modelo de sucesso teremos de examinar a informação disponível acerca da sua execução e determinar a melhor caracterização possível do projecto.

5.1.1 Pressupostos e restrições

Não foram identificados requisitos de precisão, fiabilidade e manutenção nesta fase do projecto. O grupo alvo de utilizadores considera o projecto como uma prova de conceito, por esse motivo os requisitos serão definidos depois de uma eventual demonstração de resultados.

Pressupõe-se que todos os dados fornecidos pela DSI têm uma qualidade elevada e representam a realidade. A fiabilidade dos dados é periodicamente submetida a auditorias e é um assunto que não será abordado neste relatório.

Foi exigido, por parte do negócio, a ofuscação de alguns dados, de modo a que estes não se tornassem legíveis para pessoas externas à DSI. No entanto, esta alteração foi feita de modo a não influenciar os resultados finais em termos de objectivos. Serão também omitidas algumas descrições das estruturas de dados e alguns atributos serão renomeados para se introduzir clareza. Estas alterações serão feitas de modo transparente ao leitor. Por esta razão não será

transcrita nenhuma informação constante nos repositórios da DSI, sem que esta sofra algum tipo de transformação que a torne irreconhecível (mas compreensível).

5.2 Compreensão dos dados

Nesta fase efectuou-se um levantamento dos dados que suportariam o processo. Para tal, foram analisados os oito sistemas que suportavam a metodologia. Desses foram seleccionados três. Depois de devidamente autorizado o acesso às fontes de dados, pela DSI, foi delineada uma estratégia de carregamento para um quarto repositório, que foi desenhado para o efeito. Tiveram de ser desenvolvidos métodos particulares para a extracção de certos repositórios, nomeadamente o SharePoint. Foi determinado o tamanho do histórico a utilizar e foram também definidos alguns critérios de selecção. Nomadamente quais os dados relevantes, quais os irrelevantes e a sua localização. Foram também desenvolvidas estratégias de uniformização dos repositórios e a eliminação da informação que não fosse transversal (que constasse em todo os repositórios).

5.2.1 Origem dos dados

Comecemos por descrever as diferentes fontes de informação utilizadas para familiarizar o leitor com alguns dos termos e conceitos empregues. Este ponto descreve os principais sistemas cuja informação alimentou o modelo final. Em suma, descreve os três sistemas que alimentaram a suposição inicial, levando à realização deste trabalho.

Artefactos – É o nome de uma aplicação que serve de suporte à metodologia. Contém, entre outras informações relativas ao projecto, os requisitos de negócio, as análises das equipas, os requisitos de negócio, as estimativas, o planeamento e a aceitação do cliente. Este tipo de software cai na classe de *Project Portfolio Management* (PPM) Software

Tipo de informação	Desta aplicação retiramos duas informações relevantes. A base de dados relacional e o seu *workflow* de processos.
Base de dados	Microsoft SQL Server 10.0.4064.0 Permissões (offline backup): *data base owner*
Fluxo de processo	O fluxo foi obtido através da geração automática para Microsoft Office Visio 2007 e pelo *Microsoft Visual Studio* 2010®, via código fonte
Acesso a especialista	Sim, ao arquitecto do sistema, ao especificador e ao administrador.
Acesso à documentação	Sim.

Tabela 8 - Caracterização dos Artefactos

SPI – É o nome de uma aplicação que gere a *timesheet*[29], o controlo de estados, o planeamento e a orçamentação do projecto. Os especialistas, de modo informal, chamam ao

[29] Consultar *4.5 Controlo de tempos*

SPI o registo “de facturação”. Este tipo de aplicação pertence à classe de *Time Tracking Software* (TTS)

Tipo de informação	Base de dados relacional
Base de dados	Microsoft SQL Server 2000 Permissões (*offline*): *data base owner*
Acesso a especialista	Sim, ao utilizador administrador
Acesso à documentação	Não.

Tabela 9 – Caracterização do SPI

***Microsoft Sharepoint 2007* –** Neste contexto, serve como repositório colaborativo de informação formal (aprovada e conforme a metodologia): Estudos de Viabilidade (EV), Dossier de Projecto (DP) e Relatório de Alterações de Projecto (RAP), Solução Genérica e Parecer Técnico da IARQ (consultar 4.1.4).

Tipo de informação	Documentação
Base de dados	Acedida via interface MOSS[30] Services 2007
Acesso a especialista	Sim, ao utilizador administrador
Acesso à documentação	Sim.

Tabela 10 - Caracterização do *Sharepoint*

Toda a informação necessária foi carregada no dia 08 de Janeiro de 2011 às 00:01. Tendo sido implementado, para o efeito, um *freeze*[31] aos sistemas. Um dos objectivos de fazer um único carregamento é testar os projectos reais posteriores à data referida.

Critérios relevantes para a selecção de informação e sua localização:

Localização (listados por relevância)	Critério
Artefactos, SPI	Orçamentação de projectos
SPI, Artefactos	Planificação de projectos
Artefactos, SPI	Requisitos de negócio e sua análise
Artefactos, SPI	Desvios, impactos e suas origens
Sharepoint	Documentação das diferentes fases do projecto
SPI	Informação relativa à execução de uma actividade, *timesheet* do projecto

Tabela 11- Localização da informação por tipo

[30] *Microsoft Office SharePoint Server*
[31] Um *freeze* uma indisponibilização temporária dos sistemas ao seus utilizadores. Por vezes é designado por *offlining*. A execução de um *freeze* garante a consistência dos dados.

Nesta fase, foram examinadas as estruturas das fontes de informação, feitas correlações com os processos de negócio e feito um levantamento aos documentos necessários à execução do trabalho.

A uniformização das diferentes fontes de informação levou à decisão da criação de um quarto repositório, construído apenas com a informação relevante à fase de modelação. Este repositório, doravante conhecido como *Data Mart* (DM), foi fabricado através da unificação da informação existente nos sistemas SPI, Artefactos e *Sharepoint*. Para tal, verificou-se a necessidade de criar alguns atributos que mantivessem a consistência entre as diferentes bases de dados. Estes atributos servem apenas para possibilitar o processo de criação do DM e não são considerados no processo de modelação.

Nome	Tipo*	Origem**	Descrição
EquipaID_SPI	I	A	Contém o ID da tabela correspondente em SPI
TipoProjectoID_SPI	I	A	Contém o ID da tabela correspondente em SPI
CicloID_SPI	I	A	Contém o ID da tabela correspondente em SPI
UserID_SPI	I	A	Contém o ID da tabela correspondente em SPI
EstadoID_SPI	I	A	Contém o ID da tabela correspondente em SPI
DirecçãoID_SPI	I	A	Contém o ID da tabela correspondente em Artefactos
FaseID_AV	I	S	Contém o ID da tabela correspondente em Artefactos
ClienteID_AV	I	S	Contém o ID da tabela correspondente em Artefactos
PlanoDP_ID	I	S	Contém o ID da tabela correspondente em Artefactos
EEID_AV	I	S	Contém o ID da tabela correspondente em Artefactos

Tabela 12 – Exemplo de novos atributos, referências externas

*Consultar *Apêndice 3 – Restrições de domínio de atributos*

**A- artefactos ; S -SPI

Os dados foram introduzidos fazendo corresponder os registos de um sistema para o outro. Tipicamente, estas correspondências são estabelecidas por algoritmos alfanuméricos. Vejamos o seguinte exemplo:

Atributo em SPI	Atributo em Artefactos
YYY /XXXX/ZZZZ - <texto>	YYY-XXXX-ZZZZ – <texto>

O algoritmo que faz a correspondência entre as equipas assume a seguinte forma:

```
SELECT
      T2.ID av2ID
      , T2.Descricao AS AVOriginal
      ,    RTRIM(SUBSTRING(T2.Descricao,    5,CHARINDEX    ('    -    ',
T2.Descricao)-5)) AS AVTratado
      , T1.id spiID
      , T1.Descricao AS SPIOriginal
      ,   REPLACE(REPLACE(SUBSTRING(T1.Descricao,   6,CHARINDEX   ('-',
T1.Descricao)-6),' ',''),'/','-') AS SPITratado
FROM
      T1, T2
WHERE
    CHARINDEX ('-', T1.Descricao)>0
AND CHARINDEX (' - ', T2.Descricao)> 0
AND  RTRIM(SUBSTRING(T2.Descricao, 5,CHARINDEX  (' - ', T2.Descricao)-
5)) collate SQL_Latin1_General_CP1_CI_AS LIKE
REPLACE(REPLACE(SUBSTRING(T1.Descricao,       6,CHARINDEX       ('-',
T1.Descricao)-6),' ',''),'/','-') collate SQL_Latin1_General_CP1_CI_AS
```

Tabela 13 - Listagem de código SQL, correspondência alfanumérica

As demais correspondências são fruto da utilização de técnicas semelhantes.

5.2.2 Análise volumétrica de dados

Análise volumétrica das fontes de informação	
SPI (informação recuperada a partir do *SQL Server 2008 Reports Generator*)	
Espaço de total utilizado	18.584,88 MB
Tabelas	68
Número total de registos	≈ 59.000.000
Número total de projectos úteis	872
Número médio de equipas por projecto	4
Número médio de colaboradores por equipa	7
Número de horas registadas	≈ 2.500.000
Artefactos (informação recuperada a partir do *SQL Server 2008 Reports Generator*)	
Espaço de total utilizado	3.383,25 MB
Tabelas	164
Número total de registos	≈ 10.000.000
Número total de projectos úteis	830
Número médio de requisitos de negócio por equipa por projecto	2
Tamanho médio de cada análise de requisito	1225 Palavras
Sharepoint (Informação obtida via MOSS2007)	
Número de documentos	2616
Número de EV	811
Número de DP	305
Número de RAP	136
Outros documentos	1368
Tamanho médio de documento	2.321,12 KB

Tabela 14 - Análise volumétrica dos dados

5.2.3 Caracterização de atributos e valores

Do universo de atributos disponíveis para o projecto, cujo vasto tamanho pode ser constatado pela análise do ponto anterior, foi escolhida uma pequena porção para constar no DM. Estes são os "atributos relevantes", e serão, directamente ou indirectamente, utilizados nos indicadores KPI utilizados pelo modelo e explicados mais à frente (consultar *5.3.2.*)

Porque razão foi feita esta selecção? Os conjuntos de atributos descritos nas *Tabela 15* e na *Tabela 16* foram eleitos porque:

i. Atributos necessários para calcular indicadores que, à partida, foram identificados para ser utilizados na fase seguinte. A origem dos indicadores é detalhada no ponto *5.3.2 Atributos derivados*.

ii. Opinião de peritos como gestores de projecto, especialistas em gestão e os orientadores.

SPI		
Nome	**Descrição**	**Tipo**
P1.ProjectoID	Identifica univocamente um projecto	I
P1.TipoProj	Identifica o tipo de projecto. Uma vez que os recursos imputam as horas em projectos existem vários tipos. Eis alguns exemplos: • Ausências – projectos onde são registadas as baixa médicas, férias e outras ausências. • Manutenção – projectos que representam horas de manutenção (erros de programação, indisponibilidades, etc.) de aplicações • Projecto – Novos desenvolvimento • Evolutivos – Novos desenvolvimento com uma dimensão ligeiramente menor que os "projectos" O significado deste atributo será detalhado um pouco mais á frente	I
P1.Prioridade	Um valor que designa a prioridade da aplicação depois de ser classificada pelo Comité de Sistemas (consultar *4.1.3 Procedimento operacional para novos projectos*)	I^{+}
P1.Estado	Designa o estado actual do projecto. A análise recairá apenas sobre projectos nos estados • Concluído-Prod. Isento de Aceitação • Concluído-Prod. Aceite Cliente Este atributo encontra-se detalhado mais à frente.	I

P1.DataComite	Data na qual o Comité de Sistemas aprovou deu uma prioridade a este projecto. Uma actividade pode ir a n comités. Este atributo está na forma "dd-mm-aaaa;dd-mm-aaaa;…"	S^n
P1.Origem	Direcção à qual pertence o *sponsor* do projecto	I
P1.Cliente	Cliente principal, *sponsor* do projecto	I
P1.Intercalar	Determina se a actividade é intercalar (consultar *4.1.3 Procedimento operacional para novos projectos*)	B
P1.EV_Isento	Determina se o EV[32] está isento de aprovação. Quando isento um projecto é automaticamente aprovado quando finalizado. Os projectos estratégicos ou restringidos a uma imposição legal não precisam de aprovação explícita	B
P1.DP_Isento	Idem para DP[33]	B
P1.EV_DataAprova	Data em que o EV foi aprovado	D
P1.DP_DataAprov	Data em que o DP foi aprovado	D
P1.ValorAno	Benefício anual esperado.	M
P1.Despacho	Grau de aceitação do cliente. A aceitação do cliente deriva do preenchimento de um formulário que no final calcula uma nota. Essa classificação está codificada em letras: E-excelente, B-bom, S-suficiente, M-mediocre, U-mau.	S
P1.PDP_DataPrevista	Última data contratualizada com o cliente para o fim da actividade.	D
P1.PEQ_DataPrevista	Última previsão da equipa para o término da actividade	D
E1.Descricao	Designação completa da entidade (quando se trata de um nó do qual ninguém depende a entidade é um equipa)	
E1.ParentID	Entidade hierarquicamente superior	I
E1.Resposavel	Colaborador responsável pela entidade	I
E1.TipoCusto	O custo, por hora e semana, da entidade.	I
E1.HorasDia	O número de horas por, semana, que os colaboradores da equipa cumprem (35 horas, 40 horas, etc.)	I
P2. Inicio	Data na qual foi contratualizado o início da actividade	D
P2.PreProducao	Data na qual foi contratualizada a passagem a pré-	D

[32] EV é o Estudo de Viabilidade; um estudo provisório que acompanha o pedido do projecto e apenas contém uma estimativa preliminar do esforço. Consultar ponto *4.1.3 Procedimento operacional para novos projectos*.
[33] DP significa Dossier de Projecto; o contracto feito com o cliente. Consultar ponto *4.1.3 Procedimento operacional para novos projectos*

	produção. Na pré-produção são feitas passagens em larga escala para ambientes de qualidade [34] onde é feita a aceitação final e se prepara a passagem a produção.	
P2.Producao	Data na qual foi contratualizada a passagem a pré-produção.	D
P2. Fim	Data na qual é oficialmente terminado o projecto. Depois da produção, existe um período "de garantia" que a equipa reserva para a correcção imediata de qualquer problema que possa surgir em produção. Quando esse período termina, é concluída a actividade	D
EP1.EstadoID	Históricos do estado do projecto. A data de início e a data de fim. Através desta tabela podemos detectar se a actividade sofreu muitas alterações ao procedimento normal.	I
Ep1.DataInicio		D
Ep1.DataFim		D
REH.RecursoID	Histórico sobre a que equipa pertenceu cada colaborador, por intervalo de tempo.	I
REH.EquipaID		I
REH.DataInicio		D
REH.DataFim		D
T.Data	Data a qual se refere o registo de horas	D
T.RecursoID	Recurso ao qual está a efectuar o registo de horas	I
T.ProjectoID	Projecto a que se refere o registo de horas	I
T.TipoTarefaID	Tarefa, no contexto do projecto, a que se refere o registo de horas.	I
T.EquipaID	Equipa em que o colaborador estava inserido na altura do registo	I
T.Horas	Número de horas trabalhadas	I
T.TipoDeCusto	Esta rubrica determina onde vai ser facturada a hora (nos casos onde não queremos facturar ao cliente)	I
T.Aprovado	A hierarquia do recurso aprovou o registo de horas	B
T.Comentarios	Descrição das horas. Em certos casos a aplicação utiliza este campo para facturar outras actividades introduzindo uma formatação especial. • Alguns projectos são bolsas globais para clientes e, quando existe autorização, a facturação do projecto	S

[34] Em ambientes corporativos é comum existirem três ambientes idênticos e paralelos de execução de aplicações. Desenvolvimento, Qualidade e Produção. O ambiente de qualidade é muito semelhante ao de produção (ao contrário do desenvolvimento que apenas contém dados mecanicamente gerados).

	pode ser feita aqui. • As equipas, quando autorizadas, podem utilizar as suas bolsas para imputar horas nos projectos	
CD.CustosAdicionais	Quantificação dos custos adicionais para a equipa	M
CD.TipoCustoAdicional	Este atributo qualifica o tipo de custo adicional. Este atributo é particularmente interessante se o tipo for equipa externa	I
CD.Descricao	Descrição do custo externo. Este atributo é relevante se for uma equipa externa pois aqui ficará descrito o nome da empresa que prestou o serviço.	S

Tabela 15 - Caracterização de atributos SPI

Artefactos		
Nome	**Descrição**	**Tipo**
P1.DataCriacao	Data em que foi registado o projecto da perspectiva do desenvolvimento. A partir desta data o desenvolvimento teve conhecimento desta actividade, o que significa que o Projecto já foi avaliado, analisado do ponto de vista do negócio e o desenvolvimento deve intervir	D
P1.Interna	Este booleano indica se a actividade é interna. Uma actividade interna tem como cliente a própria DSI. Estas actividades são tipicamente melhorias às infra-estruturas, estudos, optimizações, migrações, *upgrades*, etc.	B
P1.GestorCliente	Colaborador responsável pela comunicação directa com o cliente. Se a DSI se tratasse de uma *software house* seria alguém do departamento comercial. O GR, no contexto deste estudo, gere as expectativas do cliente quanto ao trabalho desenvolvido pela DSI, faz a análise aos requisitos de negócio e negoceia os projectos tentando satisfazer ambos os lados.	I
PDP.DataInicio	Data prevista para o início da documentação DP	D
PDP.DataFim	Data prevista para o fim da documentação DP	D
PDP.DataCriacao Documento	Data na qual o DP foi finalizado.	D
PDP.DataEnvio Cliente	Data na qual o documento foi enviado para aprovação do cliente	D
PDP.DataAprovacao	Data na qual o documento DP foi aprovado por todos os clientes. Esta data marca o fim da fase de negociação entre o cliente e a DSI. Ao aceitar este documento o cliente aceita apenas os desenvolvimentos que nele vêm descritos. Na primeira frase do documento pode ler-se: *"(…) está*	D

	explicitamente excluído tudo o que não estiver implicitamente incluído (…)"	
PDP.CustoTotal	Este campo descreve o custo total final que se orçamentou para o projecto.	M
PRP.Alteracao	Depois do contracto estar assinado entre o desenvolvimento e o cliente (DP aprovado), existem algumas situações que podem obrigar uma das partes a renegociar esse contracto. Essas situações podem incluir: • Alteração do âmbito do projecto por parte do cliente; • Incapacidade do desenvolvimento cumprir prazo/esforço orçamentado; • Factores externos sobrepõem-se ao projecto (como por exemplo imposições legais); • Outras situações. Estas alterações serão documentadas, pelo seu tipo e número de ocorrências.	I
PRP.TipoRP		I
OR.EquipaID	Estes atributos servem para registar as origens possíveis do desvio ao que foi estabelecido no DP. Cada registo pode indicar que a origem reside num cliente, numa equipa ou num outro factor externo. Ao contrário do indicador SPI, que só guarda o *sponsor*, este sistema mantém registos de todos os clientes.	I
OR.ClienteID		I
OR.Descricao		S
PEC.Esforco	O orçamento, de esforço previsto, em dias homem para uma equipa.	R
PEC.CustosAdicionais	O orçamento dos custos adicionais, em euros, da equipa no projecto.	M
PEC.Total	O esforço, convertido para euros (multiplicando as horas pelo preço hora) mais os custos adicionais.	M
GP.EstadoGestao	Atributo que controla o fluxo dos documentos da gestão de projectos. Como foi mencionado anteriormente os Gestores de Projecto têm uma *Framework* para a produção de documentação de acompanhamento dos projectos.	I
GP.IARQ_Emitiu Parecer	Este atributo determina se a equipa IARQ já emitiu o seu parecer para a actividade. Este perecer pode ir da validação das tecnologias utilizadas à simples menção que a IARQ não participará na actividade. Não é possível tipificar a análise feita.	B
GP.GPRO_Intervem	Determina se existe um GP designado para gerir uma actividade	B

C1.ID	Identificação do cliente envolvido na actividade	I
C1.FuncaoID	Determina a função do cliente no contexto do projecto • Leitor. É um dos recipientes da documentação produzida, mas não intervém; • Testa. Participa na fase de aceitação; dá autorização de continuação para produção do projecto; • Aprova. Este cliente negoceia com a DSI a especificação do produto a ser desenvolvido; faz o *sign-off* inicial.	I
C1.DataExecucao Funcao	Este atributo determina em que data o cliente cumpriu a sua função no contexto de um projecto.	D

Tabela 16 - Caracterização de atributos Artefactos

5.2.3.1 Estado do projecto

Durante a execução normal de um projecto este passará por várias fases. Tipicamente o projecto começa no estado "pedido", depois passa para "em execução" e finalmente para "concluído". Para cada um destes estados existem fases intermédias que o projecto pode atingir. Por exemplo, o projecto pode estar no estado "em execução, suspenso" para indicar que embora o projecto de um modo genérico esteja a decorrer naquele preciso momento, por algum motivo, se encontra suspenso.

O mesmo raciocínio pode ser aplicado ao estado "concluído". Um projecto pode estar "concluído, por desistência do cliente", ou "concluído, integrado noutro projecto". É importante entender que apenas queremos analisar os projectos cujo processo correu normalmente até ao fim e acabou com uma entrega ao cliente, dai a filtragem aos estados:

- Concluído-Prod. Isento de Aceitação
- Concluído-Prod. Aceite Cliente

O primeiro caso significa que o projecto já foi concluído e foi para produção. Este projecto, devido às suas características, está isento de aceitação. Os projectos isentos de aceitação são tipicamente imposições legais. No segundo caso o projecto já foi concluído e o cliente já emitiu o parecer quanto à sua aceitação.

5.2.3.2 Projectos e evolutivos

Até agora foi mencionada a palavra "projectos" para descrever actividades de desenvolvimento da DSI. Na realidade, os projectos estão divididos em duas subpopulações, que é importante distinguir: os projectos e os evolutivos.

As regras que os distinguem estão relacionadas com o seu tamanho em termos de esforço e duração. Existem também condicionantes relacionadas com a complexidade, quanto mais

equipas estiverem envolvidas mais complexo é considerado. É importante distinguir porque os projectos evolutivos não são atribuídos a um GP. O responsável da equipa principal faz a gestão de projectos com aquilo que se designa *Framework de gestão simplificada*; uma versão mais simples e menos burocrática do processo de desenvolvimento.

5.2.4 Formulação de suposições

5.2.4.1 Padrão de consumo de esforço

Uma das suspeitas que existia à partida, levaria a crer que a maneira como uma equipa planeia e despende o seu tempo (o consumo de esforço), ao longo do projecto, influenciaria o seu risco.

Segundo o *Rational Unified Process* (RUP), as melhores práticas da indústria apontam para um consumo de esforço, que se assemelhe ao representado na próxima imagem

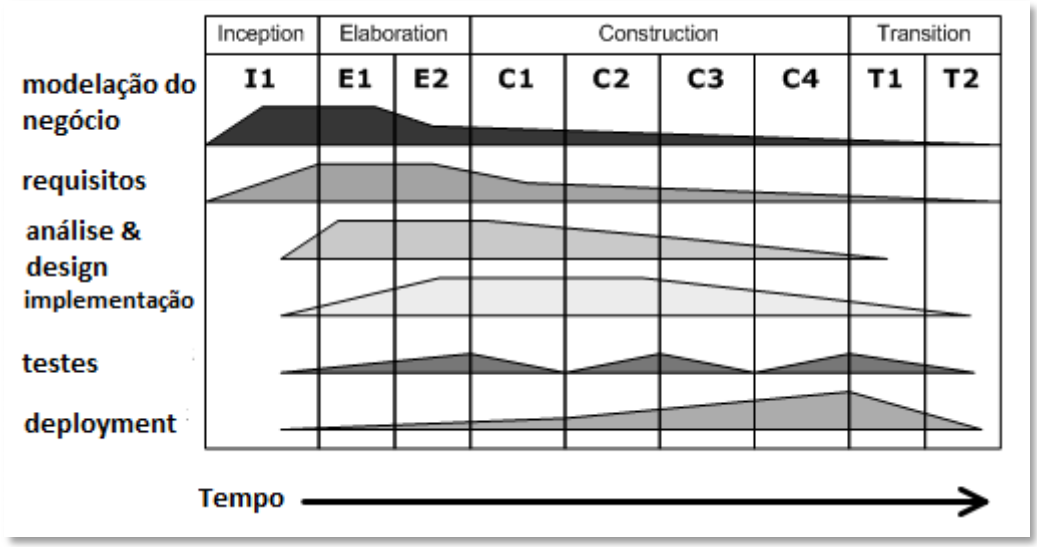


Ilustração 3 - O Processo RUP[35]

Não se pode afirmar que este comportamento é universal para todas as organizações, porque o consumo está intimamente ligado à metodologia utilizada. No caso da DSI será necessário analisar e determinar o comportamento correcto, e depois, de algum modo, classificar as equipas.

Suspeita-se que as equipas, cujo comportamento está longe deste padrão, tenham uma influência negativa no projecto.

5.2.4.2 Intervenientes na actividade

Supõem-se que os intervenientes-chave na actividade têm uma contribuição significativa para o seu sucesso. Para tal, recolheu-se no *Data Mart,* todos os intervenientes no projecto:

[35] Imagem de domínio publico adaptada de http://en.wikipedia.org/wiki/File:Development-iterative.gif

- Os clientes e respectiva direcção;
- O Gestor de Projecto;
- O Gestor de Relação;
- O responsável das equipas que participam;
- Os participantes directos na actividade.

Na fase de modelação será avaliada a contribuição de cada um para o resultado da classificação.

5.2.4.3 Grau de paralelismo do projecto

Na DSI existe uma grande divisão funcional das equipas, havendo por isso um grande encadeamento de participação de equipas. Se, num projecto, existir encadeamento de trabalho, onde uma equipa só pode intervir depois de outra ter executado uma tarefa, os atrasos causam verdadeiras "bolas de neve". Vejamos o seguinte exemplo: para uma determinada actividade a "Equipa A" deveria ter disponibilizado um serviço que continha o acesso a uma base de dados. Se a "Equipa A" se atrasar é possível que cause atrasos em todas as outras equipas. Por sua vez, as equipas que sofreram atrasos pelo primeiro projecto participam noutros projectos com outras equipas que acabam por ser influenciadas, e assim sucessivamente para outras equipas e projectos. Criando um efeito "bola de neve".

Seguindo a mesma linha de raciocínio pensemos no esforço total consumido pela DSI e no esforço pedido, através de sucessivos imperativos de negócio (projectos cuja importância faz com que sejam executados o mais rapidamente possível, por vezes interrompendo projectos menos prioritários, consultar *4.1.3 Procedimento operacional para novos projectos*). Mesmo antecipando a possibilidade de existirem projectos deste tipo ao planear o trabalho o volume de projectos abriga as equipas a aproximarem-se muito perto do limite da capacidade[36] (por vezes além dele), uma vez que esta capacidade não é infinita, quando surgem imprevistos, estes reverberam pelos projectos subsequentes, caindo no efeito "*bola de neve"* anteriormente descrito. Portanto a suspeita, traduzida em objectivos de mineração de dados, é que projectos executados em períodos de esforço muito intenso têm maior propensão para o risco.

5.3 Preparação dos dados

Nesta fase é descrita como foi feita a selecção dos dados utilizados. Descreve os critérios utilizados, como por exemplo, a relevância, a qualidade, as técnicas de mineração utilizadas, as restrições, as opiniões dos especialistas de negócio e dos orientadores. É executada a estratégia de migração com os dados já consolidados no *Data Mart*. O processo de migração é

[36] As boas práticas recomendam que a capacidade da equipa nunca seja toda utilizada no planeamento. Deve ser sempre reservada uma percentagem para ligar com imprevistos.

aproveitado para se efectuar o cálculo de alguns dos indicadores. Nesta fase são classificados todos os projectos segundo um modelo de sucesso.

5.3.1 Criação e topologia do *Data Mart (DM)*

Nota prévia: este ponto descreve um repositório designado de *Data Mart*, embora não o seja no sentido clássico [4]. Resulta de um processo ETL[37], serve um Sistema de Suporte à Decisão e contém os conceitos de factos e dimensões mas, apesar de tudo isto, não faz parte de um *Data Warehouse* corporativo maior que satisfaça necessidades de informação departamental à organização. Assim sendo não pode ser considerado um *Data Mart*. A informação lá contida serve um fim muito restrito, alimentar o modelo. Na falta de um termo melhor, optou-se por empregar esta designação.

Como foi mencionado anteriormente, num determinado ponto da evolução do trabalho, surgiu a necessidade de criar um repositório de dados contendo todas as transformações. O DM criado contém o conjunto de todos os dados que queremos analisar, todos os KPI e o *dataset*[38] final utilizado no modelo.

O conjunto de atributos que alimenta o modelo é construído à custa de projecções feitas às diferentes tabelas. A razão pela qual se adoptou por um sistema com sucessivas projecções de dados desnormalizados[39] foi porque um dos algoritmos, utilizados para a criação do modelo, obriga a que o número de atributos seja fixo (consultar 3.5.1.1 *Árvores de decisão, o algoritmo J48*). Esta limitação tem impactos no tratamento dos dados pois ao incluir os atributos das equipas estamos a aumentar o *data set*, vejamos o seguinte exemplo:

- Supondo que dispomos de informação apenas com a caracterização de projectos. A multiplicidade do modelo será:
    - $Projecto_{projecto}$
- Para enriquecermos o modelo com a informação dependente da equipa teremos de incluir essa informação, ou seja, para cada projecto, incluir os atributos de cada equipa participante:
    - $Projecto_{projecto} \bowtie Equipa_{projecto}$.
- Supondo que queremos melhorar ainda mais o nosso modelo e incluir informação acerca dos executantes da actividade. Para essa situação a multiplicidade será:

$$(Projecto_{projecto} \bowtie Equipa_{projecto})_{projecto,equipa} \bowtie Colaborador_{projecto,equipa}$$

[37] ETL, *Extract, Load and Transform*. Conjunto de operações de transformação extracção e carregamento que submetem os dados colocados no *Data Mart*.
[38] Neste contexto o *dataset* (ou *data set*) é um conjunto de registos gerados a partir do DM.
[39] A palavra desnormalizado é empregue no sentido que se utilizaria em teoria dos conjuntos [39].

Projecto	Atributo 1	Atributo 2	Atributo 3	…
Projecto 1	A	A	A	…
Projecto 1	A	A	B	…
Projecto 1	A	B	C	…
Projecto 2	B	C	C	…

Tabela 17 - *Pivot table* de dados

5.3.2 Atributos derivados

Os atributos derivados neste trabalho, denominados KPI [5], são indicadores que ajudam a organização a medir o seu desempenho. A maneira como estes indicadores são compilados tende a variar consoante a indústria analisada.

Além dos especialistas na gestão, do processo, da metodologia, do negócio e a indicação dos orientadores existem repositórios que guiam a criação KPI. Um destes repositórios, dos quais saíram alguns indicadores utilizados neste relatório, é o *KPI Library*[40].

KPI	Descrição
Rácio entre o consumo real consumo previsto por trimestre por projecto	Este indicador visa descobrir paragens, ou reduções significativas, no consumo de projectos. O objectivo é descobrir se o tipo de distribuição de horas afecta a classificação final do projecto $\frac{Consumo_{previsto} - Consumo_{real}}{Consumo_{real}}$
Return On Investment (ROI)	Cálculo do ROI a três anos. $ROI = \frac{Ganho\ previsto - Custo\ previsto}{Custo\ previsto}.3$
Duração planeada do projecto em dias	A data de início subtraída da data do fim
Esforço planeado projecto em dias homem	O esforço total orçamentado para o projecto, por todas as equipas, em horas a dividir pelo número de horas que um recurso trabalha por dia.
Número de equipas	O somatório de equipas participantes no projecto, excluindo as que não passaram ao desenvolvimento.
GPRO participa	Determina se a equipa GPRO participou no projecto na função de GP. Para verificar esta situação é necessário ver se foram imputadas horas na tarefa gestão de projectos
Número de GP	Contagem de gestores para este projecto
Houve mudança de GP	Se a contagem de GP > 1, este indicador verifica se

[40] http://kpilibrary.com/

	ambos trabalharam simultaneamente ou se houve uma mudança.
Grau de paralelismo do projecto	Para todos os dias que este projecto esteve a decorrer qual foi a média de projectos abertos (consultar ponto *5.3.2.4 Grau de paralelismo do projecto*)
Período de esforço intenso	Certos períodos de tempo são considerados críticos. Este atributo verifica se o projecto foi executado num desses períodos.
Desvio de esforço para gestão	Estes indicadores calculam o desvio, para uma fase de um projecto, de esforço. O resultado é relativo $$Desvio_{fase} = \frac{Esforço_{real} - Esforço_{previsto}}{Esforço_{previsto}}$$ (existe um KPI semelhante para a equipa/projecto)
Desvio de esforço para desenvolvimento	
Desvio de esforço para acompanhamento à pré-produção	
Desvio de esforço para suporte a produção	
Desvio de esforço para a certificação	
Desvio de duração para gestão	Estes indicadores calculam o desvio, para uma fase de um projecto, de esforço. O resultado é relativo $$Desvio_{fase} = \frac{Duração_{real} - Duração_{previsto}}{Duração_{previsto}}$$ (existe um KPI semelhante para a equipa/projecto)
Desvio de duração para o desenvolvimento	
Desvio de duração para acompanhamento à pré-produção	
Desvio de duração para suporte à produção	
Desvio de duração para a certificação	
Detalhe do esforço	O esforço para este projecto foi detalhado em tarefas ou só existe um esforço total para a equipa.
Número de requisitos implementados	Contagem do número total de requisitos de negócio implementados.
Esforço da tarefa foi diferente do Esforço do Plano	Este indicador diz-nos se foi indicado algum detalhe ao planear a actividade; ou se a actividade foi orçamentada como um todo e se esse detalhe foi cumprido na execução.
Número de intervenientes na actividade	Contagem do número total de pessoas que participou na actividade
Quantos desvios houve ao	Contagem de desvios reportados ao cliente

acordado no DP	
Principal causador dos desvios	Conta-se o número de origens de cada tipo e indica-se o que tiver a maior contagem. Em caso de empate o indicador é nulo.
Padrão de consumo de esforço	Se houve um consumo ideal de esforço na actividade, consultar *5.3.2.3 Padrão de consumo* de esforço

Tabela 18 - KPI, tabela 1

5.3.2.1 KPI específicos para a análise temporal

Em determinado ponto, influenciado pela possibilidade de criar um modelo que utilizasse séries temporais [27], adoptou-se uma aproximação temporal ao DM. Isto significava que em vez de serem criados indicadores sobre os projectos (o seu resultado final); a sua execução era analisada em pequenos intervalos e os indicadores eram calculados sobre os projectos decorrentes apenas nesses períodos.

Essa aproximação levou à criação de alguns KPI especificamente orientados para executar essa análise temporal.

KPI	Descrição
Data inicial da iteração	Data de início para a qual se referia a análise
Data final da iteração	Data de fim para a qual se referia a análise
Consumo previsto para desenvolvimento	Estes indicadores, para o período de tempo indicado, previam quanto seria consumido. A fórmula que determinava este indicador era: $\frac{Total_{esforço\ Orçamentado}}{Duração_{prevista}}.Intervalo\ dias\ iterado$. Quando a data de início ou final se situa entre os intervalos de iteração é necessário calcular o número de dias exactos até à próxima iteração.
Consumo previsto passagem à produção	
Consumo previsto para acompanhamento à produção	
Consumo real para desenvolvimento	Contagem do esforço realmente consumido no período de tempo analisado
Consumo real para passagem à produção	
Consumo real para o acompanhamento à produção	
Diferença entre o consumo previsto e o realizado para desenvolvimento	A diferença absoluta entre dois valores, para as diferentes fases.
Diferença entre o consumo previsto e o realizado para passagem à produção	

Diferença entre o consumo previsto e o realizado para o acompanhamento à produção	
Consumo acumulado previsto para desenvolvimento	No final do período analisado, portanto na data final da iteração, qual seria o consumo previsto acumulado entre o início previsto e fim da iteração.
Consumo acumulado previsto para a passagem à produção	
Consumo acumulado previsto para o acompanhamento à produção	
Consumo acumulado real para desenvolvimento	No final do período analisado, portanto na data final da iteração, qual foi o consumo real acumulado. Entre o início previsto e fim da iteração.
Consumo acumulado real para a passagem à produção	
Consumo acumulado real para o acompanhamento à produção	

Tabela 19 - KPI, tabela 2

5.3.2.2 KPI classificadores de sucesso do projecto

Os projectos contêm um atributo calculado que os caracteriza quanto ao seu sucesso. A sua classificação tem por base os seguintes KPI:

KPI	Descrição
Análise ao desvio do plano.	Este indicador determina o desvio relativo da duração $x_1 = \frac{(Duração_{real} - Duração_{prevista})}{Duração_{real}}$
Análise ao desvio do esforço.	Este indicador determina o desvio relativo do esforço $x_2 = \frac{(Esforço_{real} - Esforço_{previsto})}{Esforço_{real}}$
Análise do desvio à aceitação	A nota é uma valor de entre [1-5], 5 é a melhor máxima. $x_3 = 1 - \frac{Nota}{5}$
Classificação do projecto	A classificação final é uma média ponderada onde os valores de γ são obtidos pela experimentação e pela opinião dos especialistas. Utiliza-se o módulo dos valores porque tanto o sobreplaneamento como o subplaneamento são igualmente problemáticos para a organização. $x = \frac{\lvert x_1 \rvert . \gamma_1 + \lvert x_2 \rvert . \gamma_2 + \lvert x_3 \rvert . \gamma_3}{3}$

	Este atributo encontra-se detalhado na secção *5.3.2.6 Classificação de projectos*

Tabela 20 - KPI, classificação

5.3.2.3 Padrão de consumo de esforço

Analisemos a fase de desenvolvimento de algumas das equipas. O próximo gráfico mostra, para cada momento no tempo, o esforço total dispendido para uma equipa num dado projecto.

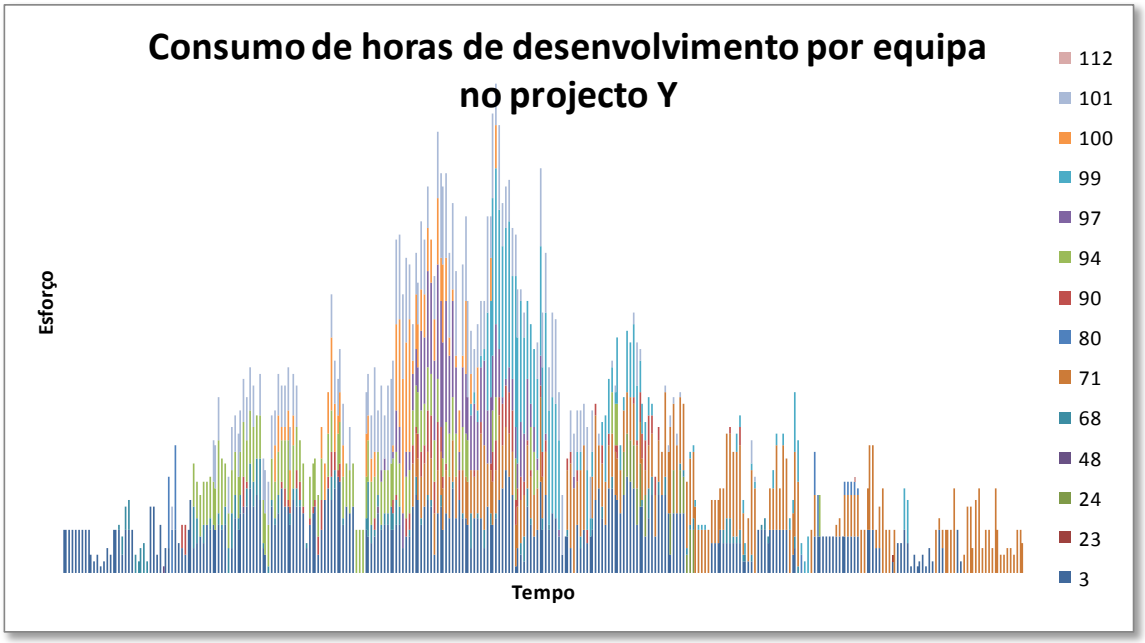


Gráfico 13 – Consumo de esforço ao longo do tempo, cada série (cor) representa uma equipa

São visíveis as semelhanças entre a *Ilustração 3 - O Processo RUP* e o *Gráfico 13*. Este representa o consumo acumulado do esforço de todas as equipas. Como podemos determinar o padrão de consumo do desenvolvimento na DSI?

O estudo da distribuição do consumo do desenvolvimento consistiu na normalização do período de tempo dispendido em todos os projectos (11 intervalos, o menor projecto encontrado) e na média de todo o esforço dispendido. O cálculo consiste em dividir o projecto em 11 períodos de tempo e obter a média de esforço dispendido para esse intervalo. Depois de obtidas as médias por projecto somar esses vectores e calcular novamente a média, seguindo a fórmula:

$$Esforço_{p,t,e} = \frac{1}{Data_{fim(p,t,e)} - Data_{início(p,t,e)}} \sum_{n=Data_{início(p,t,e)}}^{Data_{fim(p,t,e)}} Esforço(n)_{(p,e)}$$

Onde $Esforço_{p,t,e}$ representa o esforço para um período para uma equipa num projecto.

Analisando os dados constatou-se que a distribuição ideal de esforço de desenvolvimento num projecto está muito perto uma distribuição normal com $\sigma^2 = 5, \mu = -0{,}3$ e um *offset* de

$AVG(Esforço_{p,t,e})$. O Gráfico 14 ilustra uma comparação entre os valores reais de consumo de esforço e o consumo de esforço previsto pelo indicador. O Gráfico 15 ilustra um consumo ideal de esforço para uma equipa ao longo do tempo.

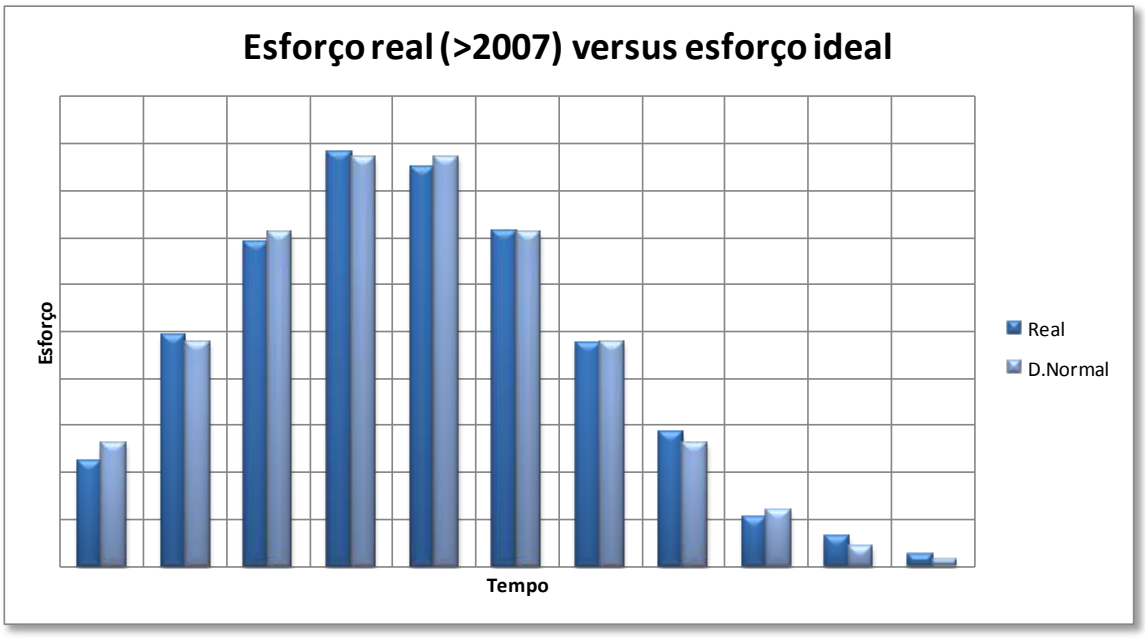


Gráfico 14 - Distribuição real *versus* distribuição normal

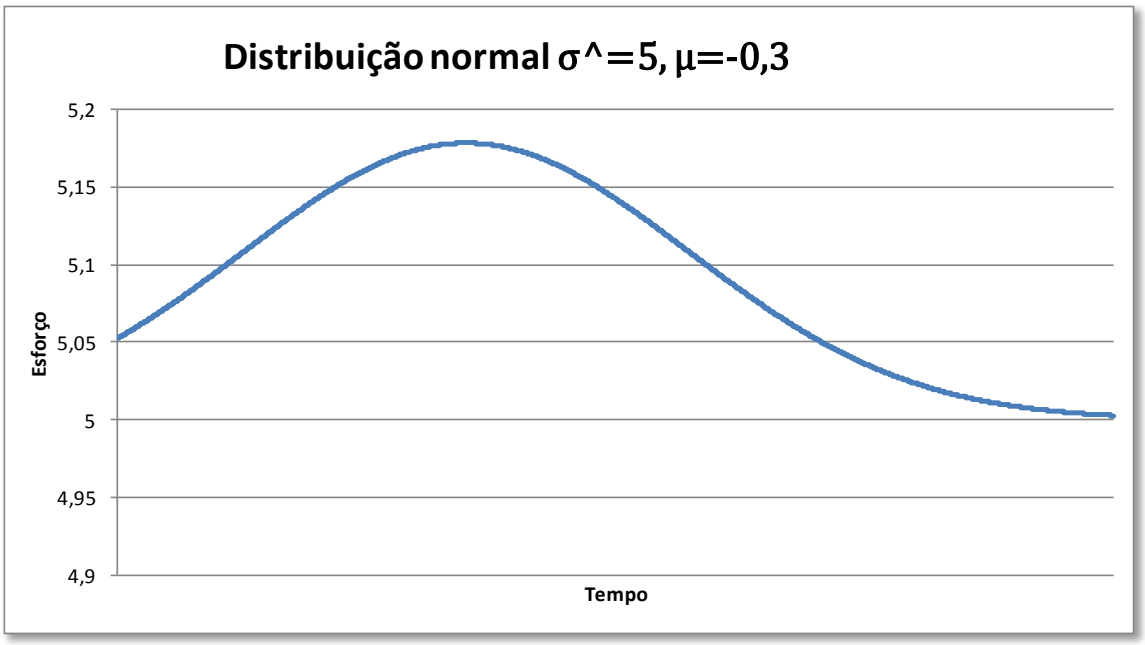


Gráfico 15 - Distribuição normal, neste exemplo o esforço médio é de 5 horas

No *Gráfico 16* analisa-se um caso prático, o consumo de esforço da "equipa X" para no contexto do "projecto Y". Este gráfico dá-nos a variação do consumo ao longo do tempo, nesta situação como foi possível determinar se o padrão de consumo desta equipa?

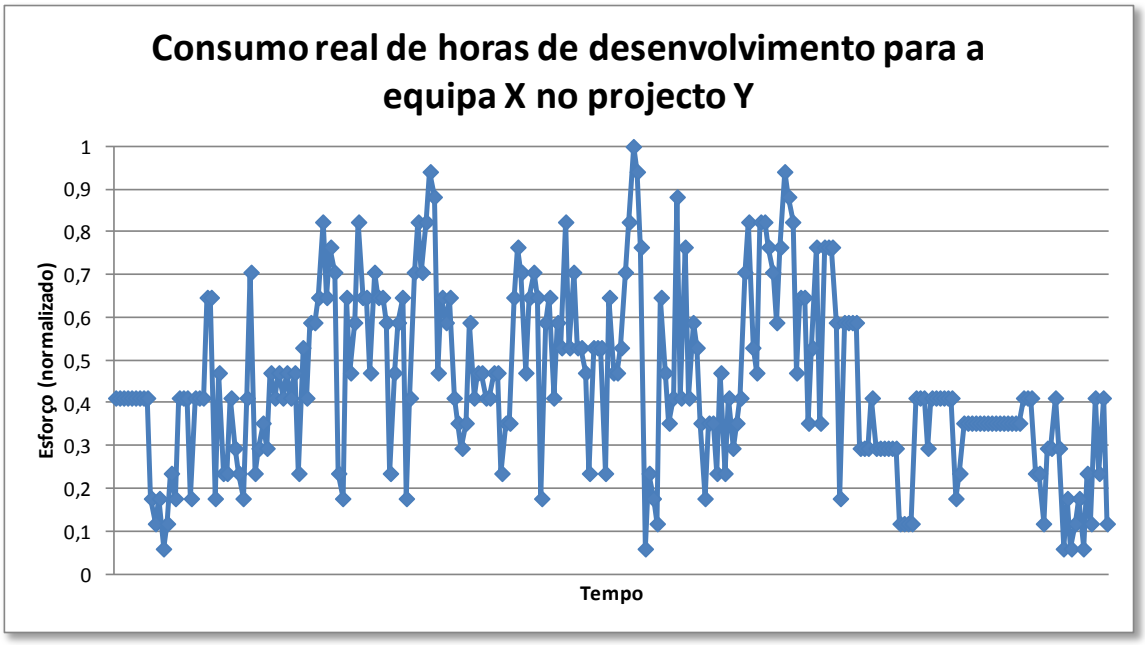


Gráfico 16 – Trabalho realizado de uma equipa num projecto

Ao sobrepor o consumo real com o consumo ideal ficamos com uma ideia mais clara do comportamento da equipa.

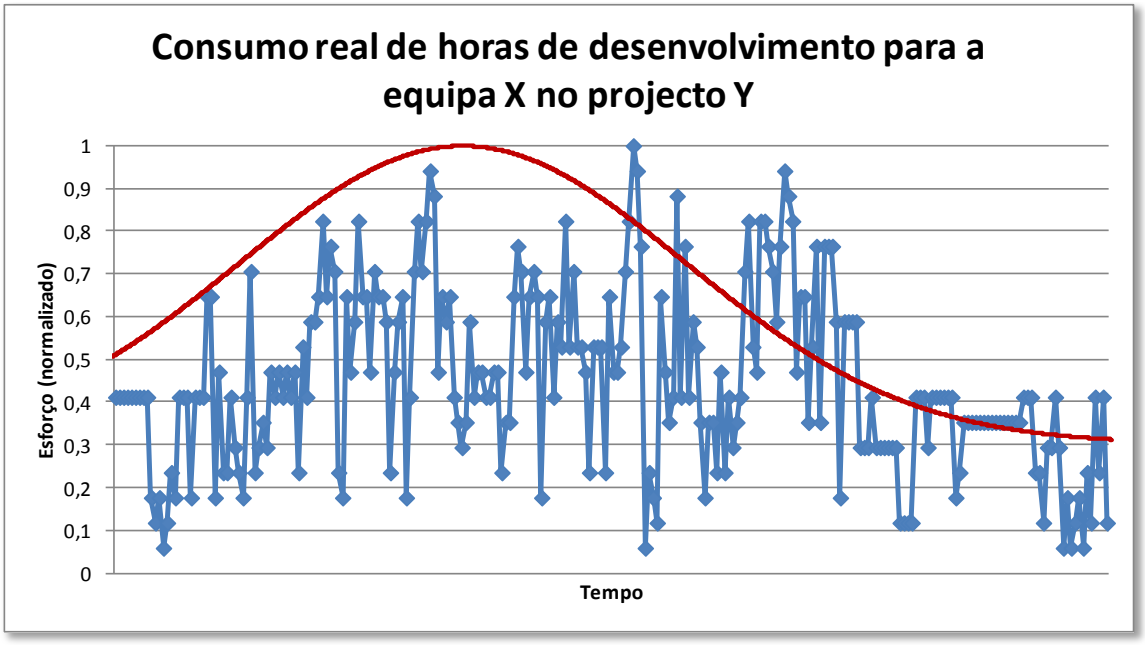


Gráfico 17 - Consumo real versus consumo ideal

Este gráfico revela que, neste caso em particular, a folga que a equipa teve no primeiro terço do projecto teve de ser compensada numa fase mais posterior. Ao analisarmos os dados descobrimos que perto do fim do projecto a equipa ainda estava a executar um esforço bastante significativo com o desenvolvimento. A descompensação inicial foi causada por um outro projecto que já estava a sofrer atrasos (um caso exemplificativo do efeito “bola de neve”).

Como determinamos então se um projecto cumpriu o consumo ideal? Através da análise de equipas com bons comportamentos chegamos à conclusão que 75% dos registos (eixo do tempo) teriam de estar entre ±10% do valor ideal (a distribuição ideal).

O valor ideal é calculado através da fórmula:

$$f(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} . MAX\left(esforco_{equipa}{}_{/projecto}\right)\right).\alpha + AVG\left(esforco_{equipa}{}_{/projecto}\right).\beta,$$

$\sigma^2 = 5, \mu = -0{,}3;\ \alpha = 4, \beta = \frac{1}{3}$,$\alpha, \beta$ foram encontradas por experimentação tendo como objectivo a observação de projectos que correram bem.

Este resultado é aplicado para cada equipa em 10 projectos escolhidos pelos especialistas de negócio. Como resultado final determinou-se que 32,5% das equipas têm, em média, um comportamento que se assemelha ao consumo ideal de esforço.

5.3.2.4 Grau de paralelismo de um projecto

Para calcular o grau de paralelismo de um determinado ponto no tempo traçaríamos uma linha vertical ao longo do eixo dos projectos e encontraríamos o número de intercepções. Para calcular o grau de paralelismo de um projecto, traçamos uma dessas linhas para cada dia de duração do projecto e no final fazemos uma média aritmética.

$$\frac{1}{duração}\sum_{i=1}^{duração}\left[intercepções_{projecto}[i]\right].$$

A *Imagem 4* foi criada através do mapa GANTT via aplicação *Microsoft Office Project 2007* e representa, para o período total que existem registos, o início e o fim (real) dos projectos.

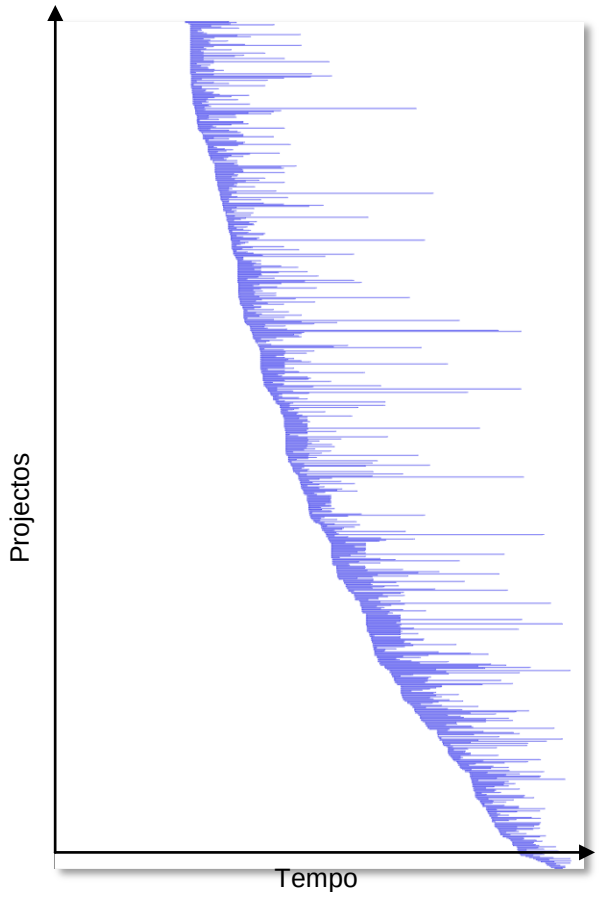


Imagem 4 - Execução de projectos, início e fim

5.3.2.5 Períodos de esforço alto

Relacionado com o grau de paralelismo está o indicador global de esforço. O fundamento por detrás deste indicador é muito simples. Foi medido o grau de esforço exigido às equipas ao longo tempo. Em seguida, com a ajuda dos especialistas, foi estabelecido um limiar para o qual se consideraria um período de intenso desenvolvimento na DSI. O resultado foi a definição do limite para o valor 0,85 do esforço absoluto normalizado para desenvolvimento. Em suma, o indicador diz-nos se um projecto foi executado simultaneamente num destes períodos, isto é: $[Data_{Inicio}, Data_{Fim}] \cap [Data\ Intenso\ Esforco_{Inicio} Data\ Intenso\ Esforco_{Fim}]$. Descobriu-se que existiam 231 períodos de esforço intenso de um total de 985 períodos possíveis. A justificação do valor 85% vem da análise da capacidade. A capacidade total de uma equipa, num determinado período, é calculada pela seguinte formula: $Cap_t = \frac{Recursos_{equipa} * horas\ dia}{dias_t} - T_{t-1} - Risco$, onde T_{t-1} representa o trabalho do período anterior que foi desviado para o ciclo actual e o $Risco$ representa um valor fixo para lidar com imprevistos. A partir do ponto 90% constatou-se que a alocação de recursos para as equipas de desenvolvimento ultrapassava a recomendação de reservar entre 15% e 30% da capacidade para tarefas que não estão relacionadas com o desenvolvimento de novos projectos ($Risco, T_{t-1}$). Considerando 0.90 um valor crítico (apenas tendo ocorrido em 73 períodos), por verificação, constatou-se que para o

valor 0.85, a maior parte das equipas, embora esteja próximo do seu limiar, ainda não o ultrapassou.

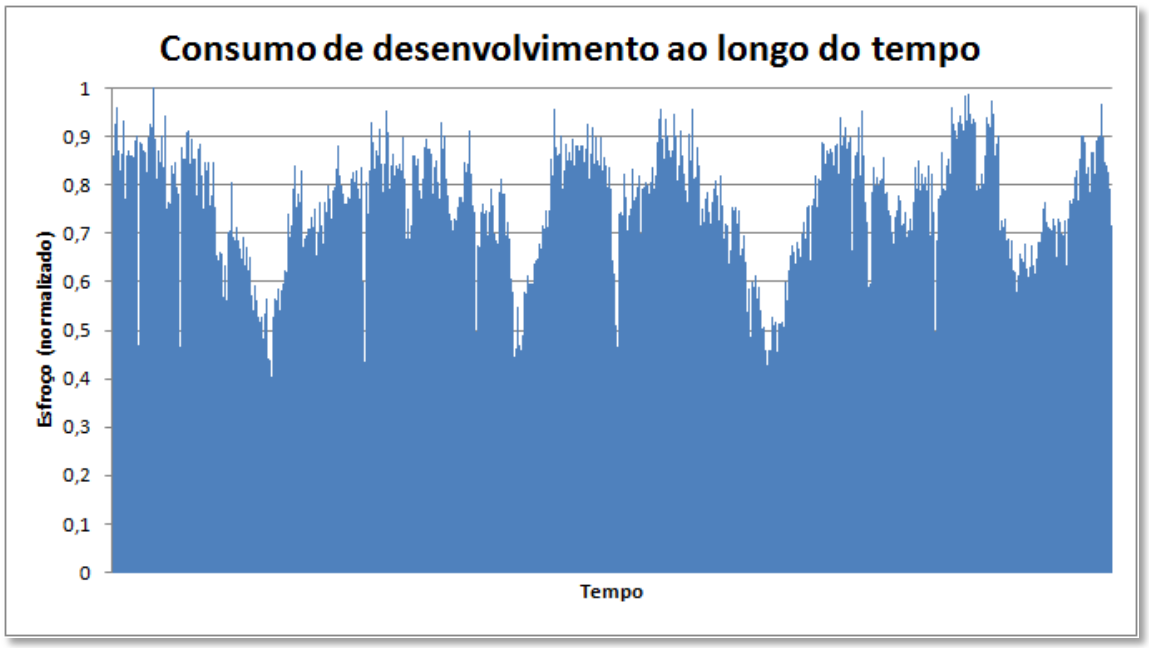


Gráfico 18 - Picos de consumo

5.3.2.6 Classificação de projectos

Utilizando o modelo "*efficiency of project execution*" [15], conseguimos definir uma forma de medir o grau de sucesso, ou insucesso, de um projecto. O modelo baseia-se na conjugação de três dimensões: está dentro do planeado, está dentro do orçamentado e está dentro da especificação?

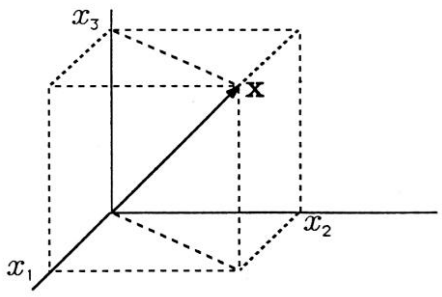


Imagem 5 - As 3 dimensões do sucesso[41]

Assim, utilizamos o KPI $\frac{|x_1|.y_1+|x_2|.y_2+|x_3|.y_3}{3}$ para classificar os projectos já terminados, obtendo a seguinte análise gráfica.

[41] Imagem retirada do sítio: http://cnx.org/content/m21467/latest/

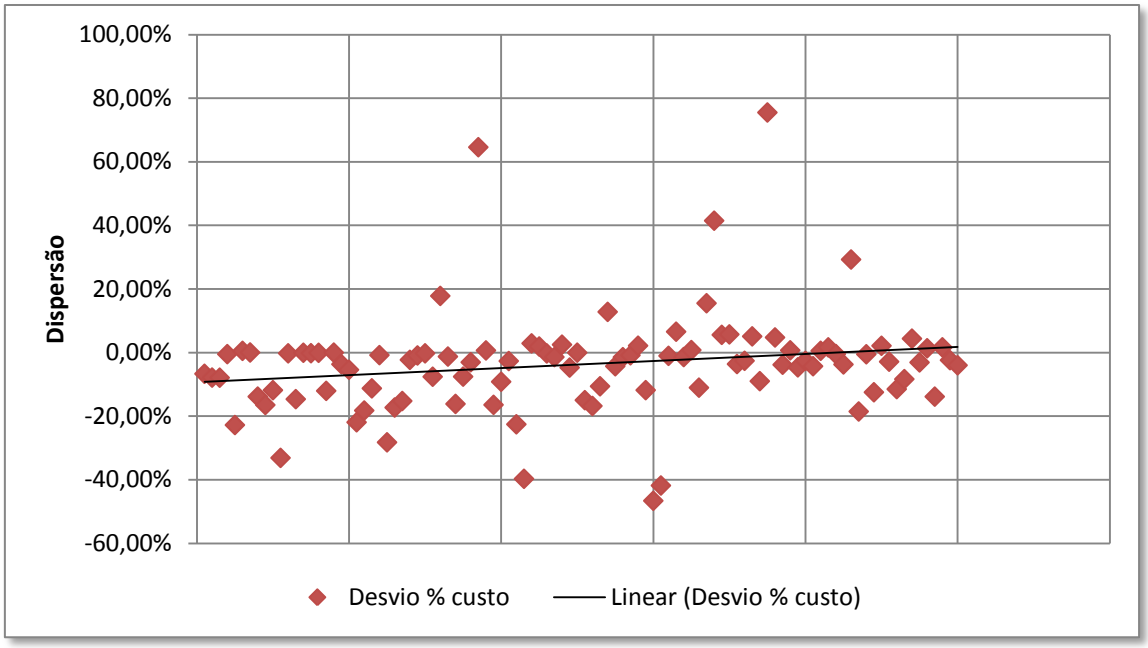


Gráfico 19 - Análise de dispersão ao desvio do esforço (amostragem de 100 projectos)

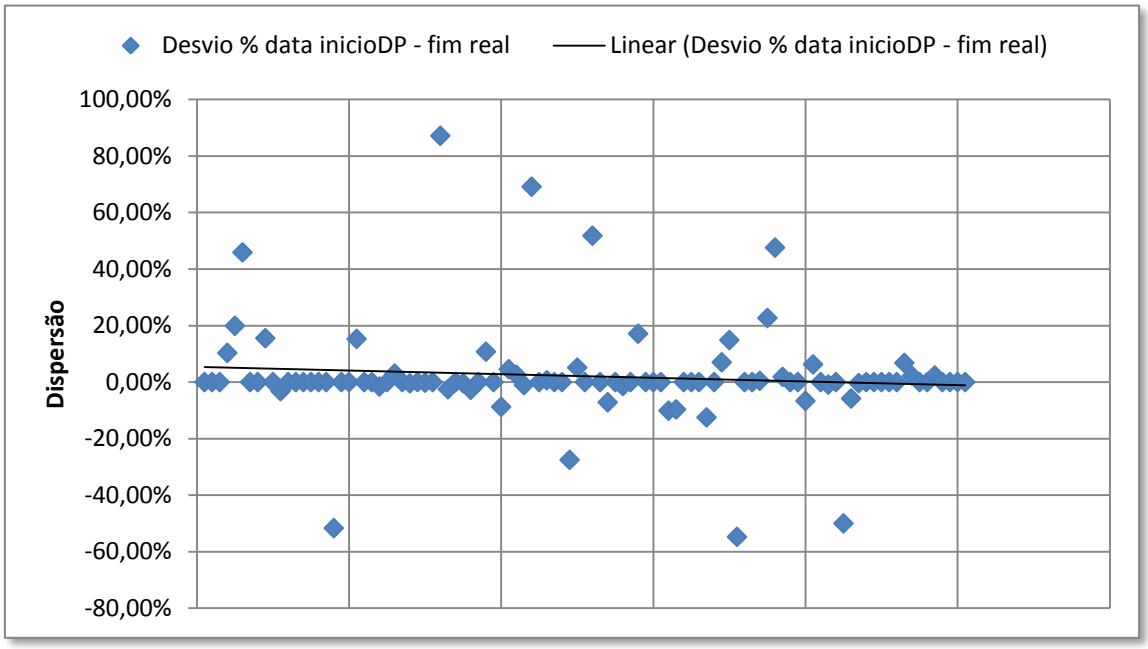


Gráfico 20 - Análise de dispersão do desvio ao plano (amostragem de 100 projectos)

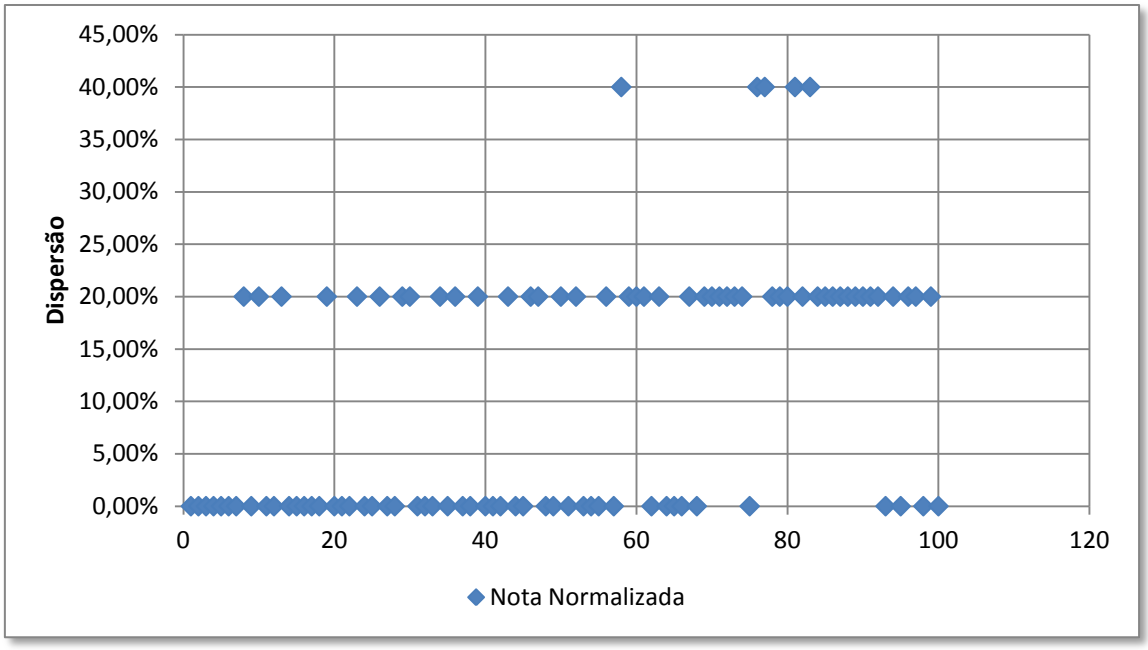


Gráfico 21 - Análise à aceitação (amostragem de 100 projectos)

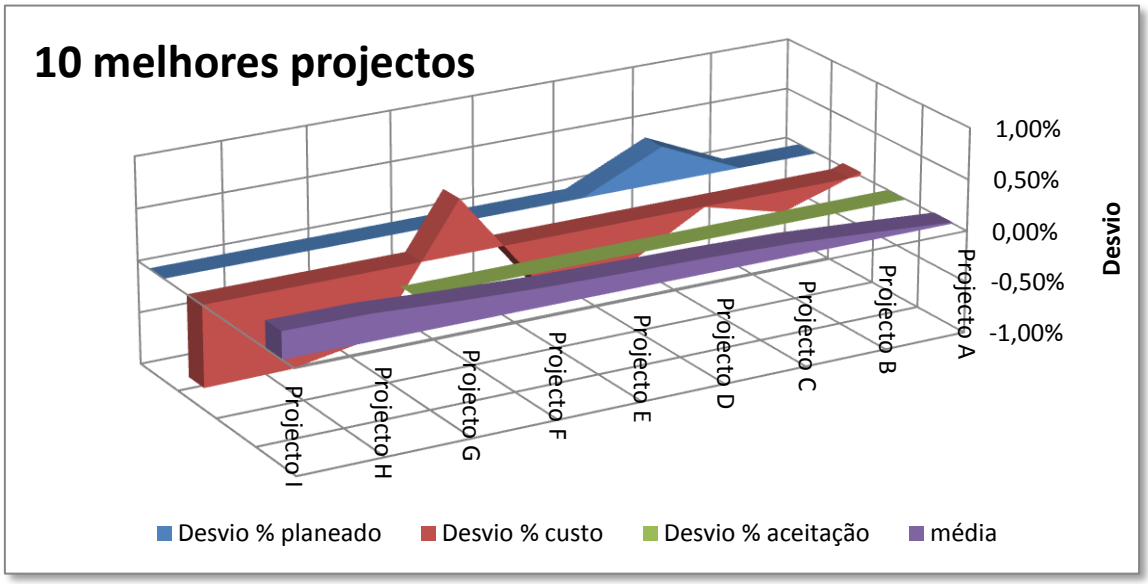


Gráfico 22 - Os 10 melhores projectos

Os valores de $\gamma_1, \gamma_2, \gamma_3$ foram obtidos pelo processo de experimentação e consulta aos especialistas. Os especialistas identificaram dez actividades que correram excepcionalmente bem e outras dez que correm terrivelmente mal. Os valores foram adaptados de modo a que houvesse garantia que essas actividades tivessem as dez melhores, ou piores respectivamente, classificações possíveis.

$\gamma_3 = 0$ *e* $\gamma_1 = \left(\gamma_1 + \frac{1}{6}\right), \gamma_2 = \left(\gamma_2 + \frac{1}{6}\right)$, quando a actividade está isenta de aceitação (consultar *5.2.3*).

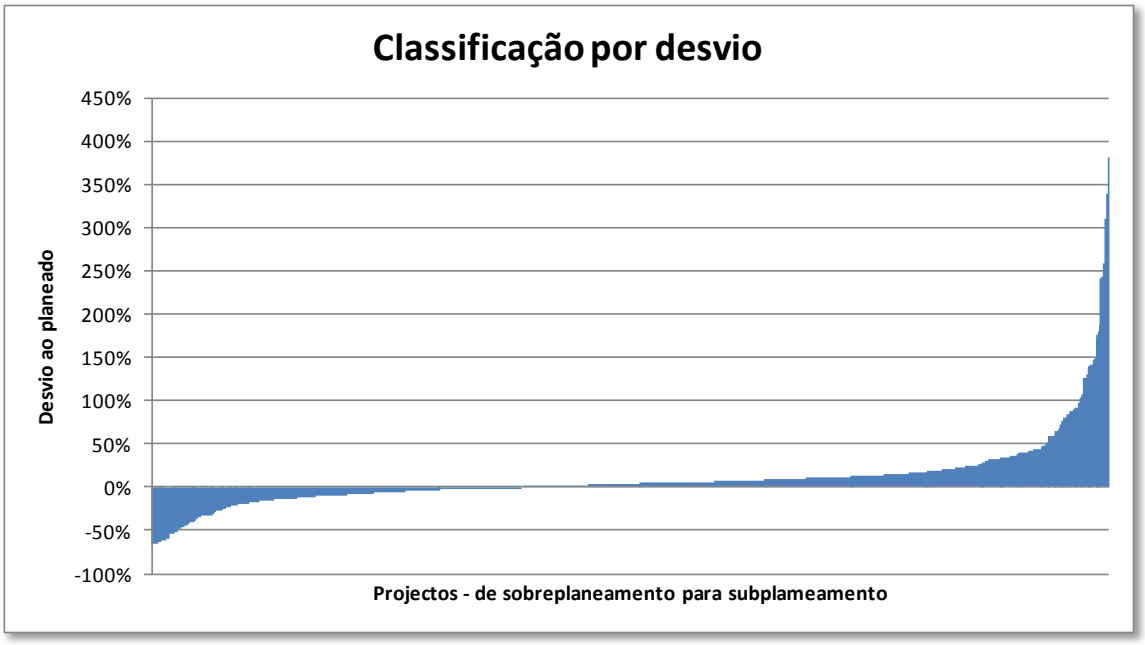


Gráfico 23 - Classificação de projectos

O *Gráfico 23* representa o domínio completo dos dados disponíveis. As margens, do eixo da longitudinal, representam os projectos cujo desfecho foi negativo e a porção central (perto de 0% de desvio) os projectos com um desfecho positivo.

Segundo os especialistas de negócio a classificação deveria estabelecer 3 classes de risco: baixo, médio e alto. Utilizando técnicas de divisão por frequência [28] os projectos foram então classificados pelas classes ilustradas no *Gráfico 24.*

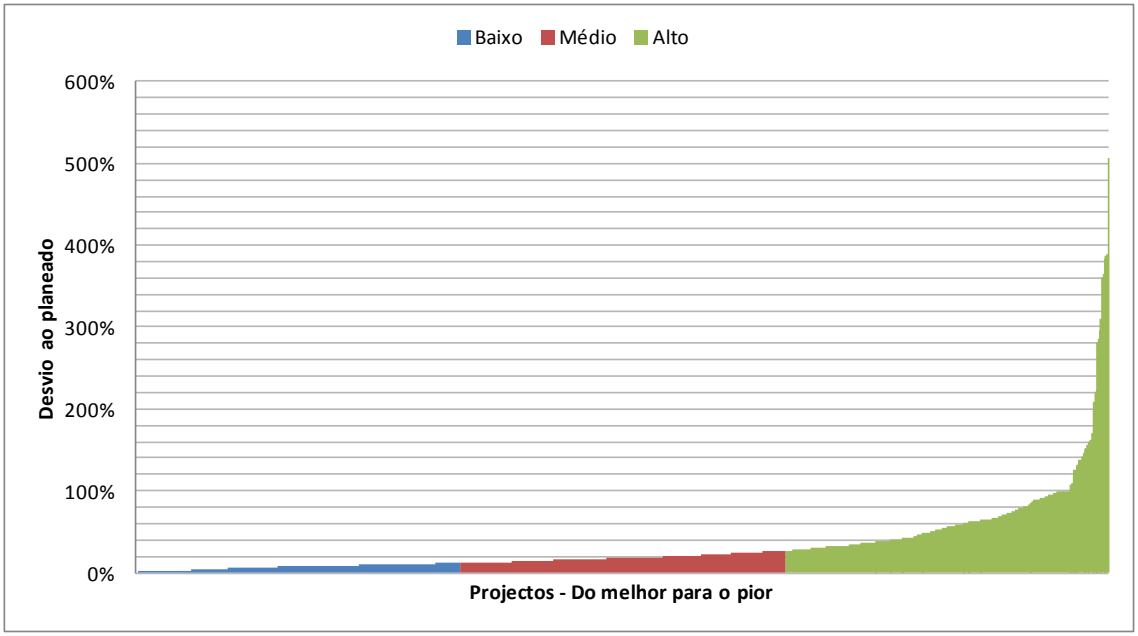


Gráfico 24 - Classificação qualitativa do risco

A classificação ficou então definida pelos intervalos presentes na *Tabela 21*.

	Baixo	Médio	Alto
Intervalo	[0.0, 0.117101]	]0.117101, 0.271231]	]0.271231, +∞[

Tabela 21 - Definição das classes

5.4 Modelação

Nesta fase foram exaustivamente testados todos os atributos e indicadores reunidos até ao momento. Cada atributo, através de um procedimento desenvolvido para o efeito, é avaliado quanto à influência no risco. O resultado final é uma lista dos melhores atributos disponíveis. São ainda explicadas algumas decisões técnicas, nomeadamente, a selecção de um algoritmo em detrimento de outros dois. Explica-se, passo a passo, como foi obtido e testado o modelo, discutem-se as técnicas e as estratégias. Discute-se a ferramenta utilizada na execução.

5.4.1 Weka 3.6

Na execução do trabalho descrito até esta fase foram utilizadas uma miríade de técnicas e ferramentas. Todo este trabalho teve como objectivo reunir a informação num suporte capaz de "alimentar" a ferramenta utilizada na fase de modelação, o Weka.

Weka[42] é o acrónimo de *Waikato Environment for Knowledge Analysis*[43] e está classificado como um software de aprendizagem automática (*machine learning software*) implementado sobre a plataforma Java. [29]

Desenvolvido na Universidade de Waikato, Nova Zelândia. Weka é um software livre disponibilizado sob a Licença Pública Geral[44].

O interface de utilização do Weka, denominado de *weka workbench,* contém uma colecção de ferramentas para a visualização, manipulação e modelação de dados.

O Weka suporta várias tarefas típicas da mineração de dados, tarefas como: pré-processamento, *clustering*, classificação, regressão e visualização de dados. A ferramenta assume que todos os dados estão disponíveis num formato "plano" onde cada "ponto" é descrito por um número fixo de atributos.

[42] Weka é também o nome de um pássaro endémico da Nova Zelândia que não voa.
[43] Ambiente Waikato para a análise de conhecimento
[44] A licença pode ser consultada neste sítio: http://www.gnu.org/copyleft/gpl.html

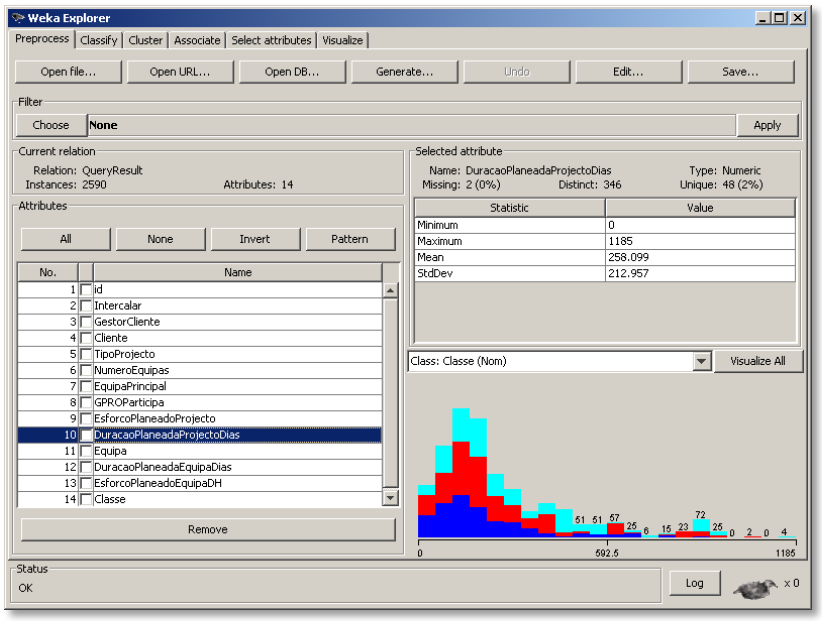


Imagem 6 - O Weka *workbench*

5.4.2 Selecção do algoritmo para a modelação

Seguindo a metodologia, iremos identificar quais são a técnicas apropriadas ao problema do universo de técnicas disponíveis.

Trata-se de um problema de classificação (consultar *Apêndice 1 – Tipificação de problemas na mineração de dados*), existindo para o efeito as seguintes técnicas de resolução disponíveis:

- Análise da discriminante.
- Métodos de regras de indução *(Rule Induction Methods)*.
- Árvore de decisão.
- Redes neurais.
- Vizinhos mais próximo de K (*K Nearest Neighbor* ou KNN).
- Raciocínio baseado em casos (*Case-based Reasoning*).
- Algoritmos genéticos.
- Séries temporais

A selecção das técnicas utilizadas foi o resultado da intercepção de três restrições: os algoritmos de modelação presentes na ferramenta, os algoritmos apropriados para a classificação e a opinião guiada dos orientadores. Esta intercepção encontra-se descrita na *Ilustração 4 - Factores de restrição na escolha do algoritmo*.

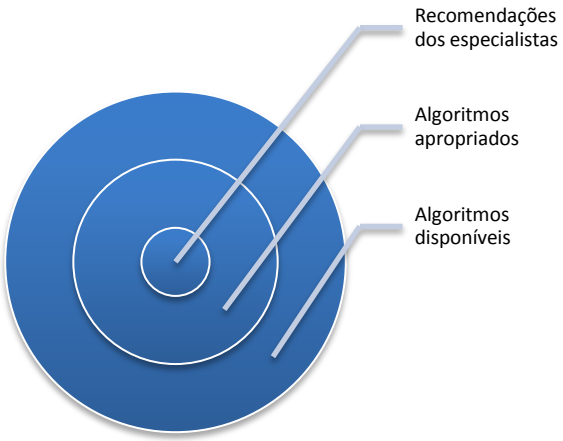


Ilustração 4 - Factores de restrição na escolha do algoritmo

Depois de aplicar a intercepção das restrições sobraram as técnicas de árvores de decisão (algoritmo J48) e de redes neuronais (perceptrão multicamada).

As redes neuronais de perceptrão multicamada foram eliminadas, na fase de escolha de atributos, pelas seguintes razões: os resultados estavam fortemente dependentes da topologia, o tempo necessário para treinar a rede era muito elevado e os resultados eram inferiores aos do J48 (a concorrência).

5.4.3 Selecção dos dados de entrada

Antes de discutir a selecção de dados interessa definir o critério que levou à sua escolha. O processo utilizado foi a medição do ganho de informação de atributos.

$$Ganho(E, A) = Entropia(E) - \sum_{v_t \in V} \frac{|E_{v_t}|}{|E|} Entropia(E_{v_t})$$

$$Entropia(V) = \sum_{i=0}^{n} -p(V_i) log_2\big(p(V_i)\big)$$

Onde V representa o conjunto de exemplos n o número de classes, $p(V_i)$ a proporção de exemplos de V pertencentes à classe [23]. Na sua essência o ganho mede a quantidade de informação que um determinado atributo representa para a classificação final.

O processo de selecção e refinamento dos dados, juntamente com o cálculo de indicadores, representaram uma porção significativa do tempo dispendido neste trabalho. Este processo, está intimamente ligado ao ponto *5.3 Preparação dos dados,* existindo muitas iterações entre os dois. Para conseguir avaliar, de um modo metódico, todos os atributos disponíveis foi desenvolvido um processo para a sistematização da sua análise. Este processo surgiu porque certos atributos e indicadores têm correlações naturais. Como exemplo de dois indicadores correlacionados pode-se apontar o indicador “GP Participa” e “Número de GP” (quando GP Participa é falso, Número de GP é sempre igual a zero). Se estivermos a testar dois grupos

distintos de indicadores onde constem estes exemplos ambos irão gerar bons resultados para o Ganho. Do ponto de vista do projecto, provavelmente, dos dois só no interessará o melhor. Por esse motivo os indicadores foram agrupados por atributo e foi aplicado o seguinte pseudo-algoritmo para sistematizar a sua solução.

1. Escolhem-se atributos/indicadores (A/I) relacionados da lista “não avaliados”. Mede-se o ganho de informação de cada A/I.
    a. Os A/I com nenhum ou baixo ganho de informação são colocados na lista “sem relevância”
    b. Os restantes A/I são colocados na lista “relevantes”
2. Se ainda existem A/I na lista “não avaliados” ir para o ponto 1. Caso contrário prosseguir.
3. Escolhem-se A/I relacionados da lista “sem relevância”. Mede-se o ganho de informação de cada A/I.
    a. Os A/I com nenhum ou baixo ganho de informação são colocados na lista “excluídos”
    b. Os restantes A/I são colocados na lista “relevantes”
4. Se ainda existem A/I na lista “sem relevância” ir para o ponto 3. Caso contrário prosseguir.
5. Escolhem-se A/I lista “relevantes”. Mede-se o ganho de informação de cada A/I.
    a. Os A/I com nenhum ou baixo ganho de informação são colocados na lista “excluídos”
    b. Os restantes A/I são colocados na lista “relevantes”
6. Se existirem atributos na lista “relevantes” ir para 5, senão terminar.

Esta aproximação garante que todos os atributos são testados, pelo menos 2 vezes. E nos atributos seleccionados são testados entre grupos diferentes.

Utilizando a ferramenta Weka 3.6, irá ser demonstrada uma iteração de selecção de dados.

5.4.3.1 Selecção dos dados do Data Mart, demonstração

Os dados são disponibilizados através uma vista no DM, consultar *Imagem 7*

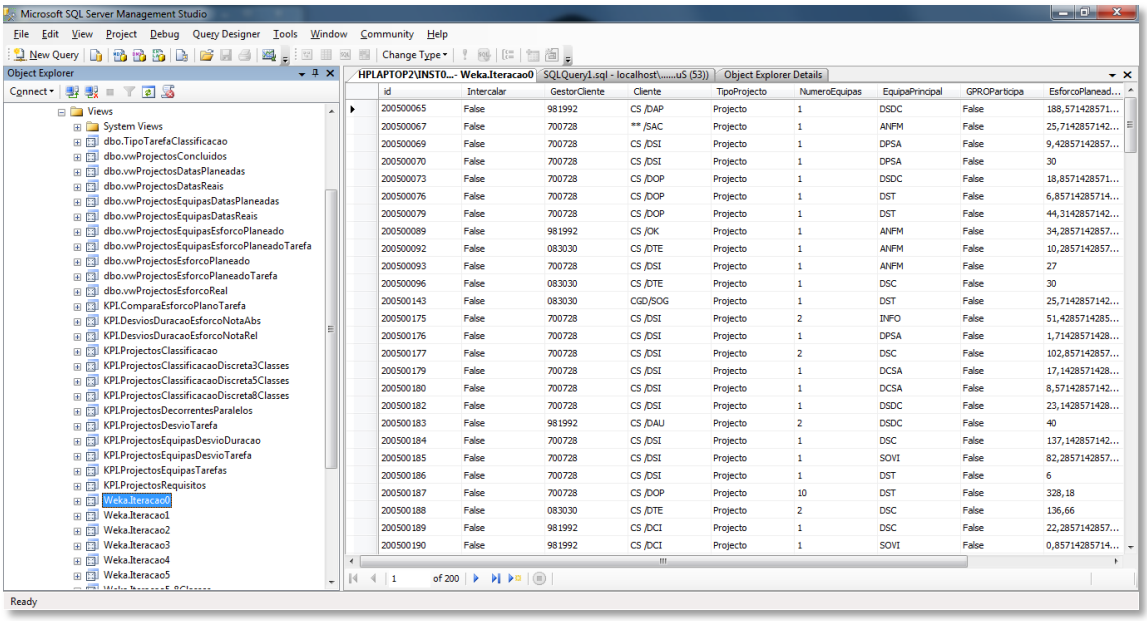


Imagem 7 - SQL Server Management Studio, vista sobre o DM

No exemplo demonstrado a listagem tem o seguinte formato

```
CREATE VIEW [Weka].[Iteracao0] AS
SELECT
          [PC].[id]
         ,[PC].[Intercalar]
         ,[R].Login AS GestorCliente
         ,[PC].[Cliente]
         ,[PC].[TipoProjecto]
         ,[PC].[NumeroEquipas]
         ,[PC].[EquipaPrincipal]
         ,[PC].[GPROParticipa]
         ,[EP].[Esforco] as EsforcoPlaneadoProjectoDH
         ,[DP].[Duracao] as DuracaoPlaneadaProjectoDias
         ,[PD].[Classe]
FROM
      dbo.vwProjectosConcluidos PC
      LEFT JOIN KPI.ProjectosClassificacaoDiscreta3Classes PD ON PC.id
= PD.ID
      LEFT JOIN dbo.vwProjectosEsforcoPlaneado EP ON EP.ProjectoID =
PC.id
      LEFT JOIN dbo.vwProjectosDatasPlaneadas DP ON DP.ProjectoID =
PC.id
      LEFT JOIN R ON R.id = [PC].[GestorCliente]
```

Tabela 22 - Listagem de código

Sem entrar em grandes detalhes sobre a listagem de código é possível observar os mecanismos discutidos anteriormente. Existe uma projecção de dados que representa a classificação (`KPI.ProjectosClassificacaoDiscreta3Classes`), uma contendo a informação do Projecto (`dbo.vwProjectosConcluidos`), outra as datas e assim sucessivamente.

Os dados são adquiridos pelo Weka e transformados num ficheiro .arff. Esta operação é efectuada pelas seguintes razões:

- Isto torna o processamento mais eficiente, por alguma razão a leitura directa do *Data Mart*, utilizando a tecnologia ODBC[45] torna o processo consideravelmente mais lento.
- As operações de transformação do *data set* não são perpetuadas para o DM, residem na memória. Perdem-se quando o programa é fechado.
- O acesso directo via ODBC tem tendência a gerar excepções na aplicação, particularmente quando se processam *data sets* com muitos registos.

O Weka tem um módulo especializado para efectuar esta conversão, o SQLViewer, consultar *Imagem 8*

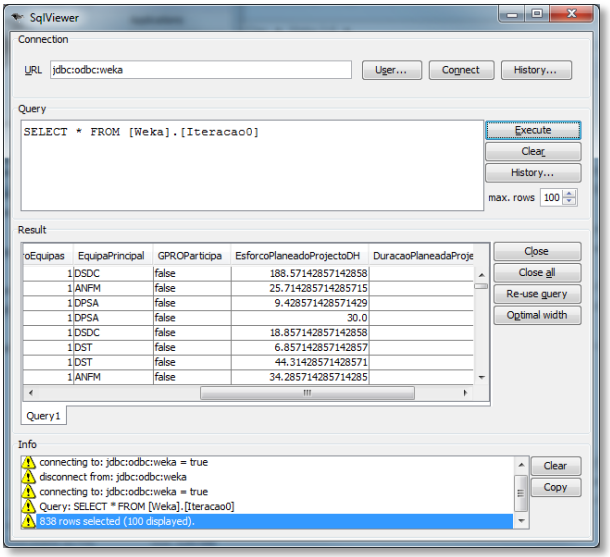


Imagem 8 - Conversão do *data set* para .arff

O ficheiro criado é depois aberto no *Explorer* (consultar *Imagem 9*) e classificado pelo algoritmo *InfoGainAttributeEval*, método *Ranker* (consultar *Imagem 10*).

[45] ODBC - *Open Database Connectivity*, uma tecnologia com o propósito de tornar o acesso à informação independente da linguagem, sistema operativo, aplicação, etc.

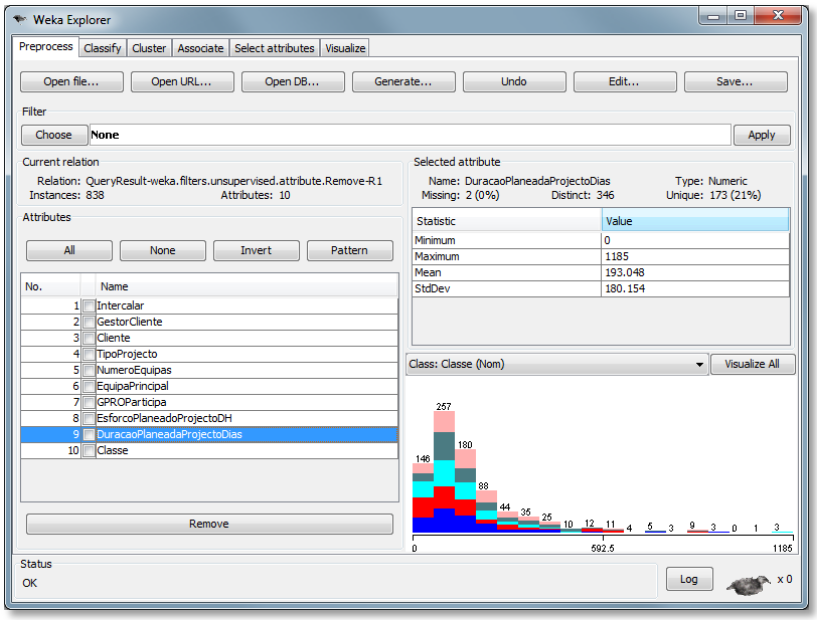


Imagem 9 - Manipulação do .arff

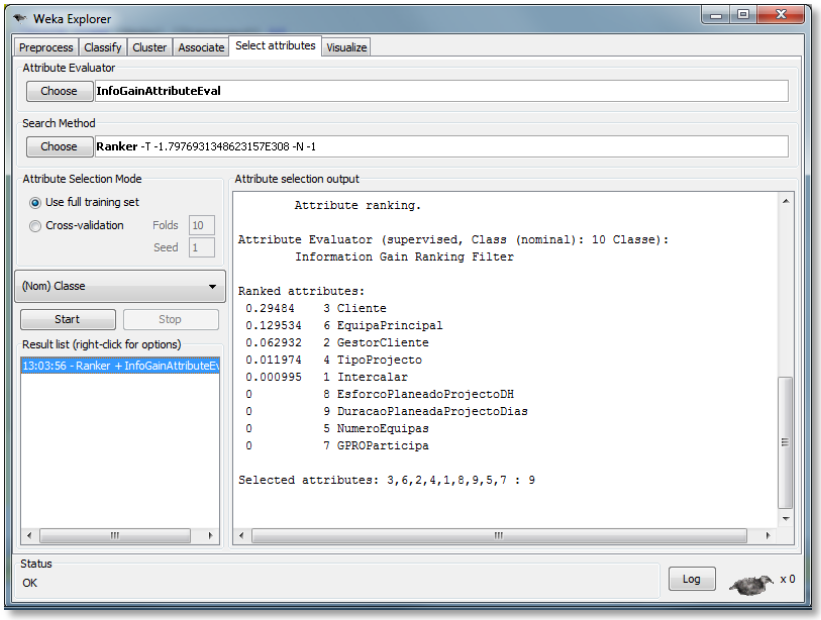


Imagem 10 - Classificação de uma iteração

O resultado da classificação da *Imagem 10* está transcrito na *Tabela 23*

Info Gain Attribute Eval	Ranked attributes: 0.29484 3 Cliente 0.129534 6 EquipaPrincipal 0.062932 2 GestorCliente 0.011974 4 TipoProjecto 0.000995 1 Intercalar 0 8 EsforcoPlaneadoProjectoDH

	0 9 DuracaoPlaneadaProjectoDias 0 5 NumeroEquipas 0 7 GPROParticipa Selected attributes: 3,6,2,4,1,8,9,5,7 : 9

Tabela 23 - Resultado da classificação pelo Ganho

Foi feita a seguinte separação de atributos:

Relevantes	Sem relevância
Cliente	EsforcoPlaneadoProjectoDH
EquipaPrincipal	DuracaoPlaneadaProjectoDias
GestorCliente	NumeroEquipas
TipoProjecto	GPROParticipa
Intercalar	

Tabela 24 - Resultados da classificação

O processo em seguida decorreu para todos os atributos e populou as listas "relevantes" e "sem relevância". Uma vez que é um processo mecânico e extenso não será descrito na íntegra, em vez disso iremos dar um "salto" para a iteração onde avaliamos a lista "sem relevância". A *Tabela 25* mostra-nos os atributos desta lista já classificados pelo Weka.

Info Gain Attribute Eval	Ranked attributes: 0.64683 3 EsforcoPlaneadoProjectoDH 0.30293 9 DuracaoPlaneadaProjectoDias 0.13826 2 NumeroEquipas 0.09551 4 EquipaPrincipal 0.03744 1 Equipa 0.01444 6 DuracaoPlaneadaEquipaDias 0.00686 5 EsforcoPlaneadoEquipaDH 0.00630 7 GPROParticipa 0 8 DesvioAcompPreProducao 0 10 DesvioCertificacao 0 11 DesvioSuporteProducao Selected attributes: 3,9,2,4,1,6,5,7,8,10,11 : 10

Tabela 25 - Classificação da lista "sem relevância"

A *Tabela 26* mostra quais são os atributos permanentemente excluídos (lista "excluídos") e os atributos que são promovidos para a lista de relevantes.

Relevantes	Excluídos
EsforcoPlaneadoProjectoDH	DesvioAcompPreProducao

DuracaoPlaneadaProjectoDias	DesvioCertificacao
NumeroEquipas	DesvioSuporteProducao
EquipaPrincipal	
Equipa	
DuracaoPlaneadaEquipaDias	
EsforcoPlaneadoEquipaDH	
GPROParticipa	

Tabela 26 - Divisão dos atributos

Como foi mencionado anteriormente a selecção de atributos influencia a multiplicidade do *data set*. Vejamos o seguinte exemplo: $[Projecto_{atributos}].[Equipas_{atributos}]$.

Sabendo que existem cerca de 800 projectos analisáveis cada um terá informação, em média, sobre 14 equipas. O conjunto final de dados preparados que foi utilizado para construir os modelos é constituído por, aproximadamente, 115.000 registos. Este valor subiu até ≈500.000, depois de se ter incluído a informação acerca dos executantes na actividade, mas este indicador aumentava o erro do modelo gerado subsequentemente.

Atributos finais, depois da selecção.

Info Gain Attribute Eval	Ranked attributes: 0.1767608 8 EsforcoPlaneadoProjecto 0.1288926 9 DuracaoPlaneadaProjectoDias 0.1130865 12 EsforcoTarefasDiferenteEsforcoPlano 0.1113637 11 EsforcoProjectoTarefa 0.1056253 7 GPROParticipa 0.0622686 14 DuracaoPlaneadaEquipaDias 0.0593552 15 EsforcoPlaneadoEquipaDH 0.0508711 3 Cliente 0.049812 5 NumeroEquipas 0.0235809 6 EquipaPrincipal 0.0234721 2 GestorCliente 0.0226544 17 EsforcoEquipaTarefa 0.0222692 4 TipoProjecto 0.0202582 1 Intercalar 0.0087026 13 Equipa 0.0000694 10 TipoTarefaProjecto 0.0000694 16 TipoTarefaEquipa Selected attributes: 8,9,12,11,7,14,15,3,5,6,2,17,4,1,13,10,16 : 17

Tabela 27 - Os atributos finais seleccionados

5.4.4 Gerar o design de teste

À medida que os atributos foram sendo seleccionados, foram sendo gerados modelos teste para validar o comportamento do modelo após a mudança de indicadores.

5.4.5 Limitações

Devido ao elevado número de atributos e ao número de registos envolvidos a rede neuronal foi considerada inadequada para a resolução deste problema. O tempo médio de construção do modelo é de 58 segundos mas a aprendizagem levava, em média 12 horas. A rede tinha 1 camada escondida com o número de percepções igual ao número de entradas. Todas as entradas estavam ligadas a todas os perceptrões da camada escondida.

Para tentar melhorar o desempenho da rede neuronal foram eliminadas algumas ligações das entradas menos relevantes (com menor ganho de informação) aos perceptrões das camadas escondidas.

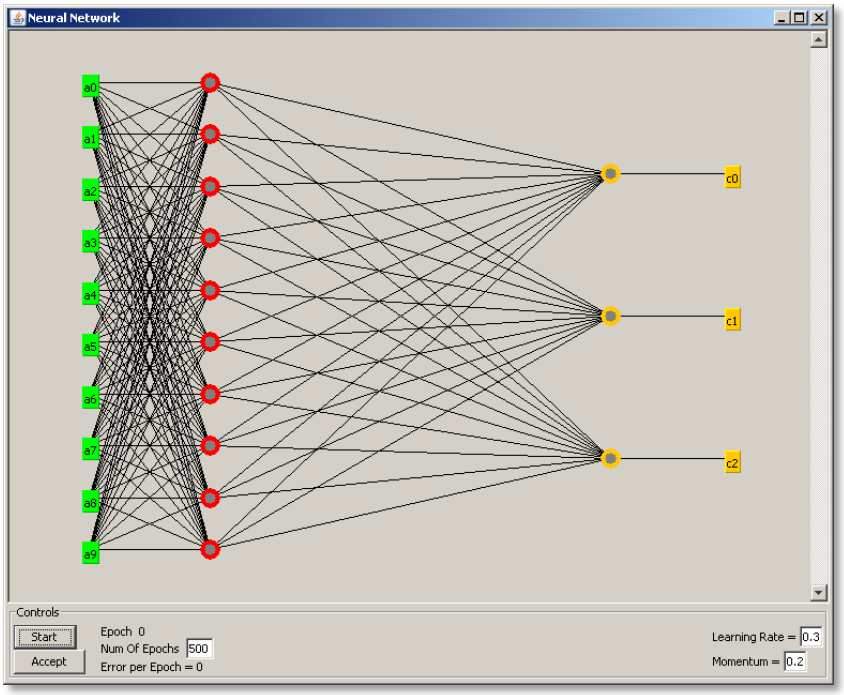


Imagem 11 - Camada escondida da rede neuronal

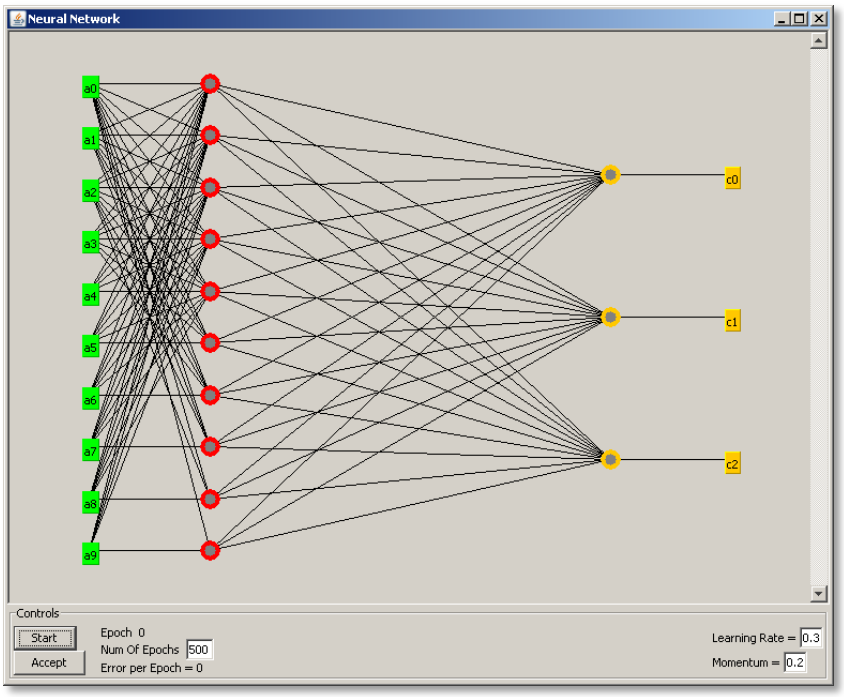


Imagem 12 - Camada escondida da rede neuronal (menos ligações)

Na *Imagem 12* são visíveis menos ligações que na *Imagem 11* - Camada escondida da rede neuronal mas mantêm-se muitas ligações aos atributos com maior ganho de informação. Com esta topologia esperava-se aumentar a eficiência mas não o erro do modelo.

Em última análise as redes neuronais de perceptrão de multi-camada não estavam a produzir resultados tão bons como o J48 e exigiam um esforço exponencialmente maior. Não se pode afirmar, no entanto, que as redes neuronais não são adequadas à resolução deste problema; contudo, depois de inúmeras tentativas, não foi encontrada uma topologia que se equiparasse à eficiência do J48. Por essa razão, as redes neuronais de perceptrão multi-camada foram abandonadas.

5.5 Validação do modelo

Nesta fase é descrita a validação do modelo e algumas das suas características como a precisão e o erro. Além das simulações ao modelo é apresentada uma validação com casos reais.

A árvore de decisão, gerada pelo modelo, contém um total de 1133 folhas e está ilustrada na *Imagem 13*.

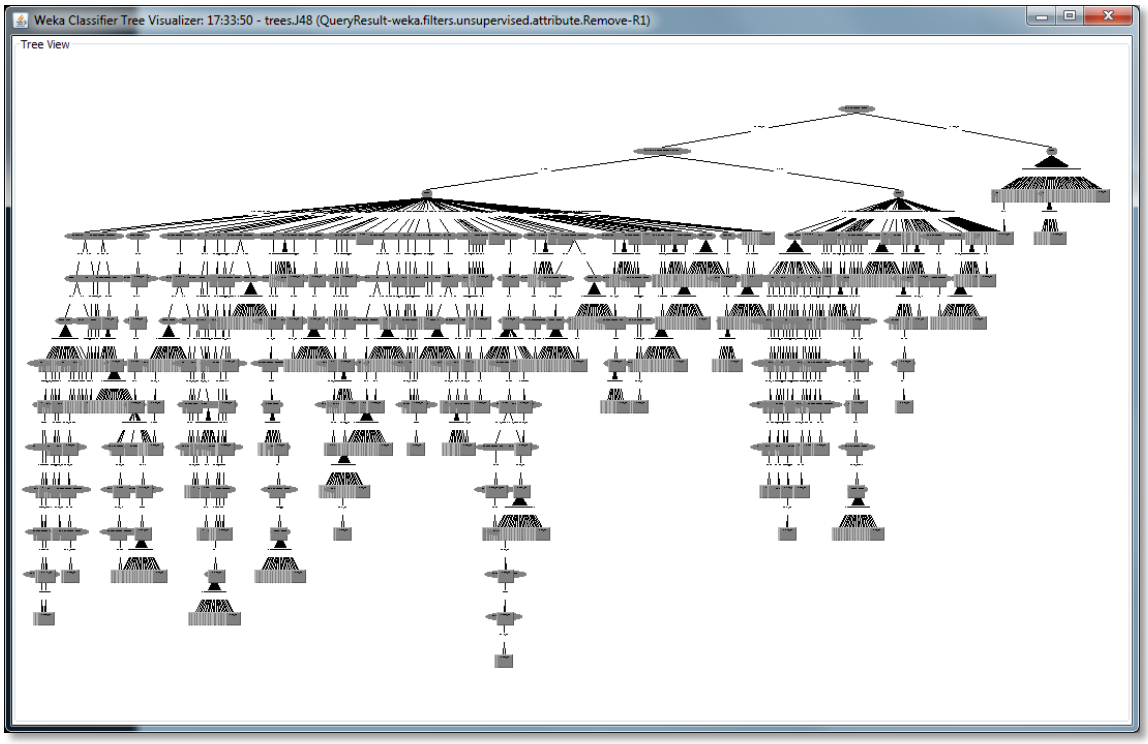


Imagem 13 - A árvore do modelo final

Depois de escolhidos os atributos (consultar *Tabela 27*) temos de avaliar o nosso modelo utilizando a técnica de validação cruzada, descrita no ponto *3.5.4.1 Validação cruzada (Cross-validation)*, dividindo o *data set* em 10 conjuntos.

Na *Tabela 28* onde está visível a matriz de confusão (consultar *3.4.4.2 Matriz de confusão*) ter em conta que A – representa a classe “bom”, B – a classe “Médio” e C- a classe “Mau”.

```
=== Summary ===
Correctly Classified Instances      115286               99.7301 %
Incorrectly Classified Instances       312                0.2699 %
Kappa statistic                          0.9959
Mean absolute error                      0.0018
Root mean squared error                  0.0305
Relative absolute error                  0.4069 %
Root relative squared error              6.4978 %
Total Number of Instances           115598

=== Confusion Matrix ===
     a     b     c   <-- classified as
 32840    99    19 |     a = C
   129 43206    36 |     b = B
    29     0 39240 |     c = A
```

Tabela 28 - O resultado da classificação final

Ao analisarmos o resultado de validação do modelo podemos constatar que:

- Entre as 115.598 instâncias 115.286 foram correctamente classificadas, uma *Accuracy* de AC=99,7%.
- Entre as 115.598 instâncias 321 foram incorrectamente classificadas, uma *Precision* de P=99,8%

Ao analisarmos os resultados temos de ter em conta o formato do *data set*; cada instância tem vários classificadores, com uma granularidade que envolve a equipa sendo que o objectivo final é classificar o projecto. Depois da classificação de novos projectos ser efectuada depende do GP analisar os dados. Vejamos o seguinte exemplo:

Projecto	Equipa	Previsão	Comentário
Projecto X	Equipa 1	A	
Projecto X	Equipa 2	A	
Projecto X	Equipa 3	C	Segundo o modelo esta equipa apresenta um risco maior que as demais.

Tabela 29 - Análise do resultado

Cabe ao Gestor de Projecto tomar algum tipo de acção para mitigar o risco gerado pela participação da equipa 3. A título de exemplo vamos supor que a equipa manteve o esforço mas aumentou a duração. O projecto foi novamente submetido à classificação:

Projecto	Equipa	Previsão	Comentário
Projecto X	Equipa 1	A	
Projecto X	Equipa 2	A	
Projecto X	Equipa 3	B	Ainda apresenta um risco superior

Tabela 30 - Nova classificação

O Gestor de Projecto (GP) pode agora decidir se pode conviver com o risco gerado pela equipa 3 ou se continua a alterar as características do projecto até obter uma classe de risco confortável.

5.5.1 Casos reais

A comparação com os projectos, que foram entretanto finalizados, foi igualmente precisa. Para classificar os projectos actuais, que não contavam no conjunto de treino do modelo, foram executados os seguintes passos.

1) Criar um ficheiro com os atributos dos projectos que pretendemos classificar, o atributo classe fica representado por um "?", visto ser o atributo que esperamos classificar. O formato do ficheiro pode ser consultado na *Tabela 31.*

```
@attribute Intercalar {false,true}
@attribute GestorCliente  {700728,930153,982036,...}
@attribute Cliente {...}
@attribute TipoProjecto {Projecto,Evolutivo}
@attribute NumeroEquipas numeric
@attribute EquipaPrincipal {...}
```

```
@attribute GPROParticipa numeric
@attribute EsforcoPlaneadoProjecto numeric
@attribute DuracaoPlaneadaProjectoDias numeric
@attribute TipoTarefaProjecto {'5 Acompanham em Pré-Produção','2 DP / Análise Funcional','7 Certificação','4 Desenvolvimento','6 Suporte à Produção','1 Gestão','GP: Início & Organização','GP: Execução, Monitorização e Controlo','GP: Encerramento','Manutenção Aplicativo',Infraestrutura}
@attribute EsforcoProjectoTarefa numeric
@attribute EsforcoTarefasDiferenteEsforcoPlano {false,true}
@attribute Equipa {...}
@attribute DuracaoPlaneadaEquipaDias numeric
@attribute EsforcoPlaneadoEquipaDH numeric
@attribute TipoTarefaEquipa {'2 DP / Análise Funcional','GP: Execução, Monitorização e Controlo','1 Gestão','6 Suporte à Produção','4 Desenvolvimento','GP: Início & Organização','7 Certificação','5 Acompanham em Pré-Produção','GP: Encerramento','Manutenção Aplicativo',Infraestrutura}
@attribute EsforcoEquipaTarefa numeric
@attribute Classe {C,B,A}

@data
false,981992,'A',Projecto,8,'B',0,127.76,119,'5 Acompanham em Pré-Produção',17.67,false,'F',119,4.285714,'7 Certificação',0,?
```

Nota: alguns atributos foram ofuscados. Classificação de 1 exemplo.

Tabela 31 - Ficheiro para a classificação de um projecto

2) O Weka dispõem de um modo consola, é particularmente útil para executar operações em lote ou aceder a funcionalidades disponibilizadas pela API sem ter de passar pelo interface gráfico. A consola do Weka denomina-se “SimpleCli”. Abrir a aplicação “SimpleCli”, visível na *Imagem 14*

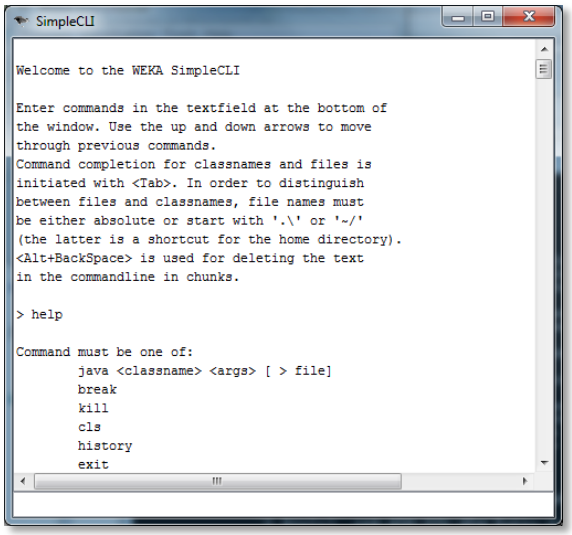


Imagem 14 - SimpleCli do weka

3) E executar o comando:

- `Java weka.classifiers.trees.J48 -p 18 -l c:\J48_2.model -T c:\test.arff`

O comando "`weka.classifiers.trees.J48`" serve para aceder à API do algoritmo J48. O parâmetro "–p 18" significa que queremos prever o 18º parâmetro e o "-l" indica o ficheiro do modelo. Neste ficheiro está definida a árvore de decisão. O parâmetro "-T" contém o ficheiro a classificar.

O resultado é uma consola contendo as previsões, consultar *Imagem 15*

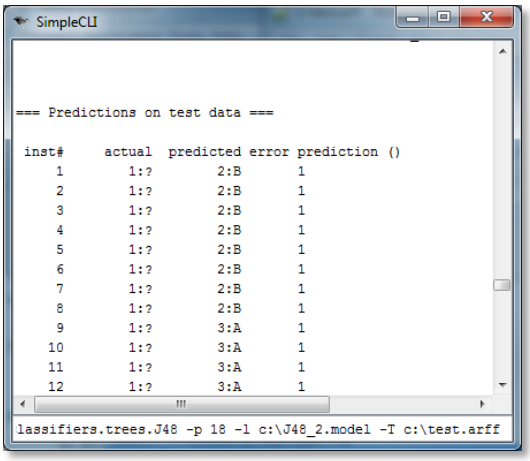


Imagem 15 - Previsões de classe

Foram testadas 115 actividades que entretanto foram encerradas.

A matriz de confusão real para os dados reais foi:

		Classe prevista		
		Bom	Médio	Mau
Classe real	Bom	33	0	0
	Médio	2	41	0
	Mau	0	0	39

Tabela 32 - Matriz de confusão dos resultados reais

No exemplo real o erro é de 1,77%. Um pouco mais alto que o modelo (0,2699%) mas mesmo assim é um valor surpreendentemente bom.

- Entre as 115 instâncias 113 foram correctamente classificadas, uma *Accuracy* de AC=98,2%.
- Entre as 115 instâncias 2 foram incorrectamente classificadas, uma *Precision* de P=98,2%

6. Conclusões

Este projecto focou-se na pesquisa de modelos e estratégias de mineração de dados para servirem de base a um Sistema de Apoio à Decisão para um contexto muito específico. A solução demonstrou que é possível classificar novos projectos quanto ao seu grau de risco.

Os resultados finais foram, surpreendentemente, bons quanto ao seu grau de precisão e baixo erro. Existe um verdadeiro potencial nas descobertas feitas ao nível financeiro uma vez que, baseado na análise feita no ponto *4.5.1 Projectos*, tende existir um desvio entre o esforço planeado e o esforço real. Uma vez que as situações problemáticas estão identificadas, está nas mãos do Gestor de Projecto fazer um acompanhamento de proximidade para antever os riscos. Neste contexto o modelo serve de guia ao Gestor de Projecto para identificar alguns dos "causadores" de risco e das suas causas prováveis.

Existe um grande potencial ao nível da tomada de decisão por parte do cliente. Este é, talvez, o maior interessado em saber se a organização consegue entregar aquilo que prometeu. Da sua perspectiva, o modelo responde à questão: "qual o grau de fiabilidade dos prazos e valores que me foram apresentados?". Isto tem um grande impacto do lado do cliente porque, além de gerir as suas expectativas, permite-lhe tomar decisões estratégicas sobre a sua área de negócio, relativizando a importância dos projectos TI.

Embora a princípio recaíssem alguma suspeitas sobre indicadores complexos e fossem feitas uma série de suposições quanto aos padrões de comportamento dos executantes, no decorrer deste trabalho, o resultado mostrou-nos que, para este modelo de classificação de sucesso, eles não contribuíam para o risco. No ponto *5.2.4 Formulação de suposições*, foram discutidos e calculados alguns indicadores cujo ganho acabaria por ser desprezável. São os casos dos indicadores de *padrões consumo de esforço* e o *grau de paralelismo do projecto*. Os indicadores com um maior contributo para o ganho (consultar *5.4.3 Selecção dos dados de entrada*) são aqueles que vêm descritos nos livros de gestão de projectos (esforço, duração, envolvimento do cliente, outros [1]…). De certo modo estes indicadores, assinalados pelo Gestores de Projecto no início do trabalho, validam o modelo obtido.

Este tipo de solução, pela maneira como foi realizada, não poderia ser generalizado para outras organizações, é muito específica nos objectivos que alcançou. Pode, no entanto, ser um manual exemplificativo para que outros alcancem os mesmos objectivos noutros contextos. Este trabalho é um ponto de partida para alguém com os mesmos objectivos, ou seja a descrição a um modelo de classificação, a caracterização das instâncias, a obtenção modelo de mineração e a classificação de risco.

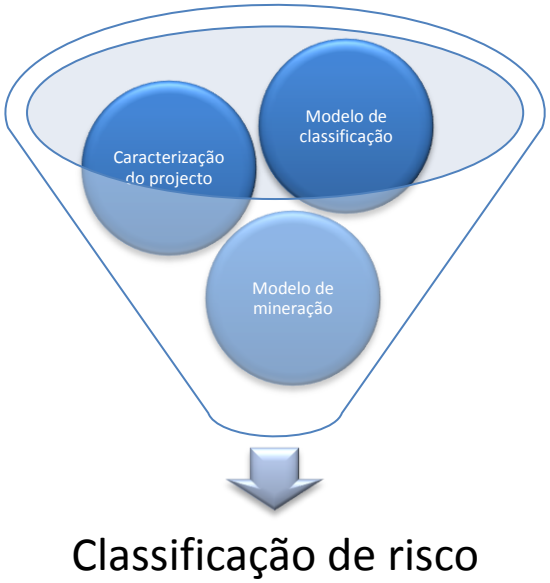


Imagem 16 - Visão geral, *inputs* e *outputs*

6.1 Trabalho futuro

O trabalho futuro a desenvolver apresenta-se dividido em duas vertentes: continuação do trabalho já realizado e exploração de novas vertentes dado o conhecimento já adquirido sobre o domínio do problema.

Na continuação do trabalho desenvolvido, seria necessário a refinar o modelo e efectuar mais testes para provar a sua precisão. Criar novos indicadores que possam acompanhar as mudanças organizacionais e alimentar o modelo com mais dados. Recolher mais informação e tornar o modelo de classificação mais completo de modo a abranger dimensões como "benefícios quantitativos" ou "custos de oportunidade" de modo a espelhar melhor o conceito de sucesso de projecto.

Um dos problemas que existe na indústria é a subjectividade do modelo de sucesso de um projecto. Como se pode medir quantitativamente algo que é fundamentalmente subjectivo. Dependerá apenas do orçamento, do cumprimento do plano e da satisfação do cliente? Ou o sucesso de um projecto também depende da sua robustez, do seu tempo de vida, da sua usabilidade e de outras características que associamos à qualidade dos sistemas TI? Empiricamente sabemos que a resposta a estas questões é "sim" mas põe-se o problema da sua medição. Um trabalho futuro poderá ser a análise e recolha de atributos que possam contribuir para a refinação de um modelo de sucesso, que englobe mais características. Se tal modelo for desenvolvido isto criará a necessidade natural de calcular mais, e melhores, indicadores recomeçando todo o processo.

Um problema que se levanta para o futuro é a questão: "seria possível generalizar a quantificação do risco para outras organizações?". Para realizar este estudo seria necessário recolher os atributos que caracterizam os projectos de outras empresas de TI. Analisar a sua

metodologia de desenvolvimento e descobrir as características comuns influenciadoras do risco. Era interessante saber que estratégia seria adoptada; como poderia o modelo de sucesso de um projecto ser generalizado ou como seriam respondidas as questões acerca dos indicadores dos projectos.

Apêndices

Apêndice 1 – Tipificação de problemas na mineração de dados

Normalmente, o projecto de mineração de dados é composto por uma combinação de diferentes tipos de problema.

6.2 Descrição e sumarização dos dados

A descrição e sumarização de dados visam reunir as características dos dados elementares ou agregados. O resultado final é uma descrição geral da estrutura dos dados. Por vezes a descrição e sumarização podem ser, por si só, o objectivo de um projecto de mineração de dados. O cliente pode estar interessado em saber, por exemplo, o volume de negócios de todos os estabelecimentos discriminados por categoria de produto ou a comparação da situação actual com a de um período anterior. Este tipo de problema está no extremo inferior da escala de complexidade dos problemas de m

ineração de dados.

No entanto, em quase todos os projectos de mineração de dados, a descrição e sumarização de dados são um objectivo secundário dos estágios iniciais. No princípio de um processo de mineração de dados, é possível que não se conheçam nem os objectivos específicos da análise nem a natureza precisa dos dados.

A análise exploratória inicial de dados pode ajudar a compreender a sua natureza e formular potenciais hipóteses acerca de padrões ocultos. Técnicas de estatística descritiva podem fornecer informações ocultas. A distribuição de idade dos clientes e suas áreas de residência, por exemplo, oferecem indicações sobre o grupo a que o consumidor pertence e, eventualmente, como serão tratados pelas estratégias de marketing.

A descrição e sumarização de dados ocorrem normalmente em paralelo com outros tipos de problema de mineração. Por exemplo, a descrição de dados pode levar à postulação de hipóteses interessantes, uma vez que a segmentação de dados é identificada e definida. É aconselhável realizar a descrição e sumarização dos dados antes de qualquer outro tipo de problema de mineração.

Muitos sistemas de informação, ferramentas para análise estatística de dados, sistemas OLAP[46] e sistemas EIS[47] podem executar a função de descrição e sumarização de dados; mas

[46] OLAP - *Online Analytical Processing* - Uma classe de ferramentas utilizada para visualizar, de um ponto de vista mais aproximado ao negócio, a informação [25].

estas ferramentas não costumam fornecer quaisquer métodos para realizar as modelagens mais avançadas.

6.3 Segmentação

A segmentação de dados tem como objectivo a sua separação em subgrupos ou classes. Um subgrupo ou classe partilha características comuns.

Ao analisar a lista de compras de um cliente, por exemplo, podem definir-se segmentos de consumidor, de acordo com os itens que elas contêm. A segmentação pode ser realizada manualmente ou automaticamente. O analista pode assumir que existem certos subgrupos com base nos resultados da descrição e sumarização de dados ou baseado no conhecimento prévio proveniente do negócio. Existem também técnicas de agrupamento automático que podem detectar estruturas.

A segmentação de dados pode ser, por si só, um objectivo de um projecto de mineração de dados. Contudo, muitas vezes, a segmentação é um passo no sentido de resolver outro tipo de problema.

O objectivo pode consistir em reduzir o tamanho dos dados para um valor aceitável, ou encontrar subconjuntos de dados homogéneos que são mais fáceis de analisar. Normalmente, quando lidamos com grandes conjuntos de dados, várias influências sobrepostas podem obscurecer os padrões interessantes. A aplicação da segmentação de dados adequada torna a tarefa mais fácil.

A análise de dependências entre os milhões de itens nas compras de consumidores é muito difícil. É muito mais fácil (e mais significativo) identificar dependências entre segmentos: que segmento compra o quê e quando.

Técnicas apropriadas a aplicar:

- Técnicas de *Clustering*.
- Redes neurais.
- Técnicas de visualização.

6.4 Descrições do conceito

A descrição do conceito visa uma descrição compreensível de conceitos ou classes. O objectivo não é desenvolver modelos completos de alta precisão de previsão, mas para obter conhecimento acerca da classe.

Por exemplo, uma empresa pode estar interessada em saber mais acerca dos seus clientes leais e desleais. A partir da descrição destes conceito (clientes leais e desleais), a empresa

[47] *EIS - Executive Information System* - Uma classe de ferramentas que visa possibilitar tomadas de decisão por parte da gestão baseadas na exploração e comparação de informação [26]

pode inferir o que poderia ser feito para manter os clientes leais ou para transformar clientes desleais em clientes leais.

A técnica de descrição do conceito tem uma ligação estreita com a segmentação e com a classificação.

A segmentação pode levar a uma enumeração de objectos pertencentes a um conceito ou classe sem qualquer descrição compreensível; normalmente, é feita antes da descrição do conceito.

Algumas técnicas, com por exemplo técnicas de *conceptual clustering* [48], executam a segmentação e a descrição do conceito simultaneamente. A descrição de conceito também pode ser usada para fins de classificação.

Por outro lado, algumas técnicas de classificação que produzem modelos de classificação compreensível podem ser consideradas como descrições de conceito.

A distinção é importante porque a classificação tem como objectivo ser completa no sentido que é aplicável a todos os casos da população seleccionada. Por outro lado, as descrições de conceito não precisam de ser completas e aplicáveis a todo o domínio do problema.

É suficiente que eles descrevam partes importantes dos conceitos ou classes.

Técnicas apropriadas:

- Métodos de indução de regras.
- *Conceptual Clustering*.

Exemplo:

Usando dados sobre os compradores de carros novos e usando uma técnica de indução de regras, podiam-se gerar regras que descrevem os clientes leais e desleais. Exemplos de regras geradas:

Se o SEXO = masculino e IDADE => 51, então CLIENTE = leal

Se o SEXO = feminino e IDADE => 21 então CLIENTE = leal

Se PROFISSÃO = gerente e IDADE <51 então = CLIENTE desleal

Se STATUS FAMÍLIA = bacharel e IDADE <51 então = CLIENTE desleal

[48] *Conceptual Clustering* é uma técnica de classificação não supervisionada que gera uma classificação conceptual para cada classe [27].

6.5 Classificação

A classificação assume que há um conjunto de objectos, caracterizados por atributos ou características, que pertencem a classes diferentes. O rótulo da classe é um valor conhecido para cada objecto. O objectivo é construir modelos de classificação que rotulem com a classe correctos objectos desconhecidos. Modelos de classificação são usados principalmente para a modelagem preditiva.

Os rótulos de classe podem ser previamente fornecidos, definidos pelo negócio ou obtidos pela segmentação.

A classificação é um dos mais importantes tipos de problema para mineração de dados. Esta técnica pode ser utilizada numa ampla gama de aplicações.

Muitos problemas do domínio da mineração de dados podem ser transformados em problemas de classificação. Um exemplo é o *credit scoring* que avalia o risco de crédito de um cliente para financiamento a crédito. Este problema pode ser transformado num problema de classificação através da criação de duas classes: os clientes bons e os maus. O modelo de classificação é gerado a partir de dados de clientes existentes e o seu comportamento relativamente à prestação do crédito. Este modelo de classificação pode então ser usado para classificar um novo cliente.

Técnicas apropriadas:

- Análise da discriminante.
- Métodos de regras de indução *(Rule Induction Methods)*.
- Árvore de decisão.
- Redes neurais.
- Vizinhos mais próximo de K (*K Nearest Neighbor* ou KNN).
- Raciocínio baseado em casos (*Case-based Reasoning*).
- Algoritmos genéticos.

6.6 Previsão

Outro tipo importante de técnica com uma ampla gama de aplicações é a previsão. É muito semelhante à classificação, a única diferença é que na previsão o atributo alvo (classe) não é um atributo qualitativo discreto, mas contínuo.

O objectivo da previsão é encontrar o valor numérico do atributo de destino para os objectos desconhecidos. Na literatura, este tipo de problema é chamado, às vezes, de regressão. Se tratar de previsão com dados de séries temporais, então é designa-se por *forecasting*.

Técnicas apropriadas:

- A análise de regressão.
- Árvores de regressão.
- Redes Neurais.
- Vizinhos mais próximo de K (*K Nearest Neighbor* ou KNN).

- Métodos de *Box-Jenkins*.
- Algoritmos genéticos.

Exemplo:

A receita anual de uma empresa está correlacionada com outros atributos, como a propaganda, a taxa de câmbio, a taxa de inflação etc. Tendo estes valores (e as estimativas confiáveis para o próximo ano), a empresa pode calcular a receita prevista para o ano seguinte.

6.7 Análise de dependências

Análise de dependência consiste em encontrar um modelo que descreva as dependências significativas (ou associações) entre os dados ou eventos.

A análise de dependências pode ser usada para prever os valores de um item de dados, baseado em informações sobre outros itens de dados. Apesar da técnica poder ser usada como um modelo preditivo, é usada principalmente para o entendimento dos dados. As dependências podem ser rígidas ou probabilísticas.

As associações são um caso especial de dependências que descrevem as afinidades de itens de dados (ou seja, itens de dados ou eventos que frequentemente ocorrem em conjunto). Um cenário típico de aplicação para as associações é a análise de cestas de compras. Uma regra como "em 30 por cento de todas as compras, cerveja e amendoim foram comprados em conjunto" é um exemplo típico de uma associação.

Algoritmos para a detecção de associações são muito rápidos e produzem muitas associações. Seleccionar os mais interessantes é um desafio. Análise de dependência tem estreitas ligações à previsão e à classificação, onde as dependências são implicitamente usadas para a formulação de modelos preditivos.

Técnicas apropriadas:

- Análise de correlação.
- Análise de regressão.
- Regras de associação.
- Redes *Bayesianas*.
- Programação de Lógica Indutiva.
- Técnicas de visualização.

Exemplo:

Usando análise de regressão, um analista de negócios descobriu que existe uma dependência significativa entre o total de vendas de um produto, o seu preço e o orçamento para o seu anúncio. Uma vez que o analista descobriu esse conhecimento, pode alcançar o nível desejado de vendas, alterando o preço e/ou as despesas em anúncios nesse sentido.

Apêndice 2 – Plano de projecto

Mapa GANT T produzido no *Microsoft Office Project 2007*

ID	Task Name	Duration	Start	Finish	Resource Initials
1	(Meses 1-2) Pesquisa bibliográfica.	45 days	Mon 01-11-10	Thu 30-12-10	TR
2	**Primeira iteração**	**130 days?**	**Wed 01-12-10**	**Tue 31-05-11**	
3	(Meses 2-3) Business Understanding	44 days?	Wed 01-12-10	Mon 31-01-11	TR;JV;FB
4	(Meses 3-4) Data Understanding.	41 days?	Mon 03-01-11	Mon 28-02-11	HP;LA;AC;JC
5	(Meses 5-6) Data Preparation.	42 days?	Wed 02-02-11	Thu 31-03-11	TR;HP
6	(Meses 6-7) Modelation.	44 days?	Tue 01-03-11	Fri 29-04-11	TR;HP
7	(Meses 7-8) Evaluation	43 days?	Fri 01-04-11	Tue 31-05-11	TR;GPRO
8	**Iterações subsequentes**	**118 days?**	**Wed 09-02-11**	**Fri 22-07-11**	
9	(Meses 2-3) Business Understanding	44 days?	Wed 09-02-11	Mon 11-04-11	TR
10	(Meses 3-4) Data Understanding.	41 days?	Mon 28-02-11	Mon 25-04-11	TR
11	(Meses 5-6) Data Preparation.	42 days?	Mon 28-03-11	Tue 24-05-11	TR
12	(Meses 6-7) Modelation.	44 days?	Mon 25-04-11	Thu 23-06-11	TR
13	(Meses 7-8) Evaluation	43 days?	Wed 25-05-11	Fri 22-07-11	TR
14	Escrita do relatório	115 days?	Fri 15-04-11	Thu 22-09-11	A;MR;AB;JZ
15	Entrega da proposta	0 days	Sat 27-11-10	Sat 27-11-10	TR
16	Relatório preliminar	0 days	Sat 05-02-11	Sat 05-02-11	TR
17	Apresentação	0 days	Wed 22-06-11	Wed 22-06-11	TR;LA
18	Entrega final	0 days	Thu 22-09-11	Thu 22-09-11	TR

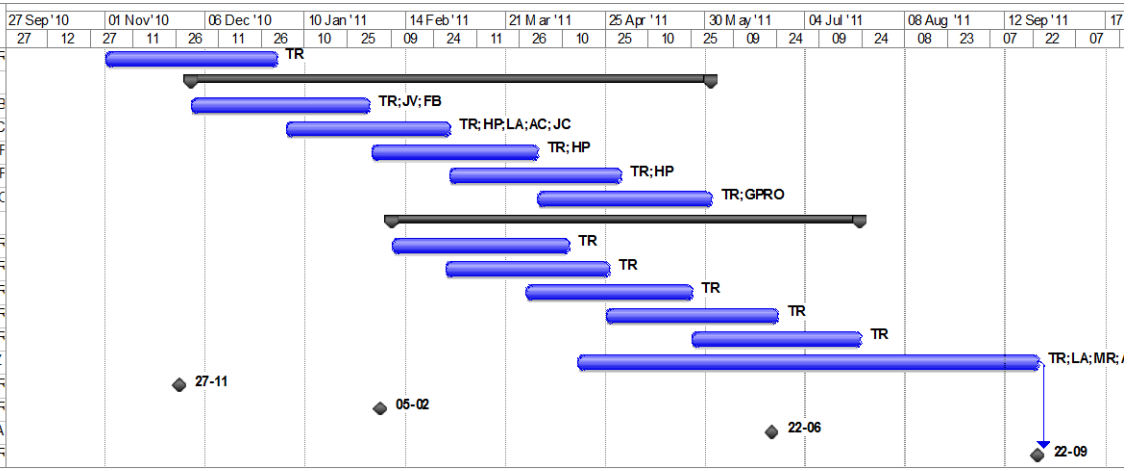


Imagem 17 - Mapa GANTT

Apêndice 3 – Restrições de domínio de atributos

Restrições de dominio	
Símbolo	Significado
I^{+}	números inteiros positivos (maiores que zero) e diferentes de nulo.
I^{+n}	números inteiros positivos.
I	números inteiros diferentes de nulo.
I^{n}	números inteiros.
I^{+0}	números inteiros positivos ou nulos e diferentes de nulo.
S	cadeia de caracteres diferentes de nulo.
$\underline{S^{n}}$	cadeia de caracteres.
D	data diferente de nulo.
D^{n}	data.
M	moeda, positivo e diferente de nulo.
R	número fraccionário.
B	valor booleano e diferente de nulo
B^{n}	valor booleano

Bibliografia

[1] The Standish Group. (1995) projectsmart.co.uk. [Online]. http://www.projectsmart.co.uk/docs/chaos-report.pdf

[2] J. L. Eveleens and C. Verhoef, "The Rise and Fall of the Chaos Report Figures," *IEEE Computer Society*, vol. 10, no. Project Management Focus, 2010.

[3] Caixa Geral de Depósitos. (2011, Janeiro) Organigrama Comporativo do Grupo Caixa Geral de Depósitos. [Online]. http://www.cgd.pt/Corporativo/Grupo-CGD/Pages/Organograma-Grupo-CGD.aspx

[4] B. Inmon. (1999, Novembro) information-management.com. [Online]. http://www.information-management.com/infodirect/19991120/1675-1.html

[5] C. T. Fitz-Gibbon, *Performance indicators BERA Dialogues (2)*. Bristo, USA: Carol Taylor Fitz-Gibbon, 1990.

[6] CRISP-DM. (Dezembro, 2011) CRISP-DM Process. [Online]. http://www.crisp-dm.org/Process/index.htm

[7] ISO/DIS 31000, *Risk management - Principles and guidelines on implementation*, 4th ed.: International Organization for Standardization., 2009.

[8] N. Crockford, *An Introduction to Risk Management*, 2th ed. Cambridge, UK: Woodhead-Faulkner, 1986.

[9] D. Hubbard, *The Failure of Risk Management: Why It's Broken and How to Fix It.*: John Wiley & Sons, 2009.

[10] Project Management Institute, *A guide to the Project Management Body of Knowledge (PMBOK Guide)*, 4. Edição, Ed. Pennsylvania, USA: ANSI/PMI, 2008, ch. 12, pp. 274-312.

[11] R. Mulcahy, *PMP® Exam Prep Rita's Course in a Book for passing the PMP Exam*, 5th ed.: RMC Publications, 2009.

[12] M. J. Carr, S. L. Konda, I. Monarch, F. C. Ulrich, and C. F. Walker, "Taxonomy-Based Risk Identification," Carnegie Mellon University/U.S. Department of Defense., Pennsylvania, USA, Technical Report 1993.

[13] L. J. Cox, "What's Wrong with Risk Matrices?," *Risk Analysis*, vol. Vol 28, No. 2, 2008.

[14] A. Management.COM. (2011, Janeiro) Key Process Indicators. [Online]. http://management.about.com/cs/generalmanagement/a/keyperfindic.htm

[15] D. v. d. Westhuizen and E. P. Fitzgerald, "Defining and measuring project success," , 2005.

[16] W. J. Pinkerton, *Project management: achieving project bottom-line success*.: McGraw-Hill, 2003.

[17] A. G. A. F. Gary Stoneburner, "Recommendations of the National Institute of Standards and Technology," Technology, National Institute of Standards and, Special Publication NSPUE2, 2002.

[18] J. Frand. (2006) What is Data Mining? Apresentação.

[19] R. L. Ackoff, "From Data to Wisdom," *Journal of Applies Systems Analysis*, vol. 16, pp. 3-9, 1986.

[20] P. Chapman et al., *CRISP-DM 1.0 Step-by-step data mining guide*, SPSS, Ed.: CRISP-DM consortium., 2000.

[21] J. Quinlan, *C4. 5: programs for machine learning*.: Morgan Kaufmann, 1993.

[22] E. Alpaydın, *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*.: MIT Press, 2004.

[23] H. Pita. Acetatos Indução. PowerPoint.

[24] Project Management Institute, *A guide to the Project Management Body of Knowledge (PMBOK Guide)*, 4th ed. Pennsylvania, USA: ANSI/PMI, 2008.

[25] Project Management Institute, *A guide to the Project Management Body of Knowledge*

(PMBOK Guide). Pennsylvania, USA: ANSI/PMI, 2008, ch. 2, pp. 28-32.

[26] ISEL. (2010, Setembro) ISEL. [Online]. http://www.isel.pt/dem/Ensino/Mestrados/pdf/REGULAMENTOGERALDEMESTRADO.pdf

[27] F. Morchen, "Time Series Knowledge Mining," , 2006.

[28] U. M. Fayyaci and K. B. Irani, "Multi-interval Descretization of Continuous-Valued attributes for classification learning," , 1993.

[29] Universidade de Waikato. Weka. [Online]. http://www.cs.waikato.ac.nz/ml/weka/

[30] Dicionário Priberam da Língua Portuguesa. Dicionário Priberam da Língua Portuguesa. [Online]. http://www.priberam.pt/dlpo/default.aspx?pal=holding

[31] Publishing Dictionary. (2011, Março) "Framework" definition. [Online]. http://www.thefreedictionary.com/framework

[32] Publishing Dictionary. (2011, Março) "Software House" definition. [Online]. http://www.publishingdictionary.com/definition/software-house.html

[33] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2*, 1995, p. 1137–1143.

[34] D. Bartholomew, "Model Uncertainty, Data Mining and Statistical Inference," *Journal of the Royal Statistical Society.* , vol. 158, no. Series A (Statistics in Society) , pp. 449-466, 1995.

[35] H. W. Ian and F. Eibe, "Data Mining Practical Machine Learning Tools and Techniques,".

[36] H. White, "Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns," 1988.

[37] R. Hoptrof, "The Principles and Practice of Time Series Forecasting and Business Modelling Using Neural Nets," 1993.

[38] S. Yochanan and W. Dorota, "Utilizing Artificial Neural Network Model to Predict Stock Markets," 2000.

[39] E. Codd, "A Relational Model of Data for Large Shared Data Banks," *Communications of the ACM*, vol. 13, p. 377–387, 1970.