ISCTE IUL
Instituto Universitário de Lisboa

DEPARTAMENTO DE CIÊNCIAS E TECNOLOGIAS DA INFORMAÇÃO

Data Mining aplicado ao ITIL®
para Previsão do Tempo de Resolução de Incidentes

Joao Vasco Silva

Dissertação submetida como requisito parcial para obtenção do grau de

Mestre em Gestão de Sistemas de Informação

Orientador:
Doutor Raul M. S. Laureano, ISCTE-IUL

Setembro, 2015

Agradecimentos

Ao Professor Doutor Raul M. S. Laureano, pela confiança, disponibilidade, determinação e exigência.

À minha querida esposa Sandra, pelo seu apoio incondicional e pela compreensão e paciência, sem o qual este trabalho não teria sido possível e ao meu filho Gustavo, que embora se tenha antecipado, nascendo no decorrer deste projeto, contribui com algumas noites de pouco sono, mas acima de tudo com muita motivação e alento nos momentos mais difíceis.

Ao Osório e à Luz, pelo seu apoio em todo o processo de extração de dados, a matéria-prima deste projeto.

Resumo

O alinhamento entre a estratégia do negócio e a estratégia das tecnologias de informação são fatores determinantes para o sucesso de uma organização, fazendo com que exista cada vez mais uma dependência das organizações dos seus sistemas de informação (SI), como ferramenta para a implantação de novas estratégias, e como forma de garantir a continuidade das operações vitais ao negócio.

Neste contexto é fácil compreender que uma organização impedida de explorar na totalidade o seu SI, devido a um incidente, é uma organização que enfrenta perdas financeiras e oportunidades de negócio, muitas vezes, irrecuperáveis.

Esta situação é frequentemente agravada pela dificuldade dos profissionais de TI, na gestão das expetativas da(s) área(s) de negócio afetada(s), nomeadamente na definição de prazos de reposição de serviço, face a um incidente nos SI da organização, o que dificulta a definição de planos de contingência com vista a minimizar o impacto desse incidente.

O ITIL surge como um modelo de Governo bem conhecido e muito utilizado por organizações na gestão dos seus serviços de tecnologias de informação. É hoje reconhecido como uma das abordagens mais eficazes para garantir o alinhamento com a estratégia, otimizar custos, aumentar a qualidade de serviços de TI, a satisfação dos clientes e a produtividade.

Embora o ITIL preconize um processo para a gestão de Conhecimento, com o objetivo de melhorar a qualidade da tomada de decisão, a aplicabilidade deste processo e a extração efetiva de conhecimento dos dados de histórico do ciclo de vida de serviços de TI, não é um padrão e nem sempre é utilizado, pelo que existe nestes dados um manancial de conhecimento, com potencial para alavancar a eficácia dos processos de ITIL.

É neste contexto, e face a esta necessidade, que o presente trabalho procurará explorar o conhecimento residente nos dados de histórico de gestão de Incidentes e com recurso a técnicas de *data mining*, criar um modelo que permita prever o tempo de resolução de um novo incidente.

Pretende-se, deste modo, apoiar gestores e profissionais de TI, a explorarem o conhecimento residente no seu histórico de incidentes, de modo a apoiar na previsão de resolução de novos incidentes e deste modo: potenciarem a eficácia e eficiência dos serviços de TI prestados, apoiar numa melhor gestão de crise face a um incidente disruptivo, ajudar a melhor gerir as espectativas das áreas afetadas e a potenciar a ativação de planos de contingência com vista à redução de impactos. Espera-se deste modo contribuir para uma maior sinergia entre áreas de suporte e as áreas de negócio e a um aumento na satisfação pelos serviços de TI prestados.

Este trabalho é um dos poucos estudos que se propõem a utilizar técnicas de d*ata mining* para estimar tempo de resolução de Incidentes, pretende-se deste modo contribuir para a disciplina

de gestão de conhecimento preconizada pelo ITIL, fornecendo ferramentas e métodos que ajudem a extrair e partilhar o conhecimento, apoiando assim a qualidade na tomada de decisão e a eficácia e eficiência dos serviços de TI prestados.

Palavras-chave: ITIL, gestão de Incidentes, gestão de conhecimento, tempo de resolução incidentes, *data mining*, previsão, modelos preditivos.

Abstarct

The alignment between business strategy and IT strategy is a determinant factor for business success, increasing the dependency that organizations have from their Information Systems (IS), as a tool to implement new strategies and as a way to guarantee the continuity of critical business operation.

In this context it is easy to understand that an organization unable to fully explore their IS, due to an incident, is an organization that faces financial loss and sometimes unrecoverable business opportunities.

This circumstances are normally aggravated by the struggle IT professionals face, to manage business expectations, regarding service recovery estimations, after an incident on the organization's IS, affecting the implementations of contingency plans, aimed to minimize the incident's impact.

The ITIL appears as a governance model well known and largely used by organizations for managing their IT services, today it is known as one of the most effectives approach to guarantee an alignment with the business strategy, to reduce costs, increase quality in IT services, customer satisfaction and productivity.

Even though ITIL advocates a process for knowledge management, created to increase quality in decision making, this process effectiveness to extract knowledge from the IT live cycle data, it is not a standard and normally not used, this suggest the existence of great knowledge in this data, with potential to thrive ITIL process efficiency.

This is the context and need that this study will focus, trying to extract knowledge from historical data of an incident management tool, and by using data mining technics, proposes to create a model that can predict the resolution delay for a new incident.

The goal is to help managers and IT professionals, to explore the knowledge residing on their incident history so they can: increase overall performance of the IT services, to better handle crises management associated with major incident, to better manage expectations of affected business areas and to enable contingency plans and reduce impact. Hoping to contribute in a better relationship between IT and the business and to increase client satisfaction with the IT services.

This study is one of the few that propose to apply data mining to forecast incidents resolution delay, the goal is to contribute to ITIL knowledge management, supplying tools and methods that help extract and share knowledge, to support in decision making and to increase IT service performance.

Key-words: ITIL, incident management, knowledge management, incident resolution delay, data mining, forecast, predictive models.

Índice

Índice de tabelas

Índice de figuras

Lista de abreviaturas

AD – Árvores de decisão

BISL – *Business Information Services Library*

BPMN – *Business Process Modeling Notation*

BSI – *British Standard Institute*

CAB – Conselho consultivo de alterações (do inglês *change advisory board*)

CCTA – *Central Computer and Telecommunications Agency* (governo do Reino Unido)

CI – Item de configuração (do inglês *configuration item*)

CMMI – *Capability Maturity Model Integration*

COBIT – *Control Objectives for Information and Related Technology*

CRISP-*DM – CRoss-Industry Standard Process for data mining*

CSI – Melhoria contínua do serviço (do inglês *continual service improvement*)

CSIP – Plano de melhoria contínua do serviço (do inglês *continual service improvement plan*)

ERP – *Enterprise resource planning*

FCCN – Fundação para a Computação Cientifica e Nacional

GSTI – Gestão de Serviços de Tecnologias de Informação

ITIL – *Information Technologies Information Library*

KDD – *Knowledge Discovery in Databases*

MOF – *Microsoft Operations Framework*

MVS – Máquinas de vetores de suporte

N – média dos valores da saída

RF – *Rrandom forest* (RF)

RM – Regressão múltipla

RNA – Rede neuronal artificia

SEMMA – *Sample, Explore, Modify, Model, Asses*

SI – Sistemas de Informação

SPSS - IBM SPSS Statistics

TI – Tecnologias de Informação

Weka - Waikato Environment for Knowledge Analysis

1. Introdução

1.1. Âmbito e problema

No contexto corporativo, o alinhamento entre estratégia do negócio e a estratégia das Tecnologias de Informação (TI) é determinante para o sucesso de uma organização. Da mesma forma, a dependência das organizações dos seus Sistemas de Informação (SI) é cada vez maior, para a implantação de novas estratégias, mas acima de tudo, para garantir a continuidade das operações vitais ao negócio.

Uma organização impedida de explorar na totalidade o seu SI, devido a um incidente, é uma organização que enfrenta perdas financeiras e de oportunidades de negócio, muitas vezes, irrecuperáveis. É nestas circunstâncias que a disponibilidade dos SI se torna uma prioridade, não só para os gestores de TI, mas também para a direção estratégica das organizações.

Um dos grandes desafios, face a um incidente nos SI que comprometa a atividade, é a gestão das expetativas da área de negócio afetada, que impossibilitada de assegurar a sua produção, procura a todo o custo estimar o tempo da paragem para poder definir planos de contingência e minimizar os seus impactos (informar o cliente para ligar mais tarde, alterar horários, alterar prioridades nas tarefas, etc.).

Contudo, para as equipas de TI nem sempre é fácil estimar o tempo de resolução de um incidente, uma vez que só identificada a causa da anomalia é possível traçar um plano de remediação e estimar o seu tempo de implementação.

É neste contexto e face a esta necessidade, que a presente investigação explora o conhecimento residente nos dados de histórico de gestão de incidentes e, com recurso a técnicas de *data mining,* cria um modelo que permita prever o tempo de resolução de um novo incidente. Consequentemente o estudo pretende ajudar a gerir melhor as expetativas das áreas de negócio afetadas, providenciando prazos estimados de resolução, procurando deste modo minimizar o impacto destes incidentes.

Para melhor compreensão dos seus fundamentos, a investigação tem por base uma ferramenta tecnológica (*software*) de gestão de incidentes, que mapeia e implementa o processo de gestão de incidentes preconizado pelas boas práticas do ITIL (*Information Technology Infrastructure Library*) (Cannon et al., 2007).

O ITIL é um conjunto de boas práticas aplicadas à gestão de serviços de TI. O modelo ITIL procura promover uma gestão com foco no cliente e na qualidade dos serviços de TI e está estruturado por processos e procedimentos, com os quais uma organização pode fazer sua gestão operacional e estratégica das TI, com vista a alcançar o seu alinhamento com a estratégia do negócio (Cartlidge et al., 2007).

A ferramenta de gestão de incidentes regista eventos comunicados pelos utilizadores, relativos a disfuncionamentos ou disrupções nos SI da organização, que de acordo com a sua criticidade são encaminhados e organizados numa fila de trabalho de uma equipa de técnicos (*Help-Desk*) que procedem à sua análise e resolução.

Dada a crescente dependência dos negócios dos seus SI, qualquer disrupção no seu funcionamento implica perdas, que em alguns casos podem ser minimizadas com a implementação de planos de contingência, que ao vigorarem nestes momentos de disfuncionamento, podem atenuar o seu impacto. Contudo, estimar tempos de resolução sem um referencial de conhecimento é complexo, principalmente em momentos de tensão quando um SI está indisponível. Logo, a gestão de expetativas das áreas de negócio afetadas é crucial nestes momentos.

1.2. Objetivos

A investigação visa contribuir para a evolução da Gestão de Conhecimento preconizada pelo ITIL e apoiar gestores e profissionais de TI a potenciar a eficácia e eficiência dos serviços de TI prestados, através da extração de conhecimento residente nas suas ferramentas de gestão de incidentes, apoiando-se no caso real de uma organização portuguesa.

Neste contexto a investigação tem como objetivo, a criação de um modelo preditivo para tempos de resolução de incidentes, numa organização suportada por um SI e com uma ferramenta de gestão de incidentes que implementa as práticas do ITIL, com recuso a técnicas de *data mining*. Em particular, definem-se os seguintes dois objetivos específicos:

- Prever o tempo de resolução de novos incidentes;
- Identificar os fatores explicativos com maior capacidade preditiva do tempo de resolução.

Este trabalho é um dos poucos estudos que se propõem a utilizar técnicas de *data mining* para estimar tempo de resolução de Incidentes, contribuindo deste modo para a gestão de conhecimento do ITIL e fornecendo a profissionais das áreas de suporte de TI, que se debatem com este problema. O modelo preditivo permitirá aos profissionais da área, conhecer uma forma de prever o tempo de resolução de um incidente, com base no seu histórico de gestão de incidentes e, deste modo, apoiá-los na gestão de expectativas, proporcionando estimativas sustentadas de tempos de resolução e contribuindo para uma maior sinergia entre áreas de suporte e as áreas de negócio.

1.3. Abordagem metodológica

A investigação tem por base um caso de estudo de uma organização financeira, que desenvolve a sua atividade em Portugal há 19 anos. Esta organização conta com aproximadamente 450 colaboradores e com um ERP *(Enterprise Resource Planning)*, desenvolvido internamente tendo em consideração todas as especificidades do negócio.

Conta com uma equipa de profissionais de TI de aproximadamente 80 colaboradores, entre técnicos, programadores, arquitetos de sistemas, analistas funcionais, gestores de projeto e gestores de TI, que asseguram todo o ciclo de vida dos SI (conceção, planeamento, implementação, operação e suporte).

Os dados utilizados na investigação foram obtidos da ferramenta de gestão de incidentes da organização, tendo sido utilizada a metodologia CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) para a extração e tratamento dos dados (fases de compreensão do negócio, compreensão dos dados e preparação dos dados (ETL)), sendo posteriormente utilizada a mesma metodologia para a fase de modelação e avaliação.

A ferramenta de gestão de incidentes em causa foi desenvolvida internamente, implementando as boas práticas do ITIL. Na análise considerou-se um histórico de cinco anos e meio (de 2010 a 2015), englobando aproximadamente 44.000 registos de incidentes.

Na fase de modelação da metodologia CRIP-DM recorreu-se a quatro técnicas: árvores de decisão, rede neuronal artificial, máquinas de vetores de suporte e métodos de regressão, com recurso à ferramentas de *data mining open source Weka* e ao *software* comercial *IBM SPSS Statistics.*

Os resultados são interpretados e discutidos tendo por base as teorias subjacentes do ITIL.

1.4. Estrutura do documento

Este trabalho está organizado por capítulos, subdivididos em diversos tópicos, para ajudar a sua estruturação e compreensão.

No capítulo 1 preparou-se o leitor para a leitura do restante documento, contextualizando-o com uma introdução, com o âmbito e o problema, os objetivos e metodologia de análise, e com a descrição da estrutura deste documento.

No Capítulo 2 identificam-se os principais modelos e metodologias para a gestão e governo das TI, com destaque para o ISO/IEC 20000 e para o quadro de referência ITIL v3, e respetivas alternativas.

Ainda no mesmo capítulo é feita uma revisão geral relativamente ao *Business Intelligence* e ao *data mining* onde é abordado o processo de descoberta de conhecimento em bases de

dados, modelos e técnicas de *data mining* e as principais metodologias SEMMA (*Sample, Explore, Modify, Model, Assess*) e CRISP-DM (*Cross-Industry Standard Process for Data Mining*).

O Capítulo 3, diz respeito ao trabalho realizada, onde são apresentadas todas as técnicas de *data mining* aplicadas aos dados em estudo, ao longo das várias fases da metodologia CRISP-DM.

No Capítulo 4 são apresentadas as conclusões e discutidos os resultados obtidos.

2. Fundamentação teórica

Este capítulo apresenta uma breve revisão da literatura com o intuito de estabelecer o referencial teórico que suporta a presente investigação. Está dividido em duas temáticas ITIL e *data mining.* A primeira parte, realiza a identificação dos principais quadros de referência para a gestão de TI, com destaque para o ISO/IEC 20000, apresenta o referencial e descreve o ciclo de vida dos serviços de TI, preconizados na 3ª versão do ITIL, apresenta alternativas ao ITIL e descreve as principais vantagens da sua adoção.

A segunda parte apresenta os conceitos de *Business Intelligence* (BI) e de *data mining*, respetivos processos e metodologias de implementação (*Knowledge Discovery in Databases (KDD*), *Sample, Explore, Modify, Model, Asses (*SEMMA) e *CRoss-Industry Standard Process for Data Mining (*CRISP-DM)), como técnicas e meios para a extração de conhecimento dos dados de natureza operacionais.

2.1. ITIL (The Information Technology Infrastructure Library)

O ITIL teve as suas origens no *UK Office of Governement Commerce's* (OGC) em 1980, como um conjunto de boas práticas aplicadas à gestão de serviços de TI, nomeadamente, infraestruturas, desenvolvimento e operações de tecnologias de Informação (Cartlidge, et al., 2007).

O conceito surge numa altura em que o governo britânico, insatisfeito com a qualidade e nível de serviço fornecida pelas suas instituições estatais, incumbiu a *Central Computer and telecommunications Agency* (CCTA), agora designada OGC, de criar um quadro de referência (*framework)* para promover a eficiência operacional e financeira de recursos de TI no seio do governo britânico e também no setor privado (Cartlidge et al., 2007).

O ITIL tem como principal objetivo, o alinhamento dos serviços de TI com as necessidades do negócio e fornece uma descrição detalhada de um conjunto de boas práticas, através de ações e procedimentos de fácil compreensão e de fácil adoção por qualquer organização.

A publicação original do ITIL em 1989 contava com 31 volumes, cada um dedicado a um tema de gestão de Tecnologias de Informação (OGC, 2007). Na década de 90, grandes organizações e agências governamentais na Europa, começam a adotar este quadro de referências. Em 2000 a Microsoft utiliza o ITIL como base para o desenvolvimento da sua *framework* proprietária (*MOF - Microsoft Operations Framework*) (Microsoft, 2008). Porém, foi no ano de 2001, com o lançamento da 2ª versão do ITIL com 9 volumes mais concisos, que o ITIL consolida a sua posição e em 2006 esta versão, tornaram-se globalmente aceites como uma norma de Gestão de Serviços de TI (GSTI).

Em 2007, o OGC, suportado pelo conhecimento e experiência partilhado por diversas empresas e universidades, publica a terceira versão do ITIL. Esta versão é composta por cinco

volumes fundamentais e procura promover uma gestão com foco na qualidade dos serviços de TI e no cliente. Suporta-se num conjunto de processos e procedimentos, com os quais uma organização pode estabelecer uma visão estratégica e estabelecer as ações necessárias para alcançar o alinhamento estratégico com o negócio.

O ITIL conta com um conjunto de características que em muito contribuem para o seu sucesso. Algumas dessas mais-valias são (OGC, 2007:3-4): i) contém práticas de GSTI, não proprietárias, independentes da tecnologia e aplicáveis a qualquer organização; ii) as boas práticas preconizadas são o resultado da experiência e aprendizagem ao longo de vários anos de diversas organizações e fornecedores de serviços de IT a nível mundial; e iii) podem ser adotadas e adaptadas a qualquer tipo de serviço de TI de qualquer tipo de organização.

2.1.1. Publicações do ITIL e o ciclo de vida dos serviços

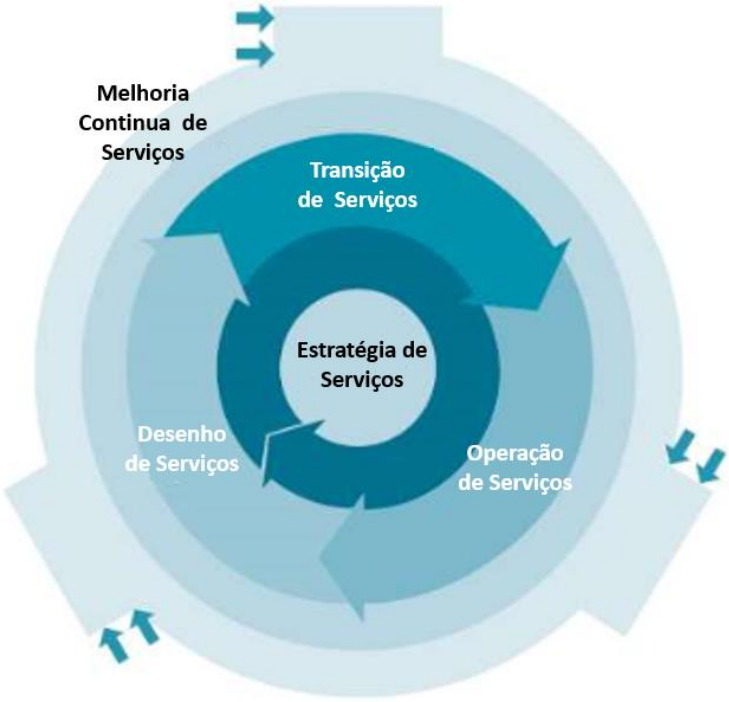
Embora a versão original do ITIL (publicada em 1989) seja diferente da versão atual (publicada em 2007), conceptualmente mantêm-se muito semelhante, focada no suporte e disponibilização de serviços de TI.

A atual versão é constituída por cinco volumes nucleares "ITIL® *Service Strategy*", (Iqbal et al., 2007), o "ITIL® *Service Design*" (Loyd, 2007), o "ITIL® *Service Transition*" (Lacy, 2007), o "ITIL® *Service Operation*" (Cannon et al., 2007) e o "ITIL® *Continual Service Improvement*" (Spalding et al., 2007). Existem outras importantes publicações complementares, incluindo um guia introdutório (Cartlidge et al., 2007), guias de bolso (Bon , 2007) e outros guias complementares com a aplicação do ITIL em cenários específicos.

O livro de estratégia de serviço *(service strategy),* contém orientações acerca do desenvolvimento de uma estratégia de serviços de TI orientada para as necessidades do negócio. O livro de desenho de serviço (*service design*) contém orientações sobre a produção e manutenção de políticas de TI, arquiteturas e documentos para a conceção de serviços e processos de TI, adequados com a estratégia de serviço delineada. O livro de transição de serviço (*service transition)* contém orientações sobre a colocação em produção dos serviços desenhados. O livro de operação de serviço (*service operations*) contém orientações sobre o suporte das operações de uma forma contínua, mantendo os níveis de serviços acordados. Finalmente, o livro de melhoria contínua do serviço (*continual service improvement*) contém orientações sobre a avaliação e a melhoria contínua do valor dos serviços prestados.

A Figura 1 apresenta a relação das cinco publicações que compõem o ITIL, numa representação do ciclo de vida de serviços de TI, suportadas por um processo iterativo de melhoria contínua.

Figura 1: Núcleo do ITIL - Ciclo de vida de Serviços de TI



Fonte: Adaptado de OGC (2007 p. 11)

Cada uma destas publicações, fornece a orientação necessária para uma abordagem integrada, como estipulado pelo ISO/IEC 20000-1 (ISO/IEC, 2005a) e pelo ISO/IEC 20000-2 (ISO/IEC, 2005b)

O ITIL está organizado na forma de ciclo de vida, iterativo e multidimensional, desenhado desta forma para garantir estrutura e estabilidade à gestão de serviços, através de métodos e ferramentas que fornecem as bases para medir a aprendizagem e a melhoria. As boas práticas preconizadas pelo ITIL podem ser adaptadas a qualquer ambiente organizacional e estratégico, podendo o guia complementar (Bon , 2007) ser utilizado para tornar mais robusto o *core* num contexto específico.

Cada volume corresponde a uma das cinco fases do ciclo de vida dos serviços, proposto no ITIL v3. Dada a forte influência do ciclo de *Deming* (também conhecido como PDCA do Inglês *Plan-Do-Check-Act,* método iterativo de quatro passos utilizado em processos de melhoria contínua, muito comum na gestão da qualidade e gestão de segurança) na construção do ciclo de vida, nenhum dos volumes pode ser utilizado isoladamente. Cada fase do ciclo de vida exerce influência sobre as restantes, direta ou indiretamente. As organizações interessadas em adotar o ITIL v3 ou em amadurecer as suas práticas atuais, devem considerar os 26 processos e as 4 funções do ciclo de vida dos serviços na sua totalidade, para poderem

obter todos os benefícios proporcionados pelas orientações do atual quadro de referência ITIL v3 (OGC, 2006).

O ciclo de vida dos serviços inicia-se pela definição de estratégia de serviço, onde são geridos os requisitos do negócio (processo de gestão de procura) e traduzidos numa estratégia para entrega do serviço, volume Estratégia de serviço (ITIL - *Service Strategie*), onde são validados os custos associados à criação e manutenção do serviço (processo de gestão financeira das TI), que passará a fazer parte do portefólio de serviços (processo de gestão do portefólio de serviços). (Iqbal et al., 2007)

Quando a estratégia de serviço está definida, inicia-se a fase do desenho de serviço, descrita no volume Desenho de serviço (*ITIL - Service Design*), através da atribuição de requisitos de nível de serviço aos serviços (processo de gestão do nível de serviço), da análise da disponibilidade e capacidade necessárias (processo de gestão de disponibilidade e processo de gestão da capacidade), da seleção dos fornecedores que darão suporte aos serviços (processo de gestão de fornecedores), da definição da forma de manter a continuidade dos serviços (processo de gestão da continuidade de serviço), da avaliação e projeto dos requisitos de segurança (processo de gestão de segurança da informação) e da introdução do serviço no catálogo de serviços (processo de gestão do catálogo de serviços), (Loyd, 2007).

Depois do desenho de um serviço e assim que o serviço está pronto para ser colocado em produção, dá-se início à fase de transição do seu ciclo de vida, volume Transição de serviço (ITIL - *Service Transition*). O fornecedor do serviço define o plano de transição (processo de planeamento e suporte da transição) e planeia, aprova, implementa e avalia as alterações necessárias (processo de gestão de alterações). Depois, o serviço é testado (processo de validação e teste de serviços) em ambiente de teste. Se o teste for bem-sucedido, o serviço é documentado (processo de gestão do conhecimento) e todas as suas configurações são incluídas na base de dados de itens de configuração (CI do Inglês *Configuration Items*), (processo de gestão da configuração e de ativos de serviço). Finalmente, o serviço é colocado em produção (processo de gestão de liberação e implantação) e é executada uma revisão pós-implementação (processo de avaliação), (Lacy et al., 2007).

A fase seguinte do ciclo de vida, e mais relevante para o presente estudo, é a fase de operação do serviço. Este é gerido e suportado de modo a alcançar os níveis de serviço acordados, estas ações estão descritas no volume Operação de serviço (ITIL - *Service Operation*), onde através de um ponto único de contato (equipa de suporte do Inglês *Service-Desk*) é realizada a gestão dos pedidos dos utilizadores (processo de satisfação de pedidos), são detetados eventos através da monitorização (processo de gestão de eventos), restabelecidas as interrupções não programadas dos serviços (processo de gestão de incidentes), evitadas as causas dos incidentes e minimizados os impactos dos incidentes não previstos (processo de gestão de problemas), gerida a segurança de acessos aos serviços (processo de gestão de

acessos), mantidos os componentes aplicacionais (função da gestão de aplicações), executadas as atividades diárias (função da gestão de operações de TI) e suportada a infraestrutura (função da gestão técnica), (Cannon et al., 2007).

A fase de melhoria contínua do serviço, descrita no volume Melhoria contínua do serviço (*Continual Service Improvement*), é acionada durante todas as fases do ciclo de vida dos serviços. É responsável por avaliar os serviços e os processos (processo de medição de serviço) e documentar os resultados (processo de relatórios de serviço) para que seja melhorada a qualidade do serviço e a maturidade dos processos (processo de melhoria de serviço). Estas melhorias devem ser implementadas na próxima fase do ciclo de vida do serviço, que se inicia novamente pela estratégia do serviço. (Spalding et al., 2007).

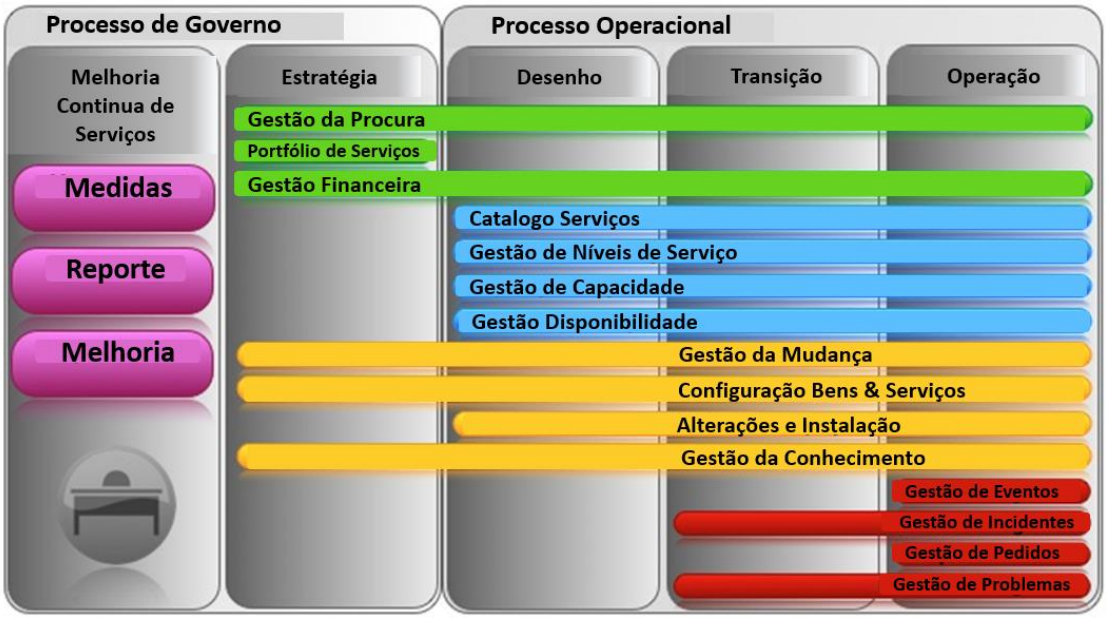
2.1.1.1. Processos no ITIL

O ITIL define Serviços como um meio de entregar valor a um cliente, agilizando os resultados esperados pelo mesmo, exonerando-o de custos ou riscos inerentes. Por outro lado, o ITIL define processos como "um conjunto estruturado de atividades desenhados para atingir um objetivo específico" (Excelos, 2013 p. 4). Um processo recebe uma ou mais entradas *(inputs)* definidas e transforma-as em saídas (*outputs)* definidas. Um processo pode incluir qualquer um dos papéis, responsabilidades, ferramentas e controlos necessários para entregar os *outputs* desejados. Um processo pode definir políticas, normas, orientações, atividades e instruções de trabalho, se forem necessários.

Logo, processos são todos os passos e atividades documentadas que são executadas para suportar um serviço e torná-lo acessível aos clientes. De modo a fazê-lo com sucesso e com confiança, é necessário definir e documentar as atividades ou passos recorrentes, envolvidos na disponibilização e suporte desse serviço.

A Figura 2 apresenta os vários processos que compõem o ITIL, associados a cada serviço e agrupados em segmentos de acordo com a sua posição no ciclo de vida, muitos destes processos subpõem-se entre os vários serviços de acordo com a sua relevância ao longo do ciclo de vida proposto pelo ITIL. Destacando-se dois desses processos (Gestão de Incidentes e Gestão de Problemas) com maior expressão para o presente estudo.

Figura 2: Relação entre Serviços e Processos no ITIL



Fonte: Adaptado de itil v3 service management lifecycle livetime.com (2014)

2.1.1.2. Processo de gestão de incidentes

Um incidente é uma interrupção não planeada para um serviço de TI, ou uma redução na qualidade de um serviço de TI. A falha de um componente (CI do Inglês *Configuration Item*), do qual ainda não se conheça o impacto no serviço, também é um incidente (Cartlidge et al., 2007)

O objetivo da gestão de incidentes é restabelecer a normalidade do serviço tão depressa quanto possível e minimizar o impacto adverso do incidente nas operações do negócio.

Os incidentes são muitas vezes identificados pela gestão de eventos, ou por utilizadores que contatam uma equipa de suporte informático. Os incidentes são categorizados no momento do seu registo por forma a possibilitar a identificação de quem (equipas, níveis de competência) deverá intervir por forma a agilizar a sua resolução, e para possibilitar uma posterior análise de tendências. Os incidentes são também priorizados de acordo com a urgência e com o impacto que têm no negócio.

Se um incidente não puder ser resolvido pela equipa de suporte a quem foi atribuído, este deverá ser escalado para uma outra equipa técnica com os conhecimentos apropriados. Após o incidente ser investigado e diagnosticado, e a resolução testada, a equipa de suporte deve assegurar que o utilizador fica satisfeito antes de o incidente ser formalmente encerrado.

Uma ferramenta de gestão de incidentes é essencial para registar e gerir a informação relativa aos incidentes.

2.1.1.3. Processo de gestão de problemas

Um problema é uma causa de um ou mais incidentes. A causa normalmente não é conhecida na ocasião do registo do problema. O processo de gestão de problemas é responsável por investigar essa causa (Cartlidge et al., 2007). Os principais objetivos da gestão de problemas são a prevenção das anomalias resultantes desses mesmos problemas, a eliminação dos incidentes recorrentes e a minimização do impacto dos incidentes que não podem ser prevenidos. A gestão de problemas inclui as atividades necessárias para o diagnóstico das causas dos incidentes, para determinar a resolução e assegurar a sua implementação. Inclui ainda a documentação da informação sobre os problemas e sobre o modo apropriado de os contornar ou resolver.

Os problemas são categorizados de um modo semelhante aos incidentes, mas o objetivo é, entender as suas causas, documentar as soluções de contorno e as alterações necessárias para a resolução definitiva dos problemas. As medidas para contornar os problemas (*Work arround*) são documentadas numa base de dados de erros conhecidos (KEDB, do inglês *known error database*). A utilização dessa base de dados permite melhorar a eficácia e a eficiência da gestão de incidentes

2.1.2. O ISO/IEC 20000 e o ITIL

Desenvolvida em 2005 pelo BSI (*British Standards Instituton*), é a primeira norma mundial destinada especificamente à Gestão de Serviços de TI (GSTI). Teve como base a norma britânica BS 15000 e é composta por duas partes: a ISO/IEC 20000-1:2011, que consiste na especificação formal e define os requisitos para a gestão do fornecimento de serviços de TI, e a ISO/IEC 20000-2:2012, que define o código de prática para a GSTI. Embora tenha sido originalmente criada para refletir o código de boas praticas do ITIL, suporta igualmente outras práticas de GSTI, como o *Microsoft Operations Framework* (MOF) e o *Control Objectives for Information and Related Technology* (COBIT).

A ISO /IEC 20000 é um *standard* e como tal certificável. Esta certificação permite que organizações adquiram totalidade dos benefícios da utilização das boas práticas da GSTI. Muitas organizações afirmam funcionar de acordo com as boas práticas do ITIL mas frequentemente são implementações seletivas. A certificação perante o *standard* (como com qualquer outro standard) garante que a utilização das boas práticas são auditadas anualmente assegurando todos os benefícios defendidos pela utilização desta abordagem.

.

A Figura 3 demonstra a relação entre o ITIL e o ISO20000, sendo o código de boas práticas ITIL a base para as duas componentes do ISO2000, com especial relevância para o ISO20000-2, que define o código de prática para um GSTI.

.

Figura 3: Relação entre ISO 20000 e o ITIL

Fonte: Adaptado de ITIL V3 support for achieving ISO/IEC 20000 Lacy (2014)

2.1.3. Alternativas ao ITIL

Organizações que procuram tirar partido dos benefícios do ITIL, não necessitam de implementar todas as melhores práticas sugeridas pelo ITIL, até porque a reduzida dimensão ou indisponibilidade de recursos de TI de algumas organizações, pode dificultar ou até impedir que alcancem esse objetivo.

É apenas fundamental adotar as práticas estritamente necessárias, para alcançar os níveis de qualidade na GSTI, que atendam adequadamente às necessidades do negócio e aos objetivos da organização. No entanto, dificilmente se conseguirá alcançar um elevado grau de maturidade dos processos, sem a adoção de todas as melhores práticas do ITIL.

2.1.3.1. Modelos baseados no ITIL

O MOF da Microsoft é derivado do ITIL. Apesar de ser diferente do ITIL, contém algumas partes que são muito semelhantes. Sendo a principal diferença o cariz prescritivo do MOF,

uma vez que foca mais nas indicações sobre "como fazer", quando comparado com o ITIL e é algo ligada aos produtos da Microsoft (Microsoft, 2008).

O PRM-IT (*Process Reference Model for IT*) é um modelo prescritivo e proprietário da IBM, que ajuda a avaliar, desenhar e implementar os processos de TI, com o intuito de auxiliar as organizações a cumprir os seus propósitos e a alcançar os seus objetivos. Este modelo inclui conceitos do ITIL, conceitos do COBIT, do CMMI e da tecnologia *Rational Unified Process* da IBM, e outras práticas aceites pela indústria (IBM, 2009).

O *HP Service Management Reference Model* incorpora muitas das melhores práticas do ITIL, mas é um modelo prescritivo e proprietário da HP, suportado pelos seus próprios produtos (e.g., *HP Open View*). É constituído por cinco componentes chave, (i) estratégia e governação, (ii) desenho e planeamento, (iii) transição e controlo, (iv) operação e tecnologia, e (v) melhoria contínua do serviço (HP, 2010).

2.1.3.2. Outros modelos alternativos

O COBIT à semelhança do ITIL é também um modelo de boas práticas para a governação das TI, (ISACA, 2010), contudo serve propósitos diferentes, fornece aos gestores corporativos, aos auditores externos e aos utilizadores das TI, um conjunto de processos, medidas e indicadores relevantes, que têm o propósito de lidar com todos os aspetos das TI.

É o único modelo que abarca todo o ciclo de vida do investimento em TI e que está em conformidade não só com a lei de *Sarbanes-Oxley* (Sarbanes-Oxley, 2006), mas também com muitas outras normas e modelos, incluindo o ITIL, o CMMI e o ISO 17799., incluindo também as áreas do controlo e da auditoria. O COBIT e o ITIL não são mutuamente exclusivos, e a sua utilização conjunta pode muitas vezes ser benéfica, trazendo um maior controlo à organização dos serviços de TI.

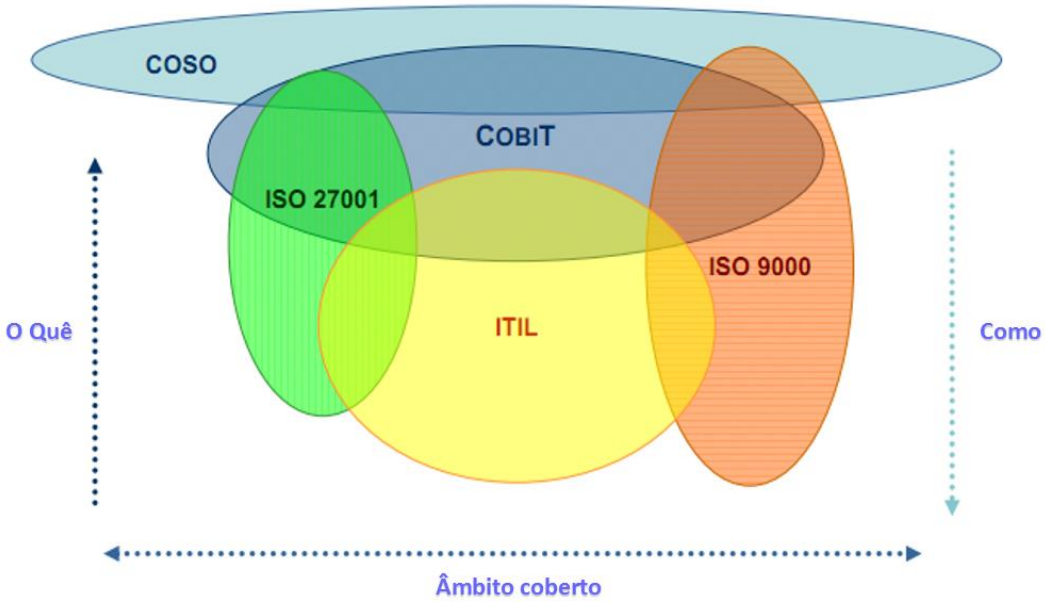
Outro modelo que pode servir como alternativa ao ITIL é o *Capability Maturity Model Integration* (CMMI). É um modelo para medir a maturidade de qualquer processo (OGC, 2007 p. 146). É utilizado na engenharia de *software* e no desenvolvimento organizacional, pretendendo fornecer às organizações os elementos essenciais para a melhoria efetiva dos processos, podendo ser utilizado num projeto, num departamento ou em toda a organização. Pode ser utilizado em três áreas distintas de interesse (SEI, 2010): desenvolvimento de produtos ou serviços (CMMI-DEV); criação, gestão e fornecimento de serviços (CMMI-SVC); e aquisição de produtos e serviços (CMMI-ACQ).

A título de resumo poder-se-á dizer que, o COBIT serve para alinhar processos de TI com objetivos de negócio, o ITIL para melhorar serviços de TI e o CMMI mais vocacionado para a avaliação da maturidade de processos, como já referido, cada um destes modelos serve

propósitos específicos diferentes, contudo quando conjugados fornecem os métodos e ferramentas para uma gestão eficaz desses serviços de TI.

A Figura 4 demonstra as sobreposições e diferenças nos domínios de ação dos vários modelos de governo, sendo que o COBIT cobre maior âmbito, contudo mais orientado para "o que fazer" e não tanto para "como fazer", é neste domínio que o ITIL ganha um maior protagonismo em virtude dos procedimentos e boas práticas que preconiza.

Figura 4: Modelos alternativos e complementares ao ITIL



Fonte: Adaptado de Bang (2010)

2.1.4. A gestão de Incidentes

Como referido, um incidente é por definição uma interrupção não planeada para um serviço de TI, ou uma redução na qualidade de um serviço de TI (Cartlidge et al., 2007). A gestão de incidentes é um processo em que o principal objetivo é restabelecer a normalidade do serviço tão depressa quanto possível e minimizar o impacto adverso do incidente nas operações do negócio. A gestão de incidentes tem uma grande visibilidade para negócio, o que torna mais fácil a demonstração do seu valor, por esta razão, é normalmente um dos primeiros processos a ser implementados em projetos de GSTI.

O processo de gestão de incidentes pode ser desencadeado de várias formas, sendo a mais comum através do utilizador (via telefone, email ou via ferramenta/aplicação de registo de incidentes), contudo, outros métodos podem identificar incidentes automaticamente através de ferramentas de gestão de eventos, que identificam comportamentos anómalos e de forma

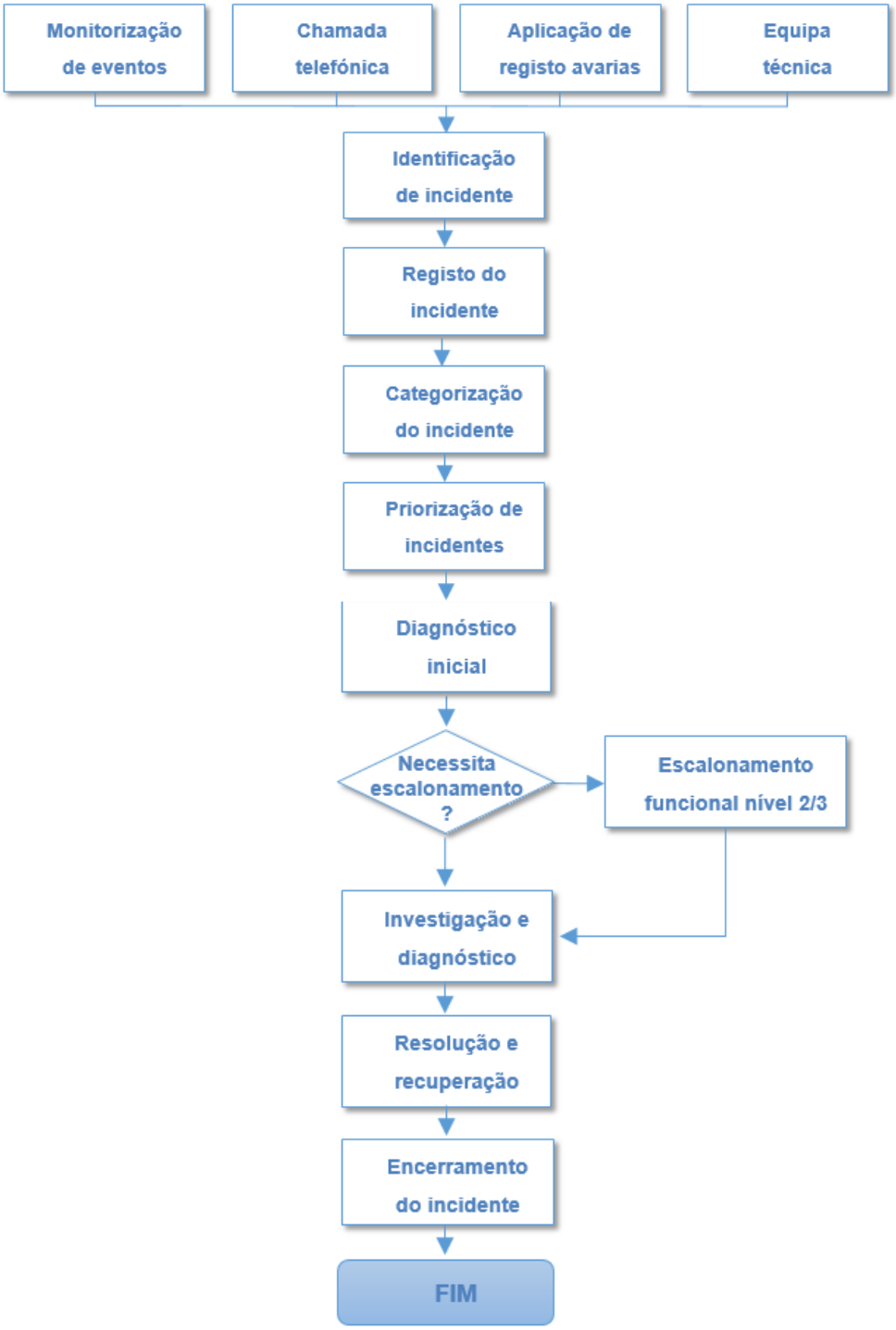
automática lançam alertas diretamente para equipas de suporte, ou via ferramenta de registo e gestão de incidentes.

A grande parte da informação utilizada na gestão de incidentes é proveniente de uma ferramenta adaptada à gestão de incidentes. Esta ferramenta deve implementar, no mínimo, capacidade de registo e gestão de incidentes, contudo se o objetivo for a implementação de um verdadeiro GSTI deve ser utilizada uma ferramenta que suporte todos os processos do ITIL (Cannon et al., 2007).

2.1.4.1. O work-flow na gestão de incidentes

A Figura 5 demonstra em detalhe os passos a serem executados ao longo do processo de gestão de incidentes, desde o momento da sua identificação e catalogação, até à sua conclusão, incluindo todo o processo de investigação, diagnostico, correção e resolução.

Figura 5: Os passos (work-flow) da gestão de incidentes



Fonte: Adaptado de Cannon, et al (2007 p. 90)

Os principais passos resumem-se de seguida :

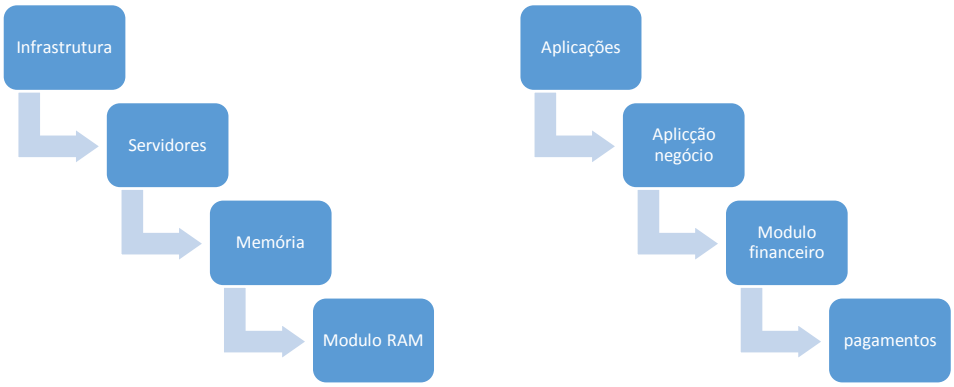
- Identificação de incidentes: é neste passo que se inicia a gestão de incidentes, preferencialmente os incidentes devem ser identificados por processos automáticos (por via das ferramentas de gestão de eventos), evitando-se assim a sua identificação pelos utilizadores do negócio. O racional aqui implícito, está ligado a uma abordagem preventiva na identificação de comportamentos anómalos de modo a corrigi-los antes que estes se tornem incidentes, ou, caso não seja possível, ativar a gestão de incidentes o mais cedo possível de modo a minimizar impactos dessa disrupção.

- Registo do incidente: todos os incidentes devem ser documentados independentemente do modo como são identificados (email, telefone, alerta visual, deteção automática, etc.). Toda a informação relevante deve ser registada e mantida (atributos do incidente), para caso o incidente seja escalado, toda a informação registada até o momento, incluindo dados de registo, diagnósticos e soluções aplicadas, esteja disponível à equipa que vai dar seguimento ao processo de resolução.

  Alguns dos atributos identificados como importantes para a gestão de incidentes (Cannon et al., 2007) incluem:
    - Identificador único
    - Categorização do incidente (normalmente dividido em duas a quatro subcategorias)
    - Urgência, impacto e prioridade
    - Data de registo
    - Identificador da pessoa que regista
    - Método de registo (telefone, email, aplicação)
    - Identificação do departamento do utilizador (localização, telefone, etc.)
    - Método de contato com o utilizador (telefone, email, etc.)
    - Descrição de sintomas
    - Estado do incidente (ativo, em espera, encerrado, etc.)
    - Componentes de TI afetados (computador, telefone, software, etc.)
    - Área de suporte ao qual o incidente está afeto
    - Problema relacionado ou erro já conhecido
    - Registo de ações realizadas para resolver o incidente
    - Data e hora da resolução
    - Categoria de encerramento
    - Data e hora da conclusão
- Categorização do incidente: parte do registo inicial deve ser dedicado à categorização do incidente, esta informação é particularmente útil para o processo de gestão de problemas e para estudos posteriores, sobre análises de tendências, recorrência de incidentes, etc.

A categorização é feita, por norma, por níveis com granularidade de três a quatro níveis. A Figura 6 demonstra dois exemplos de categorização multinível de incidentes de Infraestruturas (*Hardware*) e Aplicacional (*Software*). Em ambos os casos, a categorização inicia-se por uma categoria mais genérica que vai especificando até chegar a um componente específico (módulo dentro da memória de um servidor da infraestrutura). Esta categorização por subcategorias ajuda em análises posteriores de tipificação de incidentes ocorridos.

Figura 6: Exemplos de categorização de incidentes em 4 níveis



Fonte: Adaptado de Cannon et al. (2007 p. 93)

- Priorização de incidentes: juntamente com o registo de um novo incidente deve ser registado a sua prioridade. Esta vai definir o modo como o incidente é tratado. A prioridade é normalmente definida pela urgência (necessidade do negócio) e pelo impacto que este tem no negócio.

  A Tabela 1 apresenta um modo eficaz de relacionar urgência e impacto, de modo a obter assim uma prioridade, num intervalo de 1 (mais alta) a 5 (mais baixa), para um determinado incidente. Apresenta também um exemplo de objetivo de tempo de resolução de incidentes de acordo com a sua prioridade.

Tabela 1: Exemplo de identificação da prioridade de um incidente

PRIORIDADE	Impacto			
Urgência		Alto	Médio	Baixo
	Alta	1	2	3
	Média	2	3	4
	Baixa	3	4	5

PRIORIDADE	Descrição	Objetivo de tempo de resolução
1	Crítico	1 hora
2	Alto	8 horas
3	Médio	24 horas
4	Baixo	48 horas
5	Planear	A planear

Fonte: Adaptado de Cannon et al (2007 p. 95)

A prioridade, resultante do impacto e urgência definidos no registo de um incidente, pode depois ser utilizada para definir a primazia no seu tratamento por parte das equipas de suporte. Por exemplo, a fila de trabalho de uma equipa de suporte que esteja organizada não por ordem de chegada, mas antes por prioridade, pode estar em constante alteração, uma vez que por cada novo incidente registado, a fila estará em constante alteração, estando os incidentes de maior prioridade a serem apresentados no topo dessa lista, independentemente do seu tempo em fila.

- Investigação e diagnóstico: após o registo do incidente as equipas de suporte iniciam a investigação e diagnóstico da situação. Esta deve recorrer à consulta de histórico de incidentes e deve: averiguar como é que o utilizador identificou a anomalia, identificar a cronologia de eventos até à identificação da anomalia, testar (se aplicável) a recorrência da anomalia, e identificar eventos ou ações que possam ter desencadeado a anomalia.

  Se não for possível identificar a causa nem encontrar uma solução, o incidente deve de imediato ser encaminhado (escalado) para uma outra linha de suporte com competências necessárias à sua resolução.

- Resolução e recuperação: quando uma potencial solução é identificada, esta deve ser aplicada e testada, de modo a garantir a recuperação do componente afetado ao seu estado normal, e a total reposição do(s) serviço(s) de TI afetado(s) pelo incidente.

- Encerramento do incidente: a equipa de suporte deve garantir que o incidente está totalmente ultrapassado, que o serviço afetado está operacional e que os utilizadores estão satisfeitos e concordam com o encerramento do incidente.

Deve, igualmente, ser registada informação relativa à resolução e ao encerramento do incidente, onde deve constar; dia, hora, categoria de encerramento, descrição das ações desenvolvidas que levaram à resolução, identificação de procedimentos utilizados. Deve, ainda, ser avaliada a pertinência de escalamento a problema (abertura de problema), caso exista probabilidade de reincidência. Só após estes passos pode o incidente ter o seu encerramento.

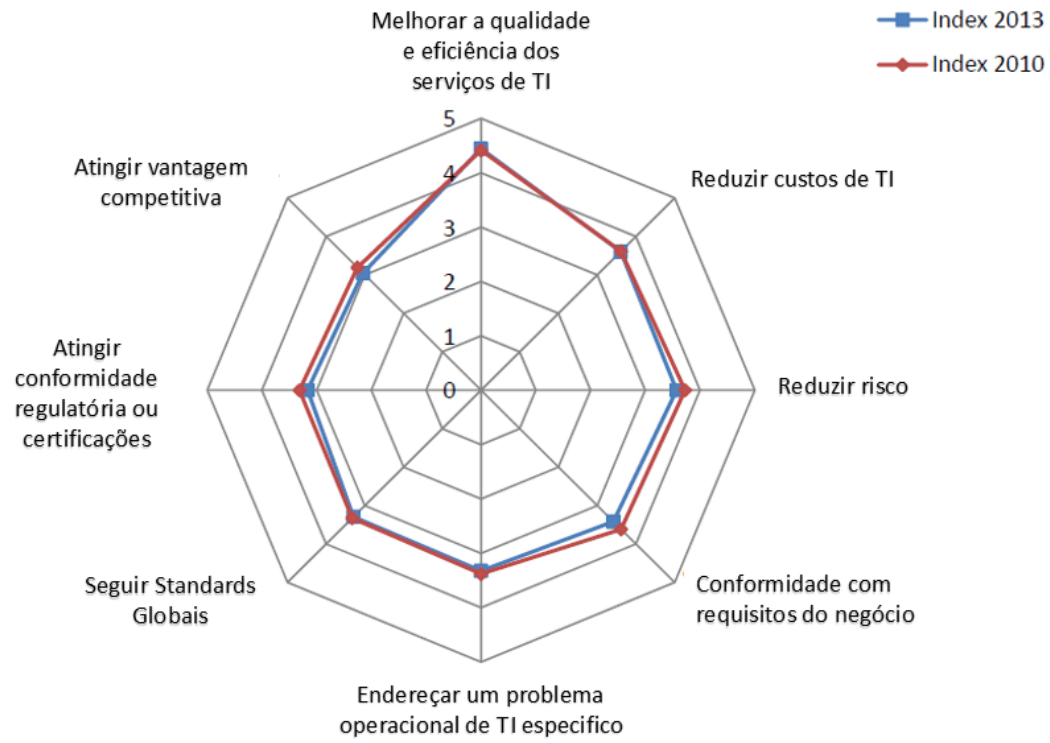
2.1.5. O panorama atual

Um estudo realizado pelo *itSMF internacional* e assistido pelo *National University of Singapore* entre 12 de Março e 30 Abril de 2013, obteve feedback de 738 profissionais de TI, de 49 países sobre a adoção dos GSTI. Este estudo é apresentado de seguida e numa abordagem comparativa, será também apresentado um estudo semelhante realizado pela mesma entidade, em 2010.

De destacar que, segundo o estudo, as principais motivações para a adoção de um GSTI são; a melhoria na qualidade e eficiência de serviços de TI e redução de custos, num contexto de governo das TI, seguidos pela redução de custos e alinhamento com requisitos do negócio, numa ótica de estratégia e alinhamento com o negócio.

A Figura 7 demonstra a evolução (comparativo) entre as razões que motivam à adoção de GSTI pelas organizações, entre 2010 e 2013.

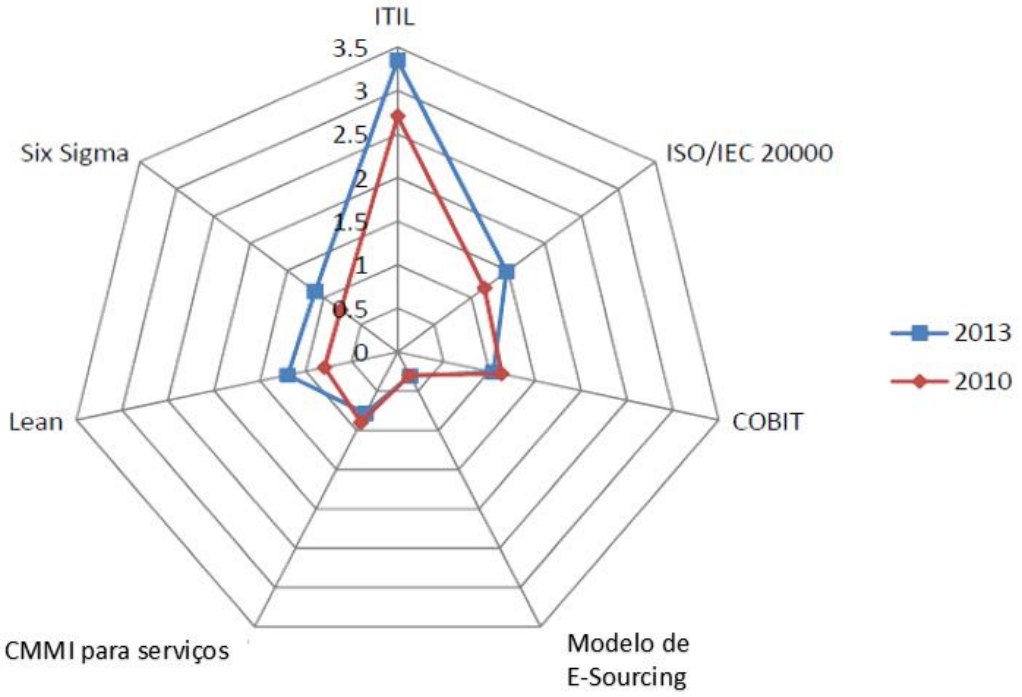
Figura 7: Razões que motivam organizações a adotar GSTI, comparativo 2010 e 2013



Fonte: Adaptado de itSMF (2013 p. 13)

No que respeita à adoção de modelos de boas práticas, o estudo aponta o ITIL com a maior taxa de penetração e utilização pelas organizações, em clara vantagem face às demais metodologias. Tendência esta que se mantem desde o estudo anterior. De notar que COBIT e CMMI, não demonstraram alterações significativas entre 2010 e 2013 (Figura 8).

Figura 8: Modelos de boas práticas adotados por organizações, entre 2010 e 2013



Fonte: Adaptado de itSMF (2013 p. 13)

Quando questionados sobre o sucesso do último projeto de GSTI, os participantes no estudo foram unânimes, tendo afirmado que mais de metade (61,5%) dos projetos tiveram muito sucesso ou foram extremamente bem-sucedidos. A tabela seguinte mostra as percentagens de reposta de acordo com grau de satisfação entre 2010 e 2013, de notar que os projetos considerados como insucessos (satisfação <0), reduziram consideravelmente ao longo dos dois estudos (de 1,1% para 0,3%).

Tabela 2: Taxa de sucesso do último projeto de GSTI

Resultado da Satisfação	Resultados 2010		Resultados 2013		Resultado do Projeto
	Contagem	Percentagem	Contagem	Percentagem	
>100%	80	7,3%	104	14,1%	Extremamente bem sucedido – melhor do que o esperado
80%-100%	479	43,9%	350	47,4%	Bem sucedido – mas dentro do esperado
20%-80%	464	42,5%	237	32,1%	Sucesso
0%-20%	56	5,1%	45	6,1%	Resultado marginal
<0%	12	1,1%	2	0,3%	Insucesso – Projeto falhado

Fonte: Adaptado de itSMF (2013 p. 18)

2.1.5.1. Benefícios da implementação das melhores práticas do ITIL

O ITIL oferece uma abordagem sistémica e profissional para a gestão e fornecimento de serviços de TI, e segundo os seus criadores (OGC, 2007) a adoção e implementação das suas orientações permite a obtenção do seguinte conjunto de benefícios:

- A redução de custos com as TI (eficiência e eficácia na aprestação de serviços de TI);
- A melhoria nos processos de serviços de TI através da utilização de melhores práticas comprovadas;
- O aumento do grau de satisfação do utilizador através de uma abordagem mais profissional na prestação dos serviços;
- A melhoria nos serviços e na comunicação através de terminologia normalizada;
- O aumento de produtividade e o maior foco nas prioridades do negócio;
- A melhor utilização das competências e experiência existentes na organização;
- A melhoria na prestação de serviços de *outsourcing* através da especificação do ITIL e ISO 20000 como padrão para a contratualização da prestação de serviços.

2.2. *Business Intelligence* e *Data Mining*

2.2.1. *Business Intelligence*

Por BI (*Business Intelligence*) entende-se uma conjunto de técnicas e ferramentas utilizadas para transformar dados em bruto, em informação útil e com significado para um determinado negócio. O objetivo do BI é permitir uma fácil interpretação de grandes volumes de dados, de modo a potenciar a extração de conhecimento existente, mas não estruturada, nas bases de dados (operacionais) de uma organização e permitir deste modo explorar novas estratégias e oportunidades de negócio, contribuindo deste modo para a geração e valor.

O uso do conhecimento é um fator crítico de sucesso para qualquer organização, auxiliando na tomada de decisão. Pode-se assim concluir que o grande benefício de uma plataforma de BI numa organização é a transformação da informação em conhecimento. De fato permite a redução de custos, acesso à informação em tempo útil, maior facilidade de análise de dados, eficiência na gestão de recursos e otimização dos investimentos em sistemas de informação (Laureano et al.,(2014).

2.2.2. Descoberta de conhecimento em base de dados e *Data Mining*

Fayyad apresentou em 2006 o processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD). Trata-se de um processo iterativo e exploratório, que permite identificar padrões e modelos com base nos dados, potencialmente úteis e compreensíveis (Fayyad et al., 1996). É um processo com cinco fases que utiliza métodos de *data mining* (numa das cinco fases do processo) para extrair conhecimento através da identificação de padrões, de acordo com a especificação e os limites da fonte de dados. As fases, apresentadas na Figura 9, são:

1. Seleção: esta etapa consiste na criação dos dados a analisar, concentrando-se num subconjunto de variáveis ou dados, no qual a descoberta de conhecimento será executado. Nesta fase, deve-se compreender o domínio dos dados e identificar os objetivos do processo de extração de conhecimento na perspetiva do cliente de modo a identificar subconjunto de dados de onde se pretende obter conhecimento.

2. Pré Processamento: esta etapa consiste na limpeza dos dados de destino e de pré-processamento com vista à obtenção de dados consistentes. Nesta fase deverá haver uma limpeza e preparação/tratamento dos dados, é uma fase tendencionalmente morosa onde deve ser decidida a estratégia para lidar com dados omissos ou erráticos.

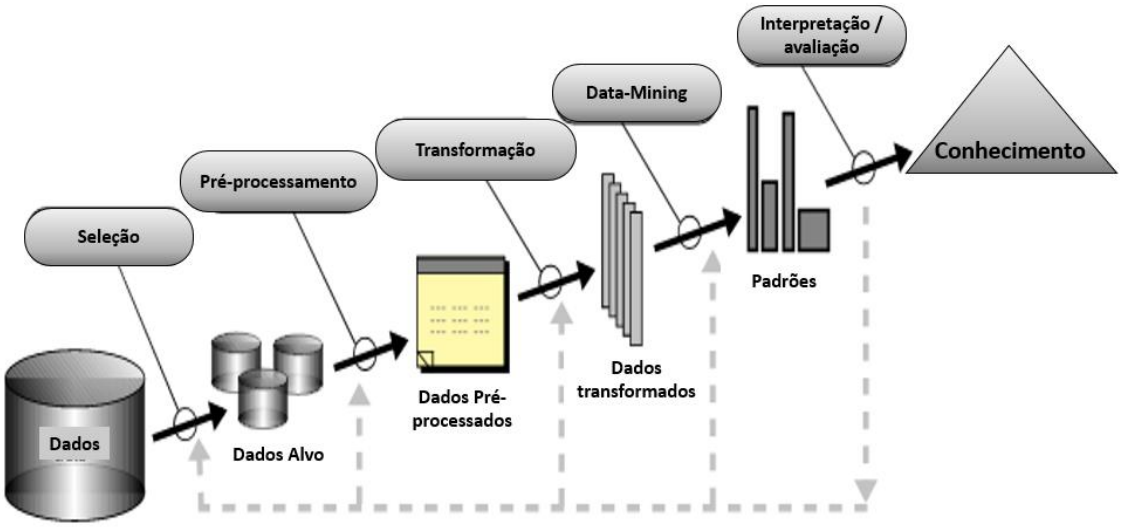
3. Transformação: esta etapa consiste na transformação dos dados usando, por exemplo, técnicas de redução da dimensionalidade ou métodos de transformação. Nesta fase devem ser identificadas caraterísticas mais representativas dos dados.

4. *data mining*: esta fase consiste na procura de padrões de interesse, dependendo do objetivo (geralmente, a previsão). Esta é a fase da análise exploratória de modelação e definição de hipóteses e onde se inicia o *data mining* propiamente dito. O *data mining* ajuda a descobrir nova informação nos dados, através da análise de grandes volumes de

dados; permite automatizar um método de descoberta de padrões nos dados e ajuda na criação de modelos e de conhecimento com base em informação já existente (Han, 2006).

5. Interpretação e Avaliação: Esta etapa consiste na interpretação e avaliação dos padrões identificados. Nesta última fase é feita a interpretação dos padrões com eventuais regressos à fase 1 e consequentes interações.

Figura 9: Processo KDD - Knowledge Discovery in Databases



Fonte: Adaptado de Fayyad et al. (1996 p. 41)

2.2.3. Metodologias de *Data Mining*

Face ao crescimento na área de *data mining*, a indústria tem feitos vários esforços no sentido de criar *standards* para a análise e extração de conhecimento dos dados, entre os quais se destacam o SEMMA e o CRISP-DM, ambas metodologias direcionadas para a implementação de soluções de *data mining*.

2.2.3.1. SEMMA

O processo SEMMA foi desenvolvido pelo *SAS Institute* e significa *Sample, Explore, Modify, Model, Assess,* e refere 5 processos para a criação de projetos de *data mining*. Apesar do SEMMA ser um processo agnóstico à ferramenta de *data mining,* está fortemente associado à ferramenta *SAS Enterprise Miner*. As suas fases descrevem-se:

1. *Sample*: Esta fase consiste em amostragens dos dados por extração de partes do universo de dados. Devem ser suficientes para conter informação significante, contudo tornando o processo mais fácil de manipular.

2. *Explore*: Esta fase consiste na exploração dos dados, procurando tendências inesperadas e anomalias.

3. *Modify*: Esta fase consiste na modificação de dados, pela criação de seleções e transformação de variáveis para o processo de seleção do modelo.

4. *Model*: Esta fase consiste na modelação dos dados através de programas de identificação de padrões que produz o resultado esperado.

5. *Assess*: Esta última fase consiste na análise crítica dos modelos no sentido de identificar a pertinência e robustez dos resultados do processo de *data mining*.

2.2.3.2. CRISP-DM

A Metodologia CRISP-DM, que significa *CRoss-Industry Standard Process for data mining,* foi criada por um consórcio composto por *DaimlerChryler, SPSS* e NCR. É uma metodologia não--proprietária e composta por 6 fases, permitindo o retrocesso a fases anteriores:

1. Compreensão do negócio (*Business Understanding):* Fase para compreender os objetivos do projeto e os requisitos de uma perspetiva de negócio, com o objetivo de converter esse conhecimento numa definição do problema de *data mining* e num plano preliminar desenhado para atingir os objetivos.

2. Compreensão dos dados (*Data Understanding):* Fase de análise e compreensão dos dados, onde se identificam problemas de qualidade de dados e/ou se detetam subconjuntos interessantes de dados para formar hipóteses sobre a informação oculta nos dados.

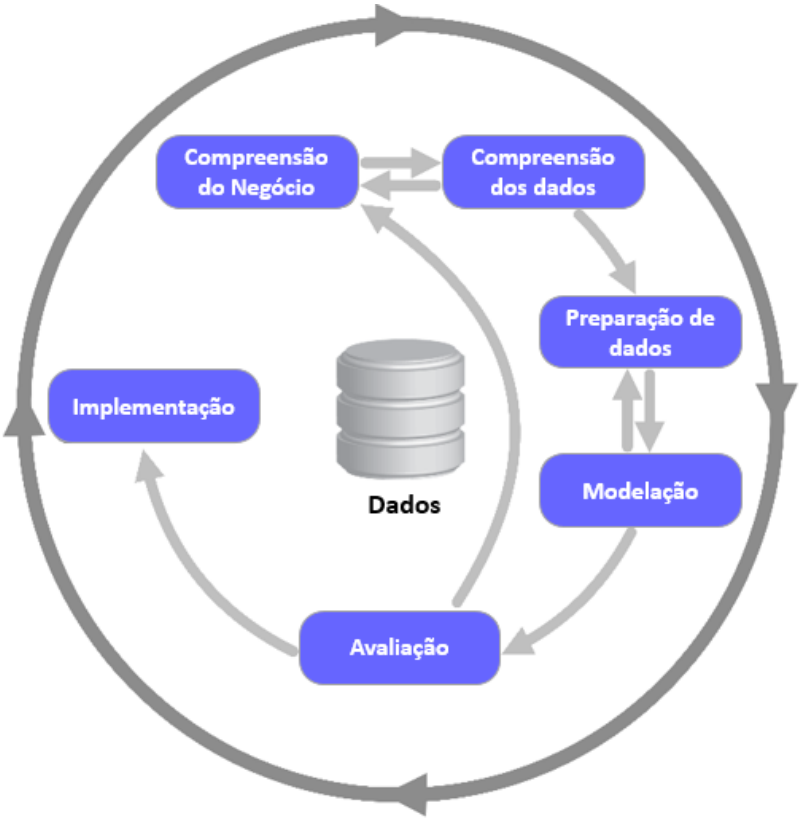
3. Preparação dos dados (*Data Preparation):* Fase de construção do conjunto de dados finais para alimentação das ferramentas de modelação. As tarefas incluem selecionar os atributos, transformar e limpar os dados para utilizar nas ferramentas de modelação.

4. Modelação (*Modelling):* Fase de seleção e aplicação das técnicas de modelação, onde os parâmetros são calibrados para valores ótimos.

5. Avaliação (*Evaluation):* Fase onde o principal objetivo é determinar se há algum problema de negócio importante que não foi suficientemente considerado. No final desta fase, a decisão sobre a utilização dos resultados de *data mining* deve ser alcançado.

6. Implementação ou Desenvolvimento (*Deployment):* Esta fase pode ser simples, tal como a geração de um relatório, ou tão complexo como a implementação de um processo de *data mining* repetível para avaliar os modelos e rever os passos executados para criá-lo, de modo a garantir que o modelo alcance adequadamente os objetivos propostos.

Figura 10: Metodologia CRISP-DM



Fonte: Adaptado de Chapman (2000)

De acordo com Azevedo et al. (2008) ambas as metodologias (SEMMA e CRISP-DM) são implementações do processo de KDD, no sentido em que ambas incluem fases e etapas em tudo semelhantes às do KDD (Figura 11). Por este motivo, muitas vezes se confundem os processos de KDD e DM.

Figura 11: Comparativo KDD, SEMMA e CRISP-DM

KDD	SEMMA	CRISP-DM
Pré KDD	--	Conhecimento do negócio
Seleção	Amostra	Conhecimento dos dados
Pré-processamento	Exploração	
Transformação	Modificar	Preparação de dados
Data Mining	Modelo	Modelação
Interpretação / Avaliação	Avaliação	Avaliação
Pós KDD	--	Implementação

Fonte: Adaptado de Azevedo (2008)

2.2.4. Métodos e técnicas de D*ata Mining*

"Somos ricos em dados, mas pobres em informação" (Han, 2006). Com esta afirmação Han refere que *data mining* serve para extrair ou "minerar" conhecimento de grandes quantidades de dados. De um modo geral podem-se distinguir dois objetivos principais no *data mining* (Fayyad et al., 1996):

- **Verificação:** verificar a hipótese do utilizador; e
- **Descoberta**: procura de novos padrões, podendo ser dividida em:
    - **Previsão:** procura de padrões que permitam prever o futuro; pertencentes a um dos seguintes problemas:
        - Classificação: encontrar uma função que faça o mapeamento dos dados em classes pré-definidas (e.g. diagnóstico de uma dada doença a partir de um conjunto de sintomas);
        - Regressão: encontrar uma função desconhecida cuja saída (ou variável dependente) tem um domínio de valores reais (e.g. previsão do valor de uma ação da bolsa com base em indicadores financeiros).
    - **Descrição:** procura de padrões que apresentem o conhecimento de forma compreensível; utiliza métodos como:
        - Segmentação (*Clustering*): procura de um número finito de conjuntos (ou clusters) que descrevam os dados;
        - Sumarização: procura de uma descrição compacta de um conjunto ou subconjunto de dados;
        - Dependência: procura de um modelo que descreva as relações entre variáveis;
        - Deteção de desvios: descobrir alterações significativas nos dados.

Os métodos de *data mining* podem ainda ser divididos em duas categorias: supervisionados e não supervisionados. A aprendizagem supervisionada, associada a métodos de previsão utilizados para prever um valor, através da relação entre um ou mais atributos de entrada (variáveis independentes) e um atributo de saída (variável dependente) e são utilizados em problemas de previsão e classificação. Já a aprendizagem não supervisionada está associada a métodos descritivos, utilizados para descobrir padrões ou afinidades entre os dados e são utilizados em problemas de *clustering* e sumarização.

Diferentes métodos servem diferentes propósitos e apresentam vantagens e desvantagens. Diferem principalmente no tempo de processamento e de construção dos modelos, na interpretação dos dados, na leitura dos resultados e na sua aplicação aos respetivos domínios. A natureza dos dados e o problema a endereçar podem determinar o método a utilizar, uma vez que o mesmo problema pode ser analisado com recurso a diferentes técnicas de *data mining,* sendo o método escolhido, aquele que melhor promover uma solução para o problema em causa (Liao, 2012).

Descrevem-se de seguida algumas das técnicas mais comuns de *data mining* para aprendizagem aplicadas a bases de dados de grandes dimensões e que terão maior incidência no presente estudo: árvores de decisão/regressão, redes neuronais e máquinas de vetores de suporte.

Árvores de decisão/regressão

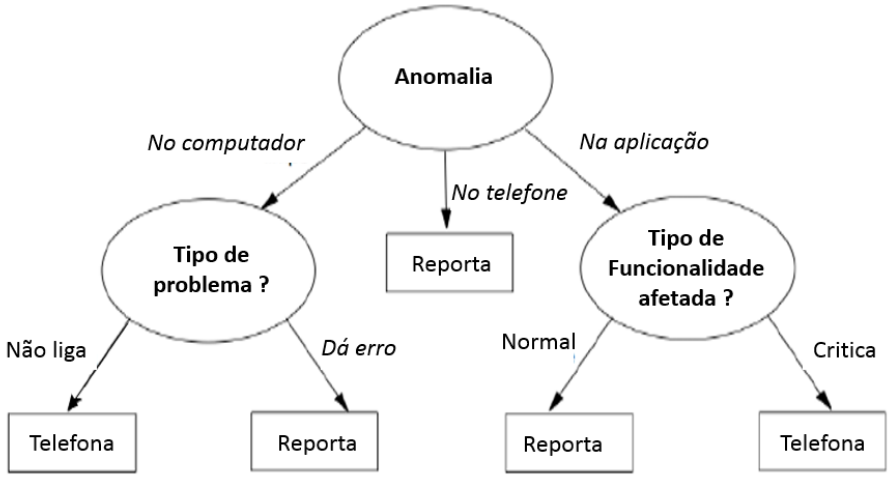
Baseia-se na hierarquização dos dados, com base em estágios/níveis de decisão (nós) e na separação de classes e subconjuntos. É composto por um nó raiz (nó com o primeiro teste); nós internos (cada um possui um teste a um atributo dos dados e têm duas ou mais subárvores que correspondem às respostas possíveis); ramos (contendo valores dos atributos) e folhas (representam as classes).

As árvores de decisão baseiam-se numa análise que testa, todos os valores dos dados para identificar aqueles que são fortemente associados com os componentes de saída selecionados para exame. Os valores que são encontrados com forte associação são os fatores explicativos, usualmente chamados de regras sobre o dado.

Representam métodos relacionados com tarefas de classificação (árvores de classificação) ou regressão (árvores de regressão) e são amplamente utilizados com diferentes tipos e grandes conjuntos de dados, com muitas variáveis e para todos os tipos de problemas de previsões, e de classificação.

As árvores de decisão constituem um dos modelos mais usados em *data mining*, particularmente em problemas de classificação, uma vez que permitem aproximar funções com um contradomínio discreto. Uma das suas principais vantagens assenta na sua facilidade de compreensão, dado que são (em geral) modelos fáceis de interpretar por qualquer ser humano (Michalewicz, 2006). A Figura 12 apresenta um exemplo de árvore de decisão, que face a uma anomalia, ajuda a decidir a melhor forma de a reportar (por reporte ou por telefone).

Figura 12: Exemple de árvore de classificação



Fonte: Adaptado de Rocha (2008)

As árvores de regressão utilizam o mesmo conceito das árvores de decisão, sendo o valor da classe substituído por um valor numérico (ao nível das folhas). Na regressão, em vez do ganho de informação, muitas vezes opta-se pela redução do quadrado dos erros ou desvio padrão (em redor de uma folha).

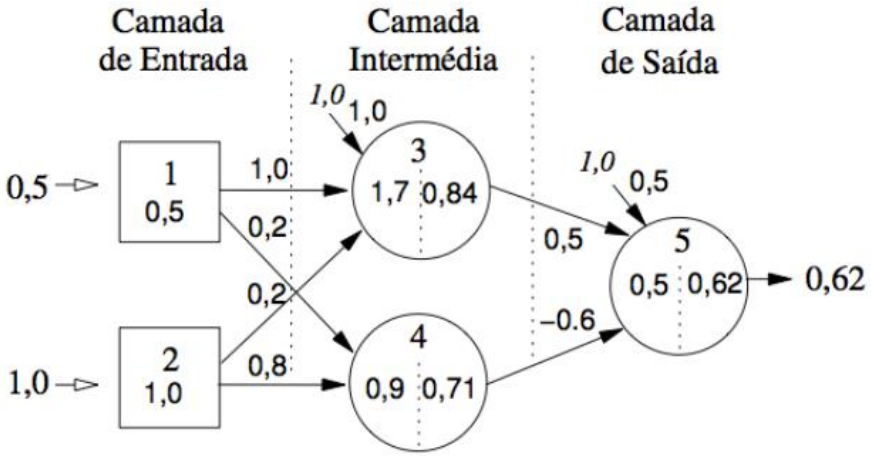
Os algoritmos da árvore de decisão mais populares são CART (c*lassification and regression trees*), C4.5 e CHAID (*Chi-square Automatic Interaction Detector*). Os três algoritmos em termos de funcionamento são muito semelhantes, sendo o CART a exceção, uma vez que é o único que constrói a árvore baseada em divisões binária, enquanto os outros dois permitem divisão múltipla. No que respeita à utilização, o CAHID é particularmente popular em segmentação de Marketing, enquanto o CART e C4.5 são mais populares noutras áreas, nomeadamente em exercícios de previsão.

Redes neuronais

As redes neuronais são inspiradas na fisiologia do cérebro e nas complexas redes neuronais biológicas, à semelhança das redes biológicas, a técnica de redes neuronais é composta por um conjunto de unidades simples de processamento designadas nós (ou neurónios artificiais, semelhante aos neurónios das estruturas biológicas Liao (2012)), que interligados formam uma rede de nós e que em conjunto aumentando a capacidade computacional da unidade. O conhecimento é adquirido a partir de um ambiente (dados), através de um processo de aprendizagem (algoritmo de treino) e armazenado nas conexões entre os nós. Representam métodos relacionados com tarefas de regressão e classificação e tendem a apresentar boas capacidades de generalização (desempenho em previsão).

No caso da aprendizagem supervisionada, a rede é constituída por uma camada de nós (ou neurónios) de entrada, uma ou mais camadas de nós intermédios, e uma camada de saída, conforme ilustrado na Figura 13.

Figura 13: Exemplo de uma rede neuronal tipo Multilayer Perceptron



Fonte: Adaptado de Rocha (2008)

As redes neuronais armazenam informação sobre a forma de "pesos", que são uma ponderação que afetará os valores produzidos por cada nó. Assim, a determinado conhecimento corresponde uma ou mais distribuições de diferentes pesos. Deste modo, ao fornecer determinados sinais de entrada na rede, ela dará uma resposta única, que corresponderá à aplicação do conhecimento que possui, sendo os pesos e a sua distribuição, obtidos através de um processo de aprendizagem da rede.

Máquinas de suporte de vetores

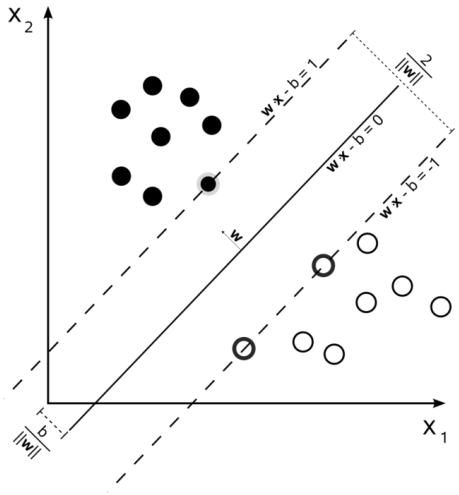
As máquinas de suporte de vetores (*Support Vector Machines*), criadas inicialmente para problemas de classificação dos dados, começaram recentemente a ser aplicadas em problemas de regressão Cruz (2007). Dispõe de um forte fundamento teórico e de um bom desempenho na construção de modelos para conjunto de dados com muitos atributos, embora seja uma técnica muito exigente em termos computacionais.

Foram criadas para problemas de classificação dos dados, mas recentemente começam a ser aplicadas também em problemas de regressão. Baseiam-se na definição e utilização de vetores de suporte que contenham apenas os exemplos mais representativos do universo de treino, e é constituída por várias fases, A primeira é designada por *mapping*, consiste em transformar o espaço dos vários atributos dos dados num espaço multidimensional, com o objetivo de permitir a separação linear dos dados que terá lugar neste hiperespaço. A dimensão deste espaço vai ser a suficiente para que a separação seja linear, portanto, feita

com um hiper-plano. O *mapping* é conseguido com recurso aos métodos de *Kernel*. A segunda etapa consiste na definição do hiper-plano. Este é definido com recurso a vetores que, por sua vez, são construídos a partir de alguns dos atributos dos dados. A seleção destes atributos é feita de modo a que o hiper-plano consiga a separação com a maior distância possível entre as classes. A última etapa consiste na apresentação dos resultados para posterior avaliação e interpretação.

A Figura 14 demonstra um exemplo da separação das classes através da maximização das margens, com recurso a máquinas de suporte de vetores. A linha de separação que garanta a maior distância entre as classes de dados é escolhida de modo a evitar erros de classificação.

Figura 14: Exemplo de máquinas suportadas por vetores



Fonte: Adaptado de wikipedia (2015)

As duas funções de *Kernel* mais comuns são o *Kernel* linear e *Kernel Gaussiano* (não-linear). Diferentes tipos de *Kernel* e diferentes escolhas dos seus parâmetros podem gerar diferentes propostas para os limites das margens. A utilização de *kernels* não lineares permite obter fronteiras não lineares com um algoritmo que determina uma fronteira linear.

2.2.5. Avaliação de modelos

Uma vez que, o objetivo da utilização de um ou mais algoritmos de *data mining* sobre um conjunto de dados é a criação de um modelo que melhores resultados obtém, torna-se assim necessária, a utilização de métodos de avaliação que permitam aferir o grau de eficácia desses modelos. Essa avaliação é feita separando aleatoriamente o conjunto de dados disponíveis

em duas partes: o conjunto de treino, utilizado para estimar os parâmetros do modelo, contêm aproximadamente 2/3 dos dados, o conjunto de teste, utilizado para avaliar a precisão do modelo, contêm os restantes 1/3 dos dados. O modelo é construído com base no conjunto de treino e depois é aplicado ao conjunto dos dados de testes, compara-se o valor da classe de cada exemplo neste conjunto com o que se obtém na previsão, esta técnica é designada por *holdout*. Uma outra técnica muito semelhante à anterior é o método de validação cruzada (Han, 2006) denominado por *k-fold,* este um pouco mais complexo que o anterior, divide o conjunto total de dados em *k* subconjuntos do mesmo tamanho e um desses subconjuntos é utilizado para teste, sendo os *k-1* restantes utilizados para estimar os parâmetros e calcular a eficácia modelo. Este processo é realizado k vezes alternando de forma circular o subconjunto de teste. Ambos os métodos, servem para avaliar o modelo e estimar a incerteza das suas previsões.

Existem várias medidas de avaliação dos modelos que permitem saber o interesse de cada modelo e avaliar a sua performance na classificação dos dados. As medidas de avaliação diferem consoante seja um problema de classificação ou de regressão. No primeiro caso, o mais relacionado com este estudo, têm-se, entre outras, a análise da matriz de confusão e a análise da curva ROC.

Matriz de confusão

A Matriz de Confusão é uma tabela para visualização dos resultados tipicamente usada em aprendizagem de classificação. Os resultados são apresentados sob a forma de tabela de duas entradas: uma das entradas é constituída pelas classes desejadas, a outra pelas classes previstas pelo modelo. As células são preenchidas com o número de instâncias que correspondem ao cruzamento das entradas. Esta técnica tem como benefícios a simplicidade de análise, principalmente se o sistema previr duas classes. Na Figura 15 exemplifica uma matriz de confusão, em que a entrada vertical são as classificações obtidas pelo modelo, e a entrada horizontal são as classificações originais dos dados. Pela análise da tabela pode-se concluir que no caso da classe B, foram classificadas corretamente 46 instâncias, e incorretamente 4 instâncias. Já no caso da classe A, todas as instâncias foram corretamente classificadas.

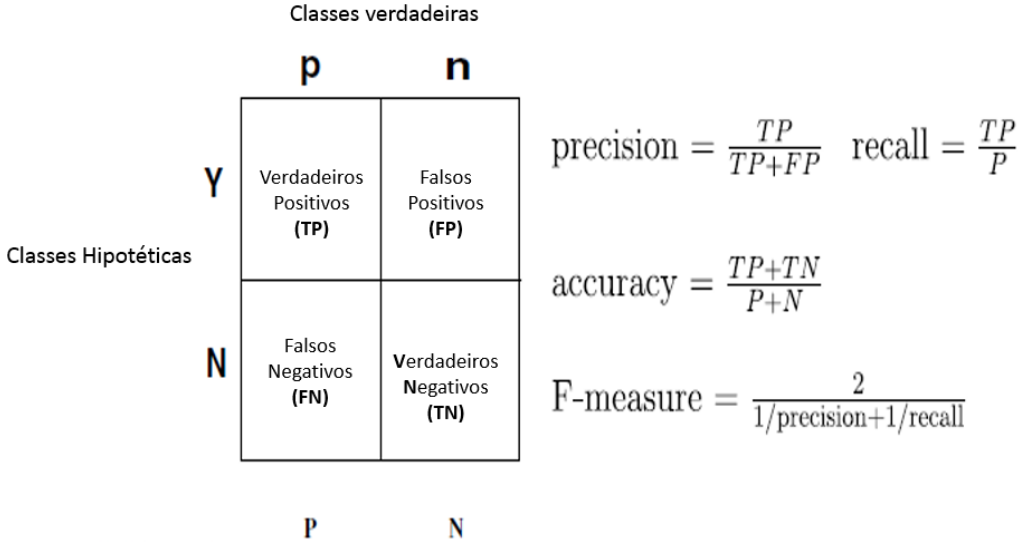
Figura 15: Exemplo de matriz de confusão

	Previsto			
Observado	A	B	C	**Corretamente Classificado**
A	50	0	0	100%
B	0	46	4	92%
C	0	1	49	98%
Percentagem global	33%	31%	35%	**97%**

No caso de sistemas que preveem mais do que duas classes, estas podem ser reduzidas a duas, a classe alvo, designada como classe positiva, e as restantes podem ser agrupadas para formar uma só classe, designada como a classe negativa. Desta forma, pode-se considerar que a matriz de confusão é uma tabela com duas linhas e duas colunas que regista o número de verdadeiros negativos (*True Negatives* -TN), falsos positivos (*False Positives* -FP), falsos negativos (*False Negatives* -FN) e verdadeiros positivos (*True Positives* -TP).

A Figura 16 mostra uma matriz de confusão e equações de um conjunto de métricas comuns que podem ser calculadas a partir da matriz. Os valores ao longo da diagonal principal representam decisões corretas, e os fora da diagonal representam os erros, "a confusão", entre as várias classes.

Figura 16: Matriz de confusão e métricas de performance calculadas através desta



Fonte: Adaptado de Fawcett (2005)

A taxa de verdadeiros positivos (também designados taxa de sucesso) é estimada por:

$$Taxa\ verdadeiros\ positivos\ (TP) \approx \frac{Verdadeiros\ Positivos\ (TP)}{Total\ positivos\ (P)}$$

A taxa de falsos positivos (também designados taxa de falsos alarmes) é estimada por:

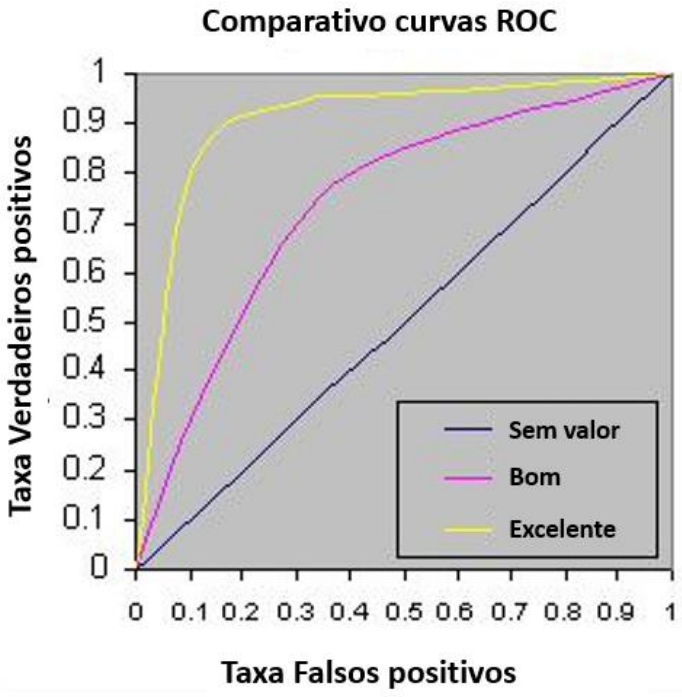
$$Taxa\ falsos\ \ positivos\ (FP) \approx \frac{Negativos\ incorretamente\ classificados\ (FP)}{Total\ negativos\ (N)}$$

Curvas ROC

As curvas ROC (*receiver operating characteristic*) – representam graficamente a relação entre taxa de verdadeiros positivos e taxa de falsos positivos, conforme descrito por Fawcett (2005).

A curva ROC é uma ferramenta que promove a comparação do desempenho dos modelos e que permite visualizar o compromisso entre os dois tipos de erros referidos. A curva ROC representa-se em duas dimensões, com o valor de verdadeiros positivos no eixo dos Y e o valor de falsos positivos no eixo dos X, permitindo desta forma a visualização desta relação. Vários pontos da curva ROC são importantes sendo o mais relevante o ponto superior esquerdo (0,1) uma vez que quanto mais perto a curva estiver deste ponto, melhor será o classificador, pois terá maior taxa de verdadeiros positivos e menor taxa de falsos positivos (Fawcett, 2005), conforme demonstrado na Figura 17.

Figura 17: Exemplo de curva ROC



Fonte: Adaptado de Tape (2015)

Para comparar classificadores pode-se reduzir a curva ROC a um valor escalar. Este valor é representado pela área abaixo da curva do gráfico que relaciona a taxa de verdadeiros positivos e taxa de falsos positivos. Este valor é designado por AUC (*Area Under Curve)*. O valor da AUC igual a 1 representa um teste perfeito. Por oposição, um valor de AUC igual a 0,5 representa um teste fraco e sem valor.

Nos modelos de regressão, para calcular o seu erro (*ei*) ou seja, estimar a diferença entre o valor previsto ($y\hat{}$) e o valor real (y), utilizam-se métricas como MAE (*Mean Absolute Error*) e o

RMSE (*Root-Mean-Square Error*), ajudando na escolha do modelo que produz resultados os mais próximos possíveis dos dados.

RMSE = $\sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y)^2}{n}}$ MAE = $\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_t - y| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$

3. Trabalho Realizado

3.1. Metodologia

Tendo por base o estado da arte apresentado, a metodologia escolhida para a sistematização e descoberta de conhecimento foi o CRISP-DM. A escolha recaiu sobre esta metodologia, em detrimento do SEMMA, por se tratar de um modelo padrão e não-proprietário e, acima de tudo, por estar mais adaptada a problemas de negócio.

A metodologia CRISP-DM preconiza seis fases, iniciados com a fase de compreensão do negócio (*business understanding*), onde é definido o objetivo de negócio, avaliada a situação, determinadas as metas de *data mining* e produzido o plano de projeto.

Na fase compreensão dos dados (*data understanding*) é feita a aquisição dos dados iniciais, e são descritos, explorados e verificados os dados, ou seja, avaliação da qualidade dos dados. Na fase seguinte, a fase de preparação dos dados (*data preparation*), são selecionados e preparados os dados para a fase de modelação. Entre as tarefas destacam-se a limpeza, formatação e integração dos dados.

Na fase de modelação (*modeling*) são selecionadas as técnicas de modelação e é gerado o projeto de teste e construído e avaliado o modelo. A esta fase segue-se a fase da avaliação (*evaluation*) onde são avaliados os resultados e revistos os processos.

Por fim, na fase implementação (*deployment*) é elaborado o planeamento da implementação, e produzido o relatório final e projeto de revisão.

3.2. CRISP-DM Fase 1: Compreensão do Negócio

3.2.1. Contexto

Nesta fase pretende-se compreender os objetivos e requisitos da perspetiva do negócio, convertendo esse conhecimento para a definição de um problema de DM, em que se define um plano e os critérios de sucesso para alcançar os objetivos.

O presente caso de estudo tem por base uma amostra de incidentes, em que os dados foram extraídos de uma ferramenta de gestão de incidentes, que implementa as boas práticas do ITIL. A organização geradora destes dados é uma empresa financeira, que integra um grupo internacional com longos anos de experiencia, com uma operação sustentada e com uma forte estratégia de crescimento por via da internacionalização. A organização desenvolve a sua atividade em Portugal há 19 anos, conta com aproximadamente 450 colaboradores e com uma estratégia comercial focada na diversificação e na relação com o cliente.

Como estratégia de TI, aposta no desenvolvimento interno e em soluções de TI desenhadas à medida no negócio. Dispõe hoje de um ERP *(enterprise resource planning)* desenhado e desenvolvido internamente com o objetivo de dotar o negócio de uma plataforma modular, escalável e de baixo custo e esforço de operação, que acompanhe e potencie o seu crescimento. O departamento de TI da organização conta com uma equipa de 80 profissionais de TI, entre técnicos, programadores, arquitetos de sistemas, analistas funcionais, gestores de projeto e gestores operacionais, que asseguram todo o ciclo de vida dos SI (conceção, planeamento, implementação, operação e suporte).

Em linha com a estratégia de TI, a ferramenta de gestão de Incidentes foi desenvolvida internamente e incorpora as boas práticas do ITIL. A primeira versão desta ferramenta foi disponibilizada em 2007 e tem sofrido várias evoluções fruto das necessidades da organização e da evolução do ITIL, estando hoje alinhada com os princípios preconizados pelo ITIL v3.

A equipa de suporte, que gere e implementa a gestão de incidentes, está organizada num subdepartamento dedicado, inserido no departamento de SI da organização. As equipas estão por sua vez organizadas em três linhas de suporte, de acordo com as suas competências (posto de trabalho, redes e infraestruturas e aplicacional).

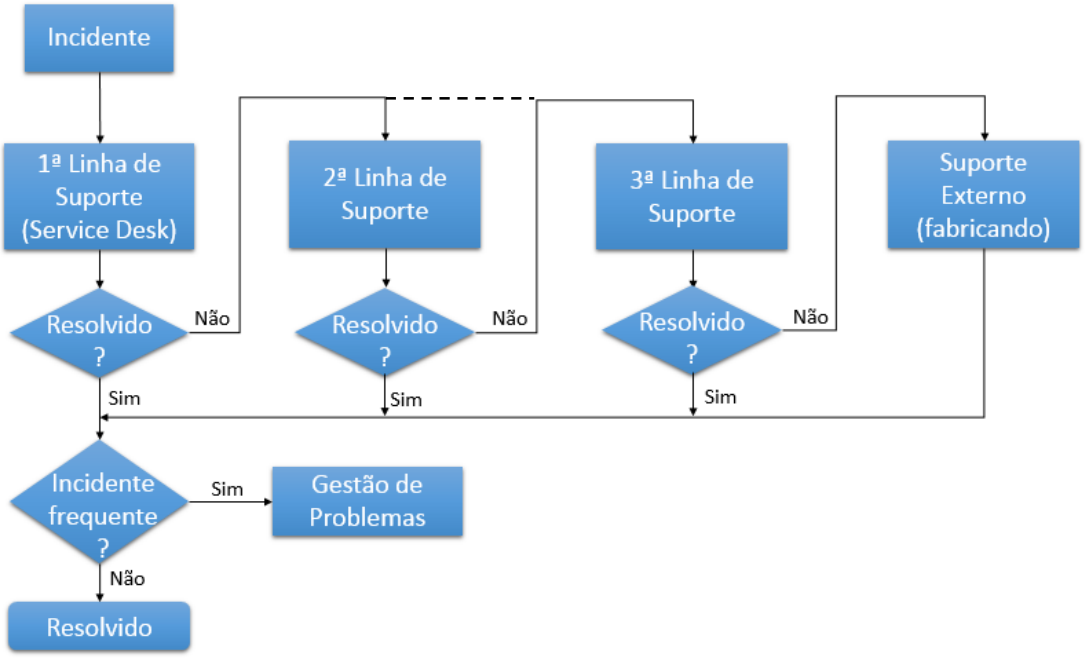
3.2.2. Equipa de Tratamento de Incidentes

A equipa de suporte responsável pelo tratamento de incidentes está organizada em 3 subequipas (ou linhas de suporte), de acordo com as competências: a 1ª linha é constituída por técnicos de informática com um forte conhecimento do negócio, cuja responsabilidade é analisar e triar todos os novos incidentes, sejam eles de infraestrutura (posto de trabalho, comunicações, telefones, etc.) ou de aplicações (aplicações de negócio, relatórios, etc.), a 2ª linha é constituída por administradores de sistemas cuja responsabilidade é a gestão dos postos de trabalho e de toda a infraestrutura de produção (servidores, bases de dados, comunicações, telefonia, etc.) e como tal responsáveis pela sua manutenção e suporte, a 3ª linha é constituída por técnicos programadores com um forte conhecimento ao nível da programação das aplicações de negócio que têm como responsabilidade o diagnóstico de anomalias aplicacionais.

O tratamento de um novo incidente começa com uma análise pela equipa de 1ª linha, que após um diagnóstico inicial avalia se tem as competências necessária para a sua resolução, e em caso positivo, efetua as ações necessárias com vista à sua resolução. Caso o incidente seja solucionado (confirmando com os utilizadores afetados) encerra o incidente, caso contrário (não tendo as competências ou não tendo conseguido uma confirmação da resolução), escala para a linha de suporte seguinte, de acordo com a natureza do problema (relacionado com infraestrutura para a 2ª linha, relacionado com aplicações de negócio com a 3ª linha). Este processo é repetido ao longo das 3 linhas de suporte existentes e eventualmente com o

fabricante ou prestador externo, até que se consiga uma resolução efetiva, um incidente só é encerrado mediante uma validação por parte do utilizador (que pode ser implícita caso não exista resposta por parte do utilizador), este fluxo pode ser observado em detalhe na Figura 18.

Figura 18: Fluxo de Gestão de Incidentes



3.2.3. Ferramenta de Gestão de Incidentes

A organização em estudo, consciente da dependência das suas operações de negócio do seu SI e dos impactos financeiro que a sua inoperância significa, procurou desde sempre implementar ferramentas e metodologias com vista à prevenção e gestão dos incidentes, neste sentido implementou o ITIL e desenvolveu uma aplicação dedicada a gerir todas as anomalias no seu SI. Esta aplicação implementa o fluxo (*work-flow*) representado na Figura 18 e é disponibilizada a todos os utilizadores da organização, existem 2 tipos de perfil de utilizadores na ferramenta, os utilizadores que reportam incidentes e os técnicos que resolvem incidentes (sendo que um técnico também pode registar incidentes).

Registo de um Incidente

No momento do registo de um incidente o utilizador (previamente identificado pela ferramenta) identifica a anomalia identificada, através da seleção de um conjunto de atributos previamente definidos, identificando o que está afetado (aplicação, ferramenta, funcionalidade, etc.) e qual o tipo de comportamento anómalo que identificou. De modo a evitar erros de inserção (erro humano) e de modo a sistematizar e uniformizar a informação registada, estes atributos são recolhidos na sua maioria por intermédio da seleção de opções previamente definidas (*drop-down list*), onde o utilizador é convidado a escolher a classificação que melhor retrata a

situação identificada, sendo registados os seguintes atributos: o tipo de anomalia a registar (Tipo de Incidente), a aplicação ou ferramenta afetada (Serviço), a funcionalidade afetada dentro da aplicação (Categoria), a anomalia que identificou na aplicação (Situação) e por fim através de campo de texto livre detalha a anomalia, de salientar a forma como estes atributos são recolhidos, iniciando-se por uma caraterização mais genérica da anomalia e que se vai detalhado ao logo do processo de registos do incidentes (numa ótica de *drill-down)*, havendo uma relação entre alguns dos atributos, uma vez que, por exemplo, a escolha de uma aplicação ou ferramenta afetada (Serviço), vai influenciar as classificações disponíveis no escolha da funcionalidade afetada (Categoria), estando apenas disponíveis (na *drop-down list*) as funcionalidades da aplicação selecionada. Todos os atributos são de preenchimento obrigatório sob pena da ferramenta não permitir o registo do incidente na eventualidade de existirem campos não preenchidos.

Para além da informação fornecida pelo utilizador, a ferramenta regista de forma automática um conjunto de outros atributos relacionados com datas (Ano, Mês, Dia), prioridade do incidente (Prioridade), identificação do utilizador (Utilizador) e área a que pertence (Área). Uma vez registado o incidente é encaminhado para a fila de trabalho da equipa de 1ª linha, que mediante a sua prioridade faz uma análise preliminar da situação. O agente começa por tentar replicar o comportamento anómalo reportado e se necessário comunica com o utilizador, uma vez compreendido o problema categoriza-o, identificando através de opções previamente definidas; o motivo (desconhecimento do utilizador, anomalia já conhecida, etc.) e a origem (desconhecido, motivado por configurações, relacionados com segurança, etc.), a partir deste momento inicia-se a análise e diagnostico propriamente ditos. Em suma o processo de registo de incidente é bastante robusto e pouco propenso a erro humano, a informação registada, pelo fato de estar pré-definida, está bem sistematização e contêm um nível de detalhe considerável, o que permite efetuar pesquisas por situações semelhantes e análises de tendências nos dados de histórico.

Prioridade no tratamento de Incidentes

Para o cálculo da prioridade é utilizado o método sugerido pelo ITIL, onde a prioridade atribuída a um incidente é calculada em função do Impacto e da Urgência. Ambas as métricas estão previamente parametrizadas na aplicação de gestão de incidentes, que atribui a cada novo incidente uma prioridade (de 1 a 5) de acordo com a matriz apresenta na Tabela 1.

O valor do impacto (alto, médio, ou baixo) está associado ao atributo que identifica a aplicação ou ferramenta afetada (Serviço) e é definido pela gestão estratégica da organização, identificando deste modo as aplicações e ferramentas essenciais à realização da estratégia da empresa (vendas, recuperação, etc.). A Urgência (alta, média, ou baixa) está associado ao atributo que identifica a funcionalidade afetada (Categoria) e é definido pelo negócio,

identificando dentro de cada aplicação/ferramenta, quais as funcionalidades mais críticas para o desempenha das suas funções.

Este modelo relaciona estes dois fatores evitando que uma funcionalidade pouco importante numa aplicação crítica tenha maior prioridade que uma funcionalidade importante numa aplicação menos critica. Esta abordagem tem como objetivo democratizar o processo de gestão de incidentes e evitar a desresponsabilização (envolvendo o negócio) e ao mesmo tempo garantir que as equipas de suporte estão focadas na resolução dos incidentes que têm um verdadeiro impacto na capacidade produtiva da empresa.

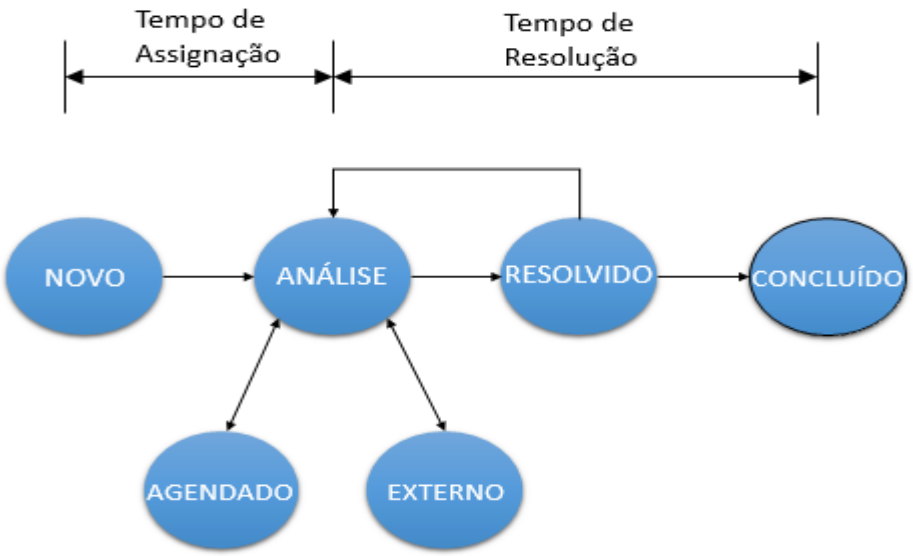
Estados do Incidente

O tratamento de um incidente geralmente começa quando é detetada uma anomalia, seja pelos utilizadores, seja pela equipa técnica, que procede ao seu registo na ferramenta de gestão de incidentes. Assim que o incidente é registado entra numa fila de trabalho e assume o estado de NOVO. Ao receber um novo incidente o agente avalia a situação, passando o incidente para o estado ANÁLISE. O tempo decorrido entre a abertura do incidente e o início dos trabalhos de resolução é designado por Tempo de Assignação.

Uma vez neste estado o agente começam a procurar soluções, o estado mantem-se até que seja encontrada uma solução. Em certas situações, pode ser útil colocar temporariamente o incidente nos estados AGENDADO ou EXTERNO, por exemplo, quando um agente solícita informações complementares ao utilizador, ou encaminha para um prestador externo (após a 3ª linha de resolução). Assim que é encontrada uma solução o estado passa para RESOLVIDO, sendo o utilizador que o registou informado da sua resolução, caso o utilizador não valide a resolução o incidente volta novamente ao estado ANÁLISE.

Assim que o utilizador confirma a resolução, o estado passa para CONCLUÍDO e o incidente é definitivamente encerrado (o incidente é igualmente concluído de forma tácita, caso o utilizador não recuse a resolução ao fim de 3 dias uteis). O tempo decorrido entre o estado de ANÁLISE e CONCLUÍDO é designado por Tempo de Resolução. A Figura 19 ilustra o diagrama de estados do processo de gestão de incidentes.

Figura 19: Diagrama de estados do processo de Gestão de Incidentes



3.2.4. Objetivos do negócio

Conforme referido, a dependência do negócio do seu SI faz com que qualquer anomalia tenha um impacto na produtividade e consequentemente no seu resultado financeiro, incidentes de menor impacto ou de resolução breve, são fáceis de gerir (o utilizador pode dedicar-se temporariamente a outra função, ou pedir a um cliente para "ligar mais tarde"), já incidentes com resoluções mais longas (mesmo os menos críticos, acabam por se tornar críticos com o tempo) colocam problemas mais graves e com maior impacto. Como tal a previsão do tempo de resolução de um incidente é a chave para gerir corretamente expetativas, evitar atritos e decidir atempadamente a ativação de medidas de contingência. Neste sentido o objetivo do negócio é conseguir saber, no momento do registo de um novo incidente, qual o prazo previsto para a sua resolução.

Para responder a este objetivo, será efetuada a descoberta de comportamentos e padrões existentes no histórico de incidentes, com recurso a técnicas de *data mining*, sendo o critério de sucesso a criação de um modelo preditivo que ajude a prever o tempo de resolução de um novo incidente, com uma taxa de sucesso superior a 60%.

No que respeita aos recursos disponíveis, será analisada informação de histórica composta por 44.000 registos de incidentes recolhidos ao longo de 5 anos e meio (entre 2010 e 2015). Ao nível de recursos de *software*, são utilizadas ferramentas *open source (*Weka) e o *software* comercial *IBM SPSS Statistics* (SPSS).

O Weka (*Waikato Environment for Knowledge Analysis*) teve início em 1993, na Universidade de Waikato Nova Zelândia, sendo depois adquirido por uma empresa no final de 2006. O Weka

encontra-se licenciado sob o abrigo da *General Public License*, sendo portanto possível aceder e alterar o respetivo código fonte.

O Weka tem como objetivo agregar algoritmos de *data mining* provenientes de diferentes abordagens/paradigmas, dedicada ao estudo da aprendizagem por máquinas (*machine learning*), disponibiliza um interface gráfico muito intuitivo, composto por uma vasto leque de algoritmos de *data mining* e uma performance e estabilidade excelentes para uma ferramenta não comercial.

O *IBM SPSS Statistics* é uma ferramenta proprietária da IBM vocacionada para a análise estatística, embora seja complementada por um produto da família SPSS, dedicado ao *data mining* (o SPSS Modeler), o *SPSS Statistics* dispõe de várias funcionalidades nesta área tornando-o uma ferramenta muito versátil e poderosa em análise de mineração de dados. Embora não seja uma ferramenta *open-source*, uma vez que a sua principal função é o estudo estatístico, já tem presença em muitas organizações, com um custo muito inferior a ferramenta comerciais dedicas exclusivamente ao *data mining*.

A amostra de dados para o presente estudo foi extraído da base de dados da ferramenta de gestão de incidentes, via SQL (*Structured Query Language*), sendo os dados depois exportados para formato Microsoft Excel. Esta extração permitiu recolher todos os registos de incidentes entre 2010 até ao presente (2015). Embora a ferramenta contenham variada informação (atributos) relacionados com o tratamento de incidentes, neste estudo apenas são considerados os dados recolhidos durante o tempo de assinação (Figura 19), ou seja, os dados disponíveis no momento do registo do incidentes e com os quais se pretende prever o tempo de resolução.

O tempo de resolução foi calculado no processo de extração e considera o tempo decorrido na perspetiva do utilizador, ou seja, desde o momento que o utilizador submete o incidente, até ao momento que o considera resolvido, de acordo com a Figura 19, este tempo representa o somatório do tempo de assignação com o tempo de resolução, tendo sido excluídos horários e dias não uteis (fins de semanas e feriados).

A fase seguinte do CRISP-DM detalha toda a preparação e transformação destes dados.

3.3. CRISP-DM Fase 2: Compreensão dos Dados

Esta fase inicia-se com a preparação e análise dos dados recolhidos, e procura identificar problemas de qualidade dos dados (valores discrepantes, omissos, etc.), identificar potenciais relações entre os dados e a deteção de subconjuntos interessantes.

Conforme já descrito, a ferramenta contêm dados relativos a todo o ciclo de tratamento do incidente (Registo, Categorização, Priorização, Investigação e diagnóstico, Resolução e recuperação, Encerramento do incidente), contudo para este estudo apenas são considerados os dados recolhidos (e conhecidos) no momento do registo e o tempo final de resolução (que corresponde à variável dependente).

A Tabela 3 apresenta cada um dos atributos extraídos (coluna atributo), descreve-os (coluna descrição) e apresenta o seu método de registos (coluna Tipo de registo)

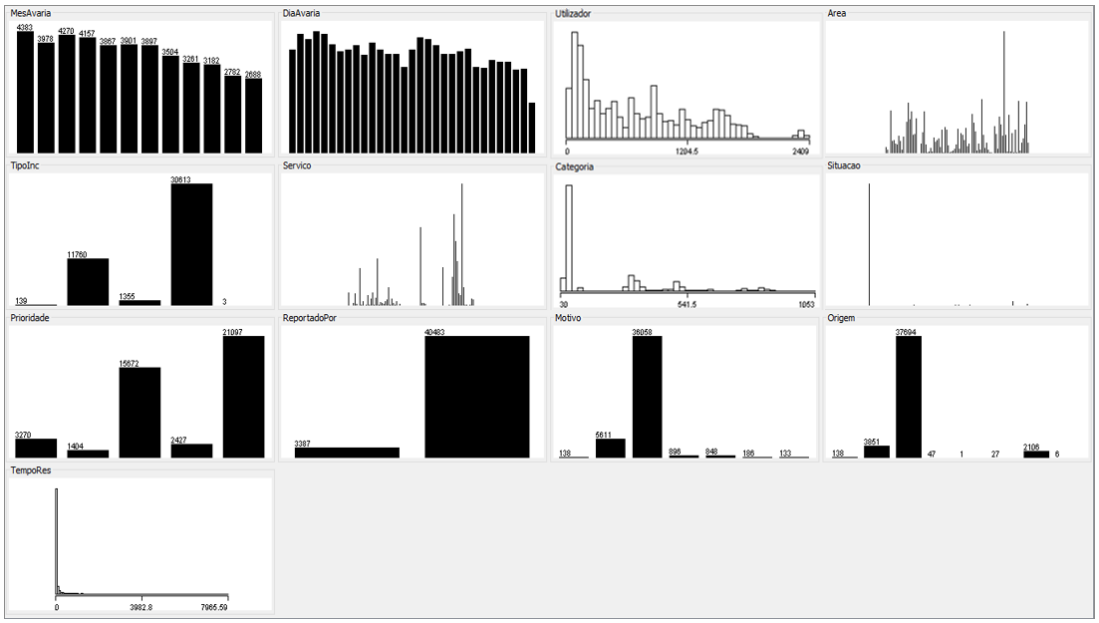
Tabela 3: Atributos registados pela ferramenta de gestão de incidentes

Atributo	Descrição	Tipo de registo
N_Reporte	Código sequencial que identifica cada incidente	Registado de forma automática, atribuindo um número sequencial a cada novo incidente.
Ano	Ano de registo do incidente	Registado de forma automática, de acordo com a data do registo do incidente.
Mes	Mês de registo do incidente	Registado de forma automática, de acordo com a data do momento de registo do incidente.
Dia	Dia de registo do incidente	Registado de forma automática, de acordo com a data do momento de registo do incidente.
Utilizador	Código do utilizador que regista o incidente	Registado de forma automática, de acordo com o identificador numérico do utilizador que submete o incidente.
Area	Código que identifica a área do negócio (departamento) onde se encontra o utilizador que regista o incidente	Registado de forma automática, de acordo com o código do departamento do utilizador que submete o incidente.
Tipo_Inc	Código que identifica o tipo de incidente (e.g. nas aplicações, no posto de trabalho, de Segurança, etc.)	Registado pelo utilizador, através da identificação do tipo de incidente, de uma lista predefinida (*drop-down list*)
Servico	Código que identifica de um modo genérico o que está afetado, utilizado para identificar uma aplicação ou ferramenta (e.g., aplicação de faturação, telefone do posto de trabalho, etc.)	Registado pelo utilizador, através da identificação do serviço, de uma lista predefinida (*drop-down list*) populada mediante escolha do Tipo_Inc
Categoria	Código que identifica em detalhe o que está afetado, utilizado para identificar uma funcionalidade numa aplicação ou ferramenta (subcategoria de serviço) (e.g., funcionalidade X da aplicação de faturação, a opção Z no telefone, etc.)	Registado pelo utilizador, através da identificação da categoria, de uma lista predefinida (*drop-down list*) populada mediante escolha do Servico
Situacao	Código que identifica o comportamento anómala identificado (subcategoria da categoria) (e.g., Serviço indisponível, quebra de performance, avaria no equipamento, dados incorretos, etc.	Registado pelo utilizador, através da identificação da situação, de uma lista predefinida (*drop-down list*) populada mediante escolha da Categoria
Prioridade	Código que identifica a prioridade do incidente (relação entre urgência e o impacto)	A prioridade é calculada automaticamente mediante a Urgência e impacto, predefinidas de acordo com registo do serviço e da categoria
Reportado_por	Código binário que identifica o perfil (utilizador ou agente) do utilizador que registou o incidente	Registado de forma automática, de acordo com o código do utilizador (utilizador ou agente)
Motivo	Código do motivo (aparente) do incidente (e.g., indisponibilidade de serviço, desconhecido, comportamento inesperado, utilização incorreta)	Registado pelo técnico, após uma triagem do incidente, através da identificação do motivo, de uma lista predefinida (*drop-down list*)
Origem	Código da origem do incidente (e.g., origem conhecida, desconhecida, problema de configuração, etc.)	Registado pelo técnico, após uma triagem do incidente, através da identificação da origem, de uma lista predefinida (*drop-down list*)
Tempo_Res	Tempo de resolução (em minutos)	Registado de forma automática, após a conclusão do incidente

De notar que os atributos N_Report e AnoAvaria, embora presentes no momento do registo de um novo incidente, assume-se que dada a sua natureza (incremental), não tem qualquer relevância para o modelo a criar, logo não são considerados na amostra. Os restantes atributos são depois alvo de análise e tratamento.

Após extração da amostra constatou-se a existência de 44.009 registos, com a dispersão apresentada na Figura 20.

Figura 20: Histogramas dos atributos da amostra



No que respeita à relação entre atributos, foi utilizado o ETA e o Spearman, como medida de avaliação do grau de associação entre os diferentes atributos (nominais e ordinais respetivamente) e o tempo (Tabela 4).

Tabela 4: Relação entre os atributos explicativos e o tempo de resolução

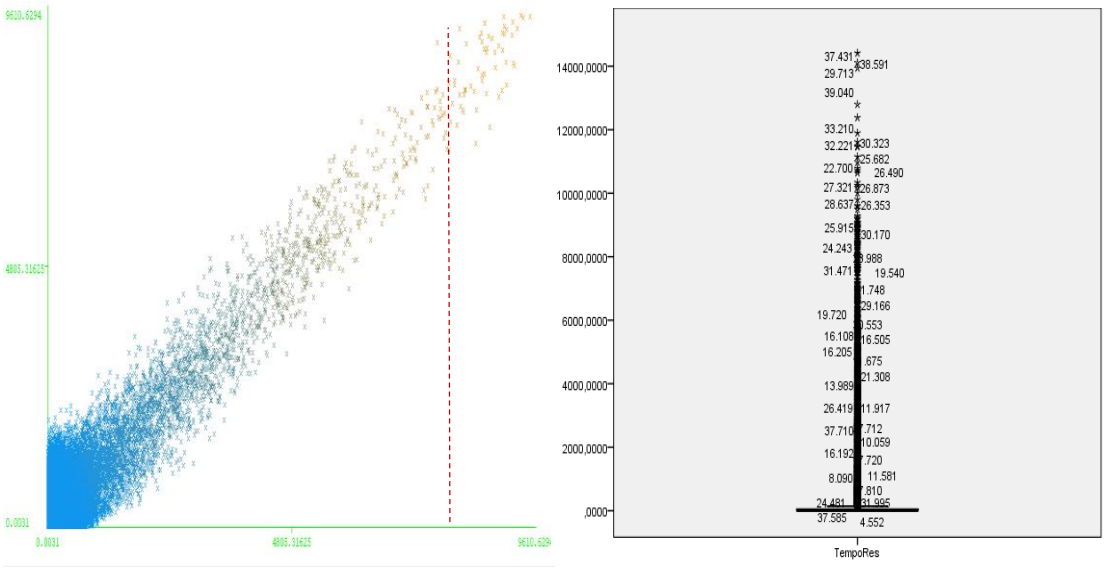
TempoRes	ETA	Spearman
MesAvaria	0,038	
DiaAvaria	0,038	
Utilizador	0,242	
Area	0,141	
TipoInc	0,159	
Servico	0,212	
Categoria	0,275	
Situacao	0,094	
ReportadoPor	0,008	
Motivo	0,071	
Origem	0,086	
Prioridade		-0,17

Dada a reduzida quantidade de registos omissos (65 registos na atributo Area), optou-se por eliminar estes incidentes do estudo. Esta equidade nos dados, conforme referido, deve-se essencialmente à robustez da aplicação de gestão de incidentes no que respeita ao seu processo de registo, uma vez que o incidente apenas pode ser submetido quando todos os campos (atributos) estão preenchidos, e estes não são inseridos manualmente (apenas por seleção de listas previamente definidas), o que elimina o fator erro humano e se traduz numa taxa residual de dados inseridos incorretamente ou omissos.

3.4. CRISP-DM Fase 3: Preparação dos dados

Nesta fase procedeu-se à transformação e limpeza dos dados e todas as atividades necessárias para construir o conjunto de dados finais (Chapman et al., 2000). Uma análise exploratória mais profunda dos atributos, por via da análise de diagramas de frequência, revelou que o atributo candidato a variável dependente (TempoRes) assumia valores muito díspares e elevada assimetria (média de 249 horas, desvio padrão de 845 horas e moda e mediana próxima das 12 horas). Embora a amostra demostre uma forte tendência para tempos de resolução baixos existem um número considerável de tempos de resolução extremamente longos (max. 14.411 horas). Assumiu-se, portanto, que estes valores discrepantes seriam erros ou situações muito pontuais e como tal foram considerados *outliers* e removidos da amostra. O critério foi tempos de resolução superiores a 8.000, o que retirou 71 registos da amostra, conforme se pode constatar na Figura 21. No diagrama de frequência à esquerda, a linha tracejada identifica os 71 registos eliminados (considerados *outliers)*. À direita é apresentada a *boxplot* deste atributo, onde se confirma a sua assimetria.

Figura 21: Diagrama de frequência e boxplot da varável dependente TempRes



3.4.1. Criação de novos atributos

Com o objetivo de potenciar o modelo preditivo, foram criados três novos atributos. Estes atributos foram gerados com base em atributos já existentes, através da agregação/agrupamento dos seus valores. Pretende-se deste modo diversificar os atributos disponíveis para o estudo, na expetativa desta diversidade melhorar o desempenho das técnicas de *data mining* a explorar.

Uma vez que a variável dependente (TempoRes) assume valores numéricos contínuos, limita a utilização de métodos de regressão. Por esta razão foi criado um novo atributo ordinal que assume valores discretos (Grupo), agrupando o tempo de resolução em quatro intervalos, de acordo com a sua relevância para o negócio. A Tabela 5 detalha as quatro classes criadas, onde são apresentados os critérios de agrupamento de tempos (coluna Intervalo de resolução), o nome do novo grupo criado (coluna Nome do Grupo) e a relevância para a atividade (coluna Descrição). Deste modo, com esta nova variável dependente discreta é possível testar técnicas de regressão (variável dependente TempoRes) e técnicas de classificação (variável dependente Grupo).

Tabela 5: Descrição do atributo e variável dependente criado "Grupo"

Intervalo de resolução (em horas)	Nome do Grupo	Descrição
[0;1[	1H	Incidentes com tempo de resolução compreendido entre 0 min. e 1 hora, para o negócio é tempo ideal uma vez que até 60min. é um tempo aceitável entre o registo e a resolução do incidente, uma vez que no decorrer deste intervalo é possível encontrar alternativas sem impacto de maior para a atividade.
[1;4[	1H4H	Incidentes com tempo de resolução compreendido entre 1 e 4 horas, são incidentes com algum impacto para o negócio uma vez que podem corresponder a meio dia de indisponibilidade, dependendo da criticidade, um tempo expetável de resolução desta dimensão pode, mediante o impacto no SI afetado, justificar a ativação de planos de contingência operacionais.
[4;8[	4H8H	Incidentes com tempo de resolução compreendido entre 4 e 8 horas, são incidentes que tendencialmente demoram entre meio-dia a um dia completo, por norma e dependendo do seu impacto, obrigam à implementação de planos de contingência operacionais.
[8;∞[	M1D	Incidentes com tempo de resolução superiores a 1 dia (8 horas) são situações mais complexas ou menos criticas (uma vez que prioridades baixas tendencialmente implicam tempos de resolução mais longos), uma vez mais, dependendo do impacto do incidente, obrigam à implementação de planos de contingência operacionais ou mesmo técnicos.

Com o objetivo de identificar eventuais padrões relacionados com sazonalidade no registo de incidentes, foram criadas dois novos atributos relacionadas com a data do registo, o dia de semana (Dia_sem) e número da semana (N_sem). A Tabela 6 descreve e demonstra o modo de criação destes atributos, a partir dos atributos Ano, Mês e Dia.

Tabela 6: Novos atributos criados "Dia_sem" e "N_sem"

Nome do atributo	Método criação	Descrição e valores assumidos
Dia_sem	Função Excel: =DIA.SEMANA(DATA(“ano”;”mês”;”dia”))	A função DIA.SEMANA devolve o dia da semana sendo a codificação por omissão: 1 (Domingo) a 7 (sábado)
N_sem	Função Excel: =NÚMSEMANA(DATA(“ano”;”mês”;”dia”))	A função NÚMSEMANA devolve o número da semana do ano em questão. Assumindo valores entre 1 e 52/53

A Tabela 7 complementa a Tabela 3 (Atributos registados pela ferramenta de gestão de incidentes), no sentido em que detalha os novos atributos criados e sumariza os atributos a serem considerados para a amostra alvo do estudo.

Tabela 7: Descrição dos atributos da amostra

Atributo	Descrição
MesAvaria	Mês de registo do incidente
DiaAvaria	Dia de registo do incidente
Dia_sem	Criado como novo atributo, com dia da semana em que o incidente é registado.
N_sem	Criado como novo atributo, com o número da semana em que o incidente é registado.
Utilizador	Código do utilizador que regista o incidente
Area	Código que identifica a área do negócio (departamento) onde se encontra o utilizador que regista o incidente
Tipo_Inc	Código que identifica o tipo de incidente (e.g. Inc. nas aplicações, Inc. posto de trabalho, inc Segurança, etc.)
Servico	Código que identifica a aplicação ou ferramenta afetada (subtipo de Tipo_Inc) (e.g. aplicações de faturação, telefone do posto de trabalho, etc.)
Categoria	Código que identifica a funcionalidade da aplicação ou ferramenta afetada (subtipo de Servico) (e.g. funcionalidade X da aplicação de faturação, a opção Z no telefone, etc.)
Situacao	Código que identifica a anomalia identificada (subtipo de Categoria) (e.g. Serviço indisponível, quebra de performance, avaria no equipamento, dados incorretos, etc.
Prioridade	Código que classifica a prioridade de acordo com a Urgência e Impacto Identificados
Reportado_por	Código que identifica se o utilizador que registou o incidente é ou não um agente (equipa técnica)
Motivo	Código no qual o técnico regista no momento inicial de análise o motivo (aparente) do incidente (e.g. Indisponibilidade de serviço, desconhecido, comportamento inesperado, utilização incorreta)
Origem	Código no qual o técnico regista no momento inicial de análise a origem (aparente) do incidente (e.g., origem conhecida, desconhecida, problema de configuração, etc.)
Tempo_Res	Tempo de resolução (desde momento do registo até a resolução, em tempo útil)
Grupo *	Transformação da variável dependente de valores contínuos (Tempo_Res) em 4 classes discretos: 1H (resolução em menos de 1h), 1H4H (resolução em meio dia), 4H8H (resolução em 1 dia) e M1D (resolução superior a 1 dia).

* Atributo complementar a ser utilizado como variável dependente, em detrimento do atributo TempoRes

Sumariza-se de seguida as ações realizadas nas fases da compressão e preparação dos dados.

Após a análise de diversos diagramas de frequência e diagrama de extremos, a variável dependente Tempo_Res foi a única a apresentar valores discrepantes (*outliers*), em 71 registos, que por serem resultantes de incidentes com um tempo de resolução anormais (possivelmente erros), foram removidos do estudo.

Foram criados três novos atributos, com o objetivo de diversificar os atributos disponíveis e potenciar a diversidade de técnicas de *data mining* a utilizar no estudo.

No que respeita à qualidade dos dados, não foram identificados atributos omissos ou incorretos, de modo geral os dados extraídos estão coerentes, sendo a única exceção o atributo Area que apresentou 65 registos omissos, tendo sido removidos.

Assim e finalizado a fase de tratamento dos dados, parte-se para a fase de modelação com um total de 15 atributos (sendo que existem duas variáveis dependentes possíveis, uma numérica continua e outra discreta ordinal) e uma amostra com 43.870 registos.

3.5. CRISP-DM Fase 4: Modelação

Nesta fase procede-se à seleção e aplicação de várias técnicas de *data mining* para a obtenção de modelos preditivos, onde os seus parâmetros são ajustados por forma a otimizar o resultado (Chapman et al., 2000). Recorre-se a diferentes técnicas de regressão e classificação, tendo sido utilizadas quatro técnicas de aprendizagem, árvores de decisão, regressão linear, redes neuronais e máquinas de vetores de suporte.

Estas técnicas foram selecionadas devido às suas características. As árvores de decisão e regressão constituem um dos modelos mais usados em *data mining*, sendo uma das suas principais vantagens a facilidade de compreensão e interpretação. Já as redes neuronais, inspiradas na fisiologia do cérebro e nas complexas redes neuronais biológicas, geram o conhecimento através de um processo de aprendizagem que armazena informação sobre a forma de “pesos”, que são uma ponderação que afetará os valores produzidos por cada nó. Embora sejam não lineares e flexíveis, são de difícil interpretação. No que respeita às máquinas de vetores de suporte, estas são consideradas o estado atual da arte no que respeita a algoritmos de classificação e de regressão. Contudo exigem um maior processamento (logo mais lentas) e são igualmente de interpretação complexa. A regressão, linear exige pouco processamento e é de fácil compreensão e tem tido uma utilização generalizada.

Para testar a qualidade e validade do modelo foram utilizados os métodos de validação *holdout* e *K-folds*. O método *holdout* divide aleatoriamente os dados no conjunto de treino e no conjunto de teste. O conjunto de treino contêm 66% dos dados (~2/3) e servirá para estimar os parâmetros do modelo. O conjunto de teste contêm os restantes 33% (~1/3) e servirá para

avaliar a precisão do modelo. O método *K-folds*, com um funcionamento muito semelhante ao anterior, divide os dados em 10 partições de igual dimensão (*K*=10) e em cada execução é testado um determinado subconjunto, sendo os subconjuntos restantes utilizados para treinar o modelo. A qualidade do modele corresponde à média dos resultados das 10 rotações.

Na primeira iteração da fase de modelação, os dados foram carregados na ferramenta Weka, de notar que a amostra é constituída na sua maioria por instâncias com atributos numéricos (codificação interna da aplicação de gestão de incidentes) contudo para efeitos do presente estudo, os atributos devem ser considerados nominais (à exceção do tempo de resolução TempRes que é um atributo numérico continuo). Neste sentido e uma vez que o Weka nestas circunstâncias assume os atributos como numéricos, foi necessário converte-los em nominais, (Filtro *NumerictoNominal* no Weka). Tendo sido de seguida executados os diferentes algoritmos de aprendizagem.

Contudo esta abordagem demonstrou-se pouco eficaz, uma vez que a execução de alguns classificadores devorou mais de 10 horas, Este comportamento assume-se ser devido à quantidade de registos e atributos (nominais) em análise pelos modelos. Neste sentido assume-se que a primeira iteração será realizada com atributos numéricos, uma vez que a comparação dos modelos possíveis de executar (em tempo aceitável), não demonstrou variações acentuadas em termos de eficácia. A Tabela 8 ilustra o comparativo dos resultados obtidos com a execução das mesmas técnicas, com atributos nominais e com atributos numéricos, quando aplicados à amostra de teste.

Tabela 8: Comparativo de modelos com atributos numéricos e nominais

Técnica	Classificador WEKA	CC com Atributos Nominais	CC com Atributos Numéricos
Regressão Linear	function.LineraRegression	0,27	0,15
Árvores de Decisão	trees.REPTree	0,23	0,23

CC – Coeficiente de Correlação (r)

Revertidos os atributos para numéricos e assumindo o atributo TempoRes como a variável dependente (expressa em valores numéricos contínuos), foi executada as diferentes técnicas de regressão, conforme apresentado na Tabela 9. Esta primeira iteração e de modo a reduzir o processamento e acelerar a análise, foi utilizado método de validação *holdout* utilizando 66% dos registos para treino e os restantes 33% para testes.

Tabela 9: Resultado de técnicas de Regressão na amostra de teste

ID	Técnica	Classificador WEKA	CC	MAE	RMSE
R1	Regressão Linear	function.LineraRegression	**0,15**	337,58	741,83
R2	Redes Neuronais	function.MultilayerPerceptronn	**0,12**	865,93	996,62
R3	Árvores de Decisão	trees.RandomTree	**0,15**	357,11	997,08
R4	Árvores de Decisão	trees.REPTree	**0,23**	317,00	739,55

CC – Coeficiente de Correlação (r); MAE - Mean Absolute Error; RMSE - Root Mean Squared Error

Dado que os resultados obtidos foram pouco promissores (coeficiente de correlação inferiores a 0,23 e erros médios de mais de 300 horas), foi realizada uma a segunda iteração, baseando-se no mesmo processo de descoberta, contudo fazendo recurso da variável dependente Grupo, que substitui o anterior (TempoRes). À semelhança da iteração anterior mantiveram-se os atributos numéricos e foi novamente utilizado método de validação *holdout* (utilizando 66% dos registos para treino e os restantes 33% para testes), a Tabela 10 apresenta os resultados obtidos com as técnicas de Classificação.

Tabela 10: Resultado de técnicas de Classificação na amostra de teste

ID	Técnica	Classificador WEKA	ICC (%)	AUC
C1	Redes neuronais	function.MultilayerPerceptronn	**54,65**	0,62
C2	Maquinas Vetores	function.SMO	**54,36**	0,53
C3	Árvores de Decisão	trees.SimpleCart	**57,86**	0,65
C4	Árvores de Decisão	trees.J48	**52,98**	0,62
C5	Árvores de Decisão	trees.REPTree	**55,83**	0,65

ICC – Percentagem de Instâncias Corretamente Classificadas; AUC – Área abaixo da curva ROC

Embora as técnicas de classificação se tenham demonstrado mais promissoras (capacidade preditiva de 57,86%) do que as técnicas de regressão é necessário compreender a relevância da capacidade previsão destes modelos, tendo em conta a dispersão e equidade das 43.870 instâncias da amostra. Neste sentido foi utilizado o classificador ZeroR do Weka (rules.ZeroR) com o objetivo de definir uma *baseline* nos dados, isto é, demostrar a capacidade preditiva mais básica, baseada unicamente na análise da frequência de registos em cada uma das quatro classes da variável dependente Grupo. As Tabelas 11 e 12, demonstram respetivamente o resultado deste classificador e a respetiva matriz de confusão.

Tabela 11: Resultado do classificador ZeroR - Baseline da amostra de dados

Descrição	Quantidade	Percentagem
Instâncias Corretamente Classificadas	23.906	54,5 %
Instâncias Incorretamente Classificadas	19.964	45,5 %

Tabela 12: Matriz de confusão do classificador ZeroR

Observado	Previsto					
	1H	1H4H	4H8H	M1D	Corretamente Classificado	Distribuição Amostra
1H	0	0	0	9707	0,0%	22,1%
1H4H	0	0	0	6119	0,0%	13,9%
4H8H	0	0	0	4138	0,0%	9,4%
M1D	0	0	0	23906	100,0%	54,5%
Percentagem global	0,0%	0,0%	0,0%	100,0%	**54,5%**	**100%**

Conforme se pode constatar, este classificador executa simplesmente uma análise à distribuição dos dados de acordo com a classe (variável dependente Grupo).Embora simples, trata-se de uma análise importante. Como se pode constatar, a assimetria da distribuição dos registos pelas várias classes permite só por si criar um modelo empírico que permite prever o tempo de resolução em 54,5% dos casos. Uma vez que os modelos criados até o momento apenas permitem prever corretamente 57,86% dos casos, pode-se afirmar que comparativamente à *baseline* (54,5%) os modelos têm uma baixa capacidade preditiva (apenas 3% acima de uma previsão empírica), havendo inclusive um modelo que apresenta resultados inferiores a esta *baseline* (Arvore decisão J48 - ID C4 - com uma eficácia de 52,98%).

Com base nestes resultados optou-se por identificar quais os atributos com menor contributo para a previsão do tempo de resolução tendo em vista reduzir a dimensionalidade do problema em estudo e viabilizar a utilização de outras técnicas analíticas, ao mesmo tempo que se simplifica o modelo final. Contudo, pretende-se que esta redução não implique compromisso da capacidade preditiva. Neste sentido recorreu-se ao SPSS para criar o modelo de classificação T e conhecer a importância para o modelo das variáveis independentes face à variável dependente Grupo (Tabela 13). O modelo selecionado foi a árvores de decisão, uma vez que foi o modelo que demonstrou melhor capacidade de previsão, tendo sido utilizado o algoritmo CRT dada a sua semelhança ao classificador SimpleCART utilizado no Weka. A Tabela 14 mostra matriz de confusão do modelo CRT, com 57,1% de Instâncias Corretamente

Classificadas, muito semelhante aos 57,89% obtidos com o SimpleCART no Weka (Tabela 10).

Tabela 13: Importância das Variáveis Independentes

Importância das Variáveis Independentes

Variáveis Independentes	Importância	Importância Normalizada
TipoInc	,012	100,0%
Motivo	,011	87,4%
Origem	,008	63,8%
Servico	,004	32,8%
Prioridade	,004	31,1%
Categoria	,003	23,2%
Situacao	,001	11,1%
DiaAvaria	,001	8,9%
ReportadoPor	,001	8,7%
MesAvaria	,001	6,2%
N_sem	,001	5,5%
Dia_sem	,000	3,0%
Utilizador	,000	1,5%
Area	,000	1,3%

Método utilizado: CRT (SPSS)

Variável Dependente: Grupo

Tabela 14: Matriz de confusão

	Previsto				
Observado	1H	1H4H	4H8H	M1D	**Corretamente Classificado**
1H	1895	0	0	7812	19,5%
1H4H	473	0	7	5639	,0%
4H8H	277	0	9	3852	,2%
M1D	750	0	0	23156	96,9%
Percentagem global	7,7%	,0%	,0%	92,2%	**57,1%**

Após a análise da relevância das variáveis independentes, optou-se por considerar os atributos com um valor de importância inferior a 10%, como candidatos a serem removidos da amostra. Neste sentido e como forma de comprovar a baixa relevância destes atributos para o modelo, foi realizado um exercício de numa abordagem de tentativa/erro, para tal foi utilizado a técnica de Árvore de Decisão de Classificação e Regressão do Weka onde o classificador SimpleCart

foi executado repetidamente com diferentes combinações dos atributos a avaliar. Para este exercício foi utilizado o método de avaliação *K-folds* com K=10, com atributos numéricos. A Tabela 15 demonstra as várias iterações realizadas, onde a coluna ICC apresenta a percentagem das Instâncias corretamente classificadas pelo classificador, a coluna AUC a área da curva ROC, as colunas intermédias apresentam os 15 atributos iniciais, identificados com "X" aqueles que foram utilizados em cada uma das iterações.

Tabela 15: Exercício de avaliação da relevância de atributos

ID	MESAVARI	DIAAVARIA	DIA_SEM	N_SEM	UTILIZADO	AREA	TIPOINC	SERVICO	CATEGORI	SITUACAO	PRIORIDAD	REPORTAD	MOTIVO	ORIGEM	GRUPO	ICC (%)	AUC
I1	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**58,12**	**0,66**
I2					X	X	X	X	X	X	X	X	X	X	X	**58,09**	0,62
I3				X	X	X	X	X	X	X	X	X	X	X	X	**58,19**	0,66
I4	X	X	X	X			X	X	X	X	X		X	X	X	**57,80**	0,65
I5				X	X	X	X	X	X	X	X		X	X	X	**57,90**	0,65
I6				X		X	X	X	X	X	X	X	X	X	X	**58,12**	0,66

Média	58,02
Desvio Padrão	0,15

ICC- Instâncias corretamente classificadas; AUC – Área abaixo da curva

A primeira iteração (I1) fez recurso de todos os atributos e serve como referencial (*baseline*) para o exercício. As restantes iterações vão avaliar o sucesso ou insucesso (incremento ou decréscimo) da percentagem de ICC e valor AUC com as diferentes combinações de atributos relativamente ao referencial (I1). Neste sentido para o primeiro teste (I2), foram retirados os atributos relacionados com a dimensão "tempo" (datas) e comparativamente ao referencial constatou-se uma redução muito residual de -0,03%. Com base neste resultado confirma-se uma baixa relevância destes atributos para o modelo a criar. Contudo e dada uma eventual redundância nestes atributos (uma vez que Dia_sem e N_sem foram criados a partir do Mesavaria e Diaavaria) optou-se por uma nova iteração (I3) apenas com o atributo N_sem como representante da dimensão "tempo", com um resultado de 58,19%, o que representa um incremento de 0,07% fase ao referencial. A iteração I4 testou a não utilização dos atributos relacionados com a dimensão "organização" (Utilizador, Area e ReportadoPor) resultando uma redução de -0,32% face ao referencial. As restantes iterações (I5 e I6) testaram diferentes combinações de atributos desta dimensão resultando em reduções de -0,22% e -0,01% respetivamente face ao referencial.

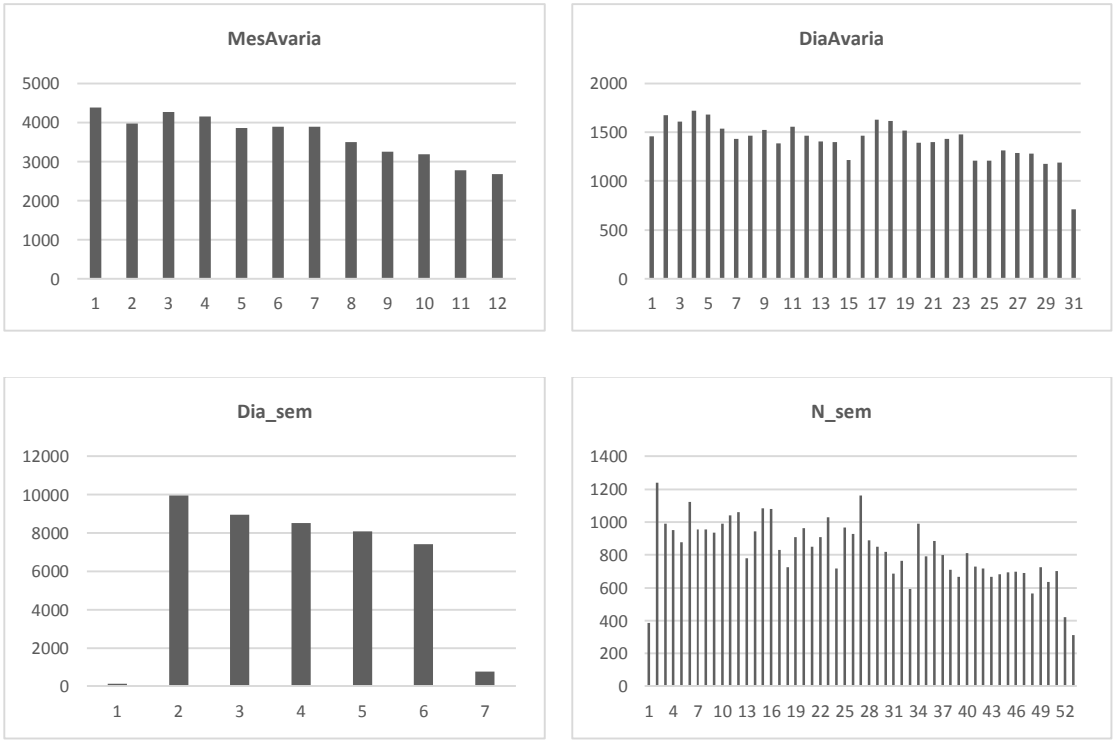
Em suma este exercício permitiu confirmar uma baixa relevância destes atributos para o modelo preditivo a criar, uma vez que retirando os atributos relacionados com tempo e com a organização a capacidade preditiva do classificador utilizado, não alterava de forma acentuada (desvio-padrão de 0,15).

Neste sentido optou-se por repetir o passo 3 Preparação dos Dados, da metodologia CRISP--DM (confirmando-se deste modo a natureza iterativa da metodologia). Pretende-se com esta nova iteração avaliar, com base na informação recolhida no passo de modelação, os atributos a considerar na amostra com o objetivo de potenciar a exploração do maior número possível de técnicas de *data mining* e obter um modelo preditivo que responda aos objetivos definidos.

2ª Iteração CRISP-DM: Passo 3: Preparação dos Dados

Face à resolução de reavaliar a pertinência dos atributos com menor importância para o modelo preditivo (Tabela 15), foram analisados os atributos: DiaAvaria, ReportadoPor, MesAvaria, N_sem, Dia_sem, Area e Utilizador. Neste sentido e para simplificar a análise, os atributos foram divididos em dois grupos: "Tempo" com os atributos relacionados com a esta dimensão; DiaAvaria, MesAvaria, N_sem e Dia_sem e "Organização" com os atributos; ReportadoPor, Utilizador e Area. As Figuras 22 e 23 ilustram respetivamente a tabela de frequência destes atributos.

Figura 22: Tabela de frequência dos atributos temporais

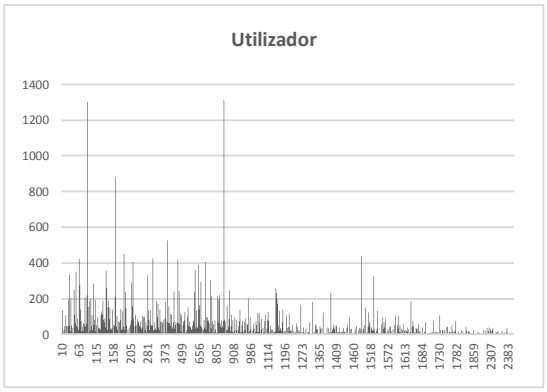


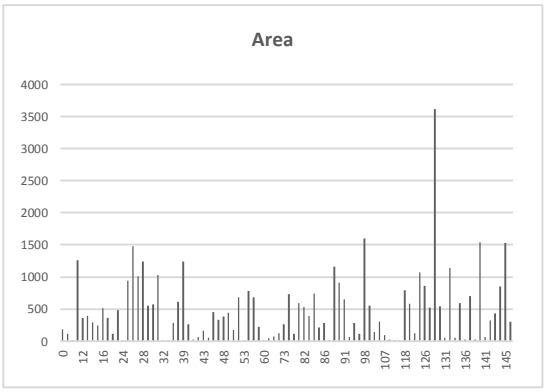
Pela análise da frequência dos registos de cada atributo, pode-se constatar que numa visão anula (atributos MesAvaria e N_sem) existe um maior número de registo de incidentes nos primeiros meses/semanas do ano, com um decréscimo ao longo do ano (-37% entre Jan. e Dez.). De igual forma, numa visão mensal (atributos Dia_sem e N_sem) semelhante comportamento, um maior número de incidentes nos primeiros dias e uma redução de 29% comparativamente aos últimos dias do mês.

Neste sentido, pode-se afirmar que do ponto de vista de negócio existe uma tendência para uma assimetria no numero de incidentes ao longo do ano e dos meses, pelo que as equipas de suporte e resolução de incidentes devem ser reforçadas nestes períodos (primeiros meses do ano e inícios de cada mês), contudo embora se possa afirmar que o volume de incidentes a resolver varie em função da data, não existem evidências para afirmar que este aumento no volume de trabalho, influencie o tempo de resolução, uma vez que foram classificadas com baixa importância na explicação da variável dependente (Tabela 13) e a sua não utilização apenas reduz a capacidade preditiva em 0,03% (Tabela 15). Contudo o atributo N_sem revelou alguma importância, fazendo inclusive aumentar a capacidade preditiva em 0,07%.

Neste sentido optou-se por remover da amostra a generalidade dos atributos que ilustram a dimensão tempo, com a exceção do atributo N_sem, que permanece como representante desta dimensão, uma vez que revelou ser uma mais-valia par o modelo.

Figura 23: Tabela de frequência dos atributos relacionados com Organização





Pela análise da frequência dos registos relacionados com organização, constata-se que existe na amostra uma elevada quantidade de utilizadores (870) e áreas (90). Contudo, ambas com baixa frequência de registos (poucos incidentes registados pelo mesmo utilizador e pela mesma área), embora o atributo Area registe maiores. Esta realidade deve-se ao fato de existir alguma rotatividade de utilizadores, externa (admissões e saídas) e internas (mobilização entre áreas), o que faz com que exista uma baixa recorrência de registo de incidentes pelo mesmo utilizador.

O atributo ReportadoPor, identifica incidentes reportados por agentes das equipas técnicas de suporte (A) e pelos utilizadores (U), como seria de esperar o volume de incidentes registados pelo negócio (utilizadores) e consideravelmente superior ao registo interno de incidentes (Agentes da equipa técnica). Embora os atributos desta dimensão tenham sido identificados como pouco relevantes para o modelo (Tabela 13), foram aqueles que no exercício de avaliação da relevância de atributos (Tabela 15) demonstraram que a sua não utilização, representa uma redução da capacidade preditiva (-0,32%). Neste sentido optou-se por manter estes atributos na amostra a utilizar no projeto. A Tabela 16 apresenta os atributos resultantes da amostra que serão considerados nas fases seguintes da metodologia e descoberta dos dados.

Tabela 16: Descrição dos atributos da amostra (2ª Iteração)

Atributo	Descrição
N_sem	Número da semana em que o incidente é registado.
Utilizador	Identifica o utilizador que regista o incidente
Area	Identifica a área do negócio que regista o incidente
Tipo_Inc	Identifica o tipo de incidente (aplicações, PC, etc.)
Servico	Identifica a subcategoria do tipo de incidente
Categoria	Identifica a subcategoria do Servico
Situacao	Identifica a subcategoria do Categoria
Prioridade	Identifica a prioridade
Reportado_por	Identifica o perfil do utilizador (técnico ou negócio)
Motivo	Identifica o motivo do incidente na análise por parte do técnico
Origem	Identifica o motivo do incidente na análise por parte do técnico
Grupo	Classifica o tempo de resolução 4 grupos

Desta forma, parte-se para a 2ª iteração da fase de modelação, com um total de 12 atributos (note-se que o atributo TempoRes foi substituído pelo atributo Grupo como variável dependente, uma vez que as técnicas de regressão se demonstraram pouco eficientes na identificação de padrões na amostra) e uma amostra com 43.870 registos.

2ª Iteração CRISP-DM: Passo 4: Modelação

A segunda iteração da fase de modelação, seguiu a mesma abordagem que a primeira, contudo baseando-se na nova amostra de onde foram retirados 3 dos atributos originais.

Mantiveram-se os atributos numéricos e o método de validação *holdout* (utilizando 66% dos registos para treino e os restantes 33% para testes). A Tabela 17 apresenta um comparativo entre os resultados obtidos com as técnicas de Classificação na nova amostra com 12 atributos (identificados na coluna - ICC 2ª Iteração) e os resultados obtidos com a amostra inicial com

15 atributos (identificados na coluna - ICC 1ª Iteração), ambas relativamente à amostra de teste.

Tabela 17: Comparativo das técnicas de Classificação na amostra de testes

ID	Técnica	Classificador WEKA	ICC 2ª Iteração (%)	ICC 1ª Iteração (%)	Delta (%)
D1	Redes neuronais	function.Multilayer Perceptronn	**54,80**	54,65	0,15
D2	Maquinas Vetores	function.SMO	**54,36**	54,36	0
D3	Árvores de Decisão	trees.SimpleCart	**57,99**	57,86	0,13
D4	Árvores de Decisão	trees.J48	**54,92**	52,98	1,94
D5	Árvores de Decisão	trees.REPTree	**56,50**	55,83	0,67

ICC – Instâncias Corretamente Classificadas

No geral todas as técnicas revelaram uma melhoria nos resultados comparativamente à iteração anterior, tendo o seu resultado médio melhorado em 0,53%, sendo que o classificador J48 (a implementação do algoritmo C4.5 no Weka) foi o que demonstrando a melhor evolução, com uma melhoria de quase 2 pontos percentuais. Embora melhor, a capacidade preditiva dos modelos continua baixa tendo em conta a *baseline* definida (ICC de 54,49%). Neste sentido a análise anterior foi repetida desta vez alterando o método de avaliação para *Cross-Validation (K-Fold*). Com K=10, visando uma melhor estimativa do desempenho do modelo. Tabela 18 mostra o comparativo das duas técnicas de validação.

Tabela 18: Comparativas técnicas de validação

ID	Técnica	Classificador WEKA	ICC *Holdout* (%)	ICC *10-Fold* (%)	Delta (%)
E1	Redes neuronais	function.Multilayer Perceptronn	54,80	54,82	0,02
E2	Maquinas Vetores	function.SMO	54,36	54,49	0,13
E3	Árvores de Decisão	trees.SimpleCart	57,99	58,19	0,20
E4	Árvores de Decisão	trees.J48	54,92	55,15	0,23
E5	Árvores de Decisão	trees.REPTree	56,50	56,93	0,43

ICC – Percentagem de Instâncias Corretamente Classificadas

Como esperado o método de validação *Cross-Validation*, embora com um tempo de execução mais longo, comprova a maior precisão na previsão do desempenho do modelo, tendo o resultado melhorado em média 0,19% em todas os classificadores. Contudo e uma vez mais não foram registados incrementos significativos na capacidade preditiva, apenas o classificador SimpleCart demonstra algum potencial, assumindo uma eficácia de 3,75% acima da *baseline* (definida em 54,45%).

Nesta fase e de modo a descartar a possibilidade dos fracos resultados serem causados pelo fato dos classificadores estarem a assumir os atributos como numéricos (o que poderia dar ao classificador uma incorreta interpretação dos dados e consequentemente afetar a execução do algoritmo), os atributos foram convertidos em nominais e foram executados novamente os classificadores.

A Tabela 19 apresenta o comparativo entre a execução do mesmo classificador com atributos nominais e numéricos. Uma vez que nem todos os classificadores puderam ser executados devido ao seu tempo de execução, e de modo a apresentar um comparativo eficaz são apresentados (Tabela 19) os resultados para método de validação *holdout* e *Cross-Validation* (*K-Fold*).

Tabela 19: Comparativo dos classificadores (atributos numéricos e nominais)

			Atributos Numéricos		**Atributos Nominais**	
ID	**Técnica**	**Classificador WEKA**	**ICC *Holdout* (%)**	**ICC *10-Fold* (%)**	**ICC *Holdout* (%)**	**ICC *10-Fold* (%)**
E1	Redes neuronais	function.Multilayer Perceptronn	54,80	54,82	N/A	N/A
E2	Maquinas Vetores	function.SMO	54,36	54,49	N/A	N/A
E3	Árvores de Decisão	trees.SimpleCart	57,99	58,19	57,98	58,48
E4	Árvores de Decisão	trees.J48	54,93	55,15	54,36	54,49
E5	Árvores de Decisão	trees.REPTree	56,49	56,93	54,33	55,04

ICC – Instâncias Corretamente Classificadas; N/A – Valor não calculado devido ao elevado tempo de execução do modelo (>10h).

Pode-se firmar que os classificadores relacionados com árvores de decisão no Weka não sofrem grande influência pelo tipo de classificador (numérico ou nominal), este exercício denota inclusive um decréscimo na eficácia dos classificadores J48 e REPTree, quando executados com atributos nominais.

Contudo, mantêm-se a baixa eficácia dos classificadores. Neste contexto foi analisado um último fator que poderá contribuir para o seu fraco desempenho, o fator de sobreajuste do modelo (do inglês *Overfitting).* Segundo Tetko (1995), este fator ocorre quando um modelo descreve erros aleatórios ou ruído, em vez de a relação subjacente entre os dados e geralmente ocorre quando um modelo é excessivamente complexo (Muitas atributos e poucas instâncias). Sendo que o maior problema do *Overfitting* é gerar modelos com baixos desempenhos preditivos, uma vez que o modelo pode acentuar pequenas flutuações nos dados e criar um modelo demasiado adaptado aos dados de treino, comprometendo assim a sua capacidade preditiva com os dados de teste ou reais.

As técnicas para evitar o *Overfitting* passam sempre por conter o crescimento da árvore, sendo o ponto ideal o compromisso entre a capacidade do modelo conseguir generalizar os padrões existentes nos dados, sem o deixar "aprender" todas as relações existentes nesses dados, o que compromete a eficácia desse modelo em novos dados. A limitação do crescimento das árvores passa por; definir à partida, a dimensão máxima da árvore (limitando o seu crescimento) ou, por via de técnicas de eliminação de nós pouco relevantes, conhecida por *Pruning* (ou poda).

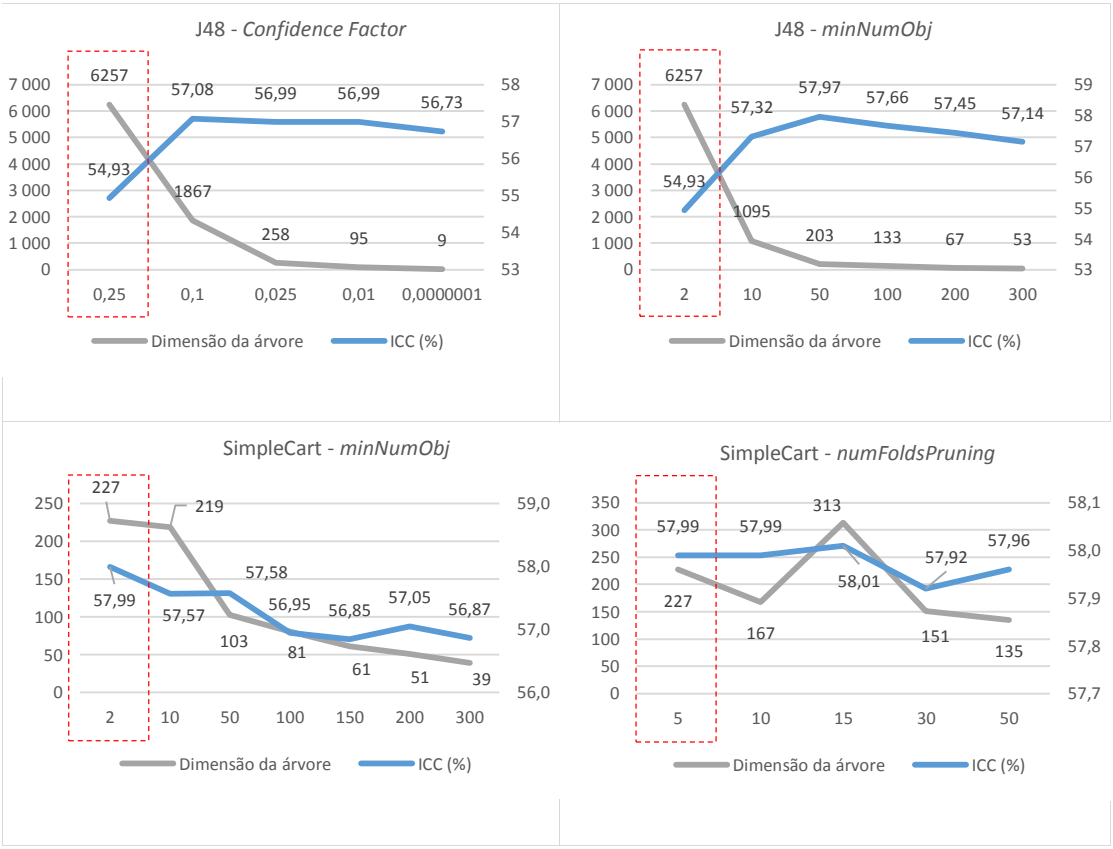
Pruning é uma técnica para reduzir o tamanho das árvores de decisão, removendo secções da árvore que fornecem baixa capacidade de classificar instâncias. Duas técnicas comuns de *Pruning são* o *Pre-Pruning* e o *Post-Pruning.* (Song, 2015) *Pre-Pruning* trava o crescimento da árvore antes desta classificar toda a amostra e *Post-Pruning* permite que a árvore classifique corretamente toda a amostra e depois elimina os nós de menor relevância. Ambas as técnicas estão implementas no Weka (Drazin et al., 2012). O *Post-Pruning* através da opção *Confidence Factor* e funciona numa ótica de *bottom-up* (de baixo para cima), eliminando nós com menor significância estatística. Já o *Pre-Pruning* (ou *Online-Pruning*) é implementado no Weka através da opção *minNumObj,* e é executado em paralelo com a geração da árvore de decisão, fundindo nós pais e nós filhos sempre que os nós filhos representem valores abaixo de um mínimo de exemplos nos dados.

Neste sentido e de modo a garantir que a eficácia dos modelos gerados não é afetada por *Overfiting*, foram testadas técnicas de limitação de crescimento e de *Pruning,* no sentido de se verificar se estas conseguem otimizar o desempenho dos classificadores e melhorar a capacidade do modelo preditivo. A Figura 24 apresenta graficamente o comparativo do resultado entre duas das árvores de decisão mais comuns do Weka (J48 e SimpleCart), onde foram testados vários valores das opções *Confidence Factor* e *minNumObj* com o objetivo de testar o desempenho das técnicas de *Post-Pruning* e *Online-Pruning,* respetivamente.

De notar que para o classificador SimpleCart não existe a opção de *Confidence Factor,* (*Post-Pruning)* logo foi testada a opção *numFoldsPruning* (uma técnica de *Pruning* por via de *Cross-Validation* interna do classificador).

No Eixo principal (à esquerda) é apresentado a dimensão da árvore (em numero total de nós e folhas) ao longo dos vários valores testados de *Pruning* e no eixo secundário (à direita), são apresentados os resultados do classificador em percentagem (ICC – Instâncias corretamente classificadas) para os diferentes valores de *Pruning.* É de realçar que o primeiro registo apresentado em cada gráfico (identificado por retângulo tracejado) representa os valores de *Pruning* por omissão (por *default*) na ferramenta para cada classificador, considerados como a *baseline* (o referencial) neste comparativo. Para o exercício foram utilizados atributos numéricos (dado que demonstra vantagem face à utilização de atributos nominais Tabela 19) e validação por *holdout*. (66% dos registos para treino e os restantes 33% para testes).

Figura 24: Comparativo de técnicas de Pruning em árvores de decisão



Conforme se pode constatar o classificador J48 demonstrou uma melhoria com ambas as técnicas, dado que a simples alteração dos parâmetros por omissão que limitam o crescimento da árvore, permitiram um acentuado decréscimo na sua dimensão (aprox. -70%), e em proporção inversa um incremento na eficácia do modelo (aumento no ICC de aprox. 5%). Em ambas as técnicas esta proporção de ordem inversa tende a esbater-se, uma vez que a partir de um certo valor, a redução na dimensão da árvore implica perda na eficácia preditiva do modelo. Já o classificador SimpleCart, demonstra menos apetência ao *Pruning*, dado que ambas as técnicas utilizadas representaram uma redução imediata na capacidade preditiva do modelo. Contudo fica implícito uma maior capacidade nativa de *Pruning* no classificador

SimpleCart do Weka (comparativamente ao J48), uma vez que com os valores por omissão, o J48 apresenta uma árvore com 6.257 nós, enquanto o SimpleCart apresenta uma arvore substancialmente mais reduzida com 227 nós e uma eficácia de 58%, valor este que não foi possível alcançar com o J48, mesmo após aplicação de várias técnicas para evitar o *Overfiting*.

Uma vez que o classificador SimpleCart não apresentou melhoria na sua eficácia por via das técnicas de *Pruning* pode-se afirmar que este classificador não sofria de *Overfiting* (eventualmente devido à sua capacidade nativa de *Pruning*), já o classificador J48 demonstrou uma melhoria considerável na eficácia à medida que a dimensão da árvore era reduzida, tornando-o um forte candidato a sofrer de *Overfiting* e como potencial para melhorar a sua eficácia preditiva. Uma vez que o teste anterior foi realizado com *holdout* como método de validação é imperativo reavaliar o classificador com o método de validação mais eficaz, o *Cross-Validation (K=10)*. A Tabela 20 apresenta o resultado desta avaliação.

Tabela 20: Técnicas para evitar Overfiting no classificador J48 com 10-Fold

ID	Classificador Weka	Método Validação	*Confidence Factor*	*minNumObj*	ICC (%)
F1	trees.J48	*holdout*	Default (0,25)	Default (2)	54,93
F2	trees.J48	*Cross-Validation* K=10	Default (0,25)	Default (2)	55,15
F3	trees.J48	*Cross-Validation* K=10	0,1	Default (2)	57,84
F4	trees.J48	*Cross-Validation* K=10	Default (0,25)	50	58,02
F5	trees.J48	*Cross-Validation* K=10	0,1	50	57,92

ICC – Percentagem Instâncias Corretamente Classificadas

Confirma-se portanto que a aplicação de técnicas para evitar o *Overfiting* melhoram o desempenho do classificador J48, uma vez que se obteve um incremento de 3% (melhor resultado obtido F4, relativamente ao referencial F1) na capacidade preditiva do modelo. Embora a validação *10-Fold* tenha melhorado o desempenho, continua abaixo do resultado obtido com o classificador SimpleCart (58,48%).

A título de resumo da fase de modelação, a Tabela 21 apresenta uma síntese (por ordem decrescente) dos resultados obtidos com cada um dos classificadores testados e respetivas técnicas de otimização utilizadas.

Tabela 21: Resumo da classificação da classe de tempo

ID	Técnica	Classificador Weka	*Método Validação*	*Técnicas Pruning*	ICC (%)	AUC
R1	Árvores de Decisão	trees.SimpleCart	*Cross-Validation* K=10	Default	58,48 *	0,67
R2	Árvores de Decisão	trees.J48	*Cross-Validation* K=10	M=50	**58,02**	0,66
R3	Árvores de Decisão	trees.REPTree	*Cross-Validation* K=10	Default	**56,93**	0,67
R4	Redes neuronais	function.Multilayer Perceptronn	*Cross-Validation* K=10	Default	**54,82**	0,62
R5	Maquinas Vetores	function.SMO	*Cross-Validation* K=10	Default	**54,49**	0,50

ICC – Percentagem de Instâncias Corretamente Classificadas; AUC – Área abaixo da curva ROC; M – minNumObj; * - Valor alcançado com atributos nominais (note-se que foi a única técnica a demonstrar melhoria nos resultados com a utilização de atributos nominais)

A técnica de árvore de decisão com recurso ao classificador Weka SimpleCart, foi o que apresentou os melhores resultados e a melhor capacidade preditiva com 58,48% das instâncias corretamente classificadas e com uma área da curva ROC de 0,67. A Tabela 22 apresenta a matriz de confusão deste classificador, relativamente à amostra de teste.

Tabela 22: Matriz Confusão classificador SimpleCart, da amostra de teste

	Previsto				
Observado	1H	1H4H	4H8H	M1D	**Corretamente Classificado**
1H	4077	64	5	5561	42,0%
1H4H	1450	69	3	4597	1,1%
4H8H	806	37	4	3291	0,1%
M1D	2321	78	2	21505	90,0%
Percentagem global	19,7%	0,6%	0,0%	79,7%	**58,48%**

Pela análise da matriz de confusão do classificador, pode-se observar que o modelo é fortemente influenciado pela distribuição assimétrica entre as classes, conforme constatado na definição da *baseline* da amostra de dados (Tabela 12) a classe M1D contem 54% dos registos da amostra o que influência o modelo, dando-lhe uma tendência para prever tempos de resolução superiores a 1 dia (M1D), o que indicia que uma parte significativa dos 58,48% de eficácia do classificador SimpleCart, é explicada pela assimetria da amostra, comprovado

pelos 90% de instâncias corretamente classificadas na classe M1D, comparativamente a apenas 43,2% no conjunto das restantes classes (1H, 1H4H e 4H8H).

De igual forma a análise da matriz de confusão permite concluir que a quantidade de instâncias corretamente classificadas das 2 classes intermédias (1H4H e 4H8H) é muito baixo, assim como a sua proporção na amostra (14% e 9% respetivamente, Tabela 12).

Face a estas duas constatações optou-se por re-testar o classificador que demonstrou o melhor desempenho (SimpleCart) com duas novas amostras que atenuem o efeito dos 2 fatores identificados (assimetria entre classes e baixa capacidade preditiva das classes intermédias) e que podem estar a comprometer a qualidade dos modelos. Apresenta-se de seguida uma sistematização do problema e possíveis soluções:

- Reduzir a propensão do modelo para prever tempos mais longos:
    - Possível solução: Nivelar distribuição da amostra equitativamente entre as 4 classes (remover aleatoriamente instâncias das classes com maiores frequências)
- Potenciar a eficácia do modelo na previsão das classes de tempo de resolução intermédias (1H4H e 4H8H)
    - Possível solução 1: Fundir as duas classes intermédias numa única (1D – Até 1dia)
    - Possível solução 2: Nivelar distribuição da amostra equitativamente entre as 3 classes (remover aleatoriamente instâncias das classes com maiores frequências)

Neste sentido foi realizada uma nova iteração na metodologia CRISP-DM, com o objetivo de gerar duas novas amostras que permitam re-testar o classificador.

3ª Iteração CRISP-DM: Passo 3: Preparação dos Dados

Com o objetivo de testar o classificador SimpleCart com duas novas amostras que permita atenuar a distribuição assimétrica de registos pelas classes e que potenciem a relevância das duas classes intermédias (1H4H e 4H8H), foram extraídas da amostra principal 3 subamostras, a Tabela 23 apresenta as 3 subamostras geradas, respetiva frequências por classe e total de registos, comparativamente à amostra original (amostra padrão).

Tabela 23: Resumo das subamostras geradas

	Amostra Padrão		SubAmostra Nivelados	
Variável Dependente Grupo	**Num. registos**	**%**	**Num. registos**	**%**
1H	9707	22%	4818	26%
1H4H	6119	14%	4634	25%
4h8H	4138	9%	4138	23%
M1D	23906	54%	4751	26%
Total	**43870**	**100%**	**18341**	**100%**

	SubAmostra 3 Classes		SubAmostra 3 Classes Nivelado	
Variável Dependente Grupo	**Num. registos**	**%**	**Num. registos**	**%**
1H	10257	23%	10257	32%
1D	9707	22%	9707	30%
M1D	23906	54%	11888	37%
Total	**43870**	**100%**	**31852**	**100%**

Foi utilizado os seguintes método para a extração das subamostras:

- SubAmostra Nivelados: Com recurso ao Excel foram selecionadas de modo aleatório (função *random*) registos das classes 1H, 1H4H e M1D de modo a nivelar os registos entre as 4 classes.
- SubAmostra 3 Classes: Com recurso ao Excel as classes 1H4H e 4H8H foram substituídas pela classe 1D.
- SubAmostra 3 Classes Nivelado: Tendo por base a subamostra 3 Classes e com recurso ao Excel foram selecionadas de modo aleatório (função *random*) registos das classes M1D de modo a nivelar os registos entre as 3 classes.

Com estas novas amostras, parte-se para a fase de modelação da 3º iteração da metodologia CRISP-DM.

3ª Iteração CRISP-DM: Passo 4: Modelação

A Tabela 24 apresenta o resultado da execução do classificador SimpleCart com as 3 subamostras geradas**.** Note-se que por razões relacionadas com o tempo de execução, o comparativo da Tabela 24, considera atributos numéricos e validação por *holdout* (66%/33%)

Tabela 24: Comparativo da eficácia do classificador SimpleCart com 3 subamostras

ID	Amostra	ICC (%)	AUC
H1	Padrão	**58,19**	0,66
H2	Nivelada	**35,90**	0,61
H3	3 Classes	**58,14**	0,66
H4	3 Classes nivelada	**48,27**	0,65

ICC – Percentagem de Instâncias Corretamente Classificadas; AUC – Área abaixo da curva ROC

Conforme se pode constatar não se obteve melhorias nos resultados, confirmando-se que o resultado de 58,19% do classificador SimpleCart é fortemente influenciado pela assimetria de registos nas várias classes, uma vez que nos dois cenários em que se nivelam os registos pelas várias classes (ID H2 e H4) a eficácia do modelo decrece abaixo dos 50%. A redução das classes, embora aumente a eficácia na nova classe intermédia (1D), não melhora a capacidade preditiva (ICC do ID-H3 inferior ao padrão), o que leva a concluir que as alterações efetuadas na amostra (criação de subamostras) em nada beneficiam a capacidade preditiva do modelo.

Neste cenário e perante os fracos resultados obtidos com os classificadores disponíveis no Weka, optou-se por explorar os algoritmos de aprendizagem disponíveis no *software comercial IBM SPSS Statistics* (SPSS).

Neste sentido e de modo a manter um fio condutor com o trabalho realizado até ao momento com o Weka, mantêm-se nesta fase de exploração de dados via SPSS, as mesmas premissas em termos de: classificadores, amostra e métodos de validação. Assim são analisados os modelos de árvore de decisão CRT, CHAID e Exaustive CHAID, caracterizados em detalhe na Tabela 25 comparativamente a alguns dos classificadores do Weka. Considera-se a amostra original, constituída pelos 12 atributos definidos na 2ª iteração do CRISP-DM, descritos na Tabela 26 juntamente com a respetiva tipologia. Como método de validação é utilizado o *Cross-Validation com K=10.* A Tabela 27 apresenta o resultado da primeira iteração realizada no SPSS com os 3 modelos de árvores de decisão.

Tabela 25: Descritivo dos algoritmos de Arvores de decisão

Métodos	**CART (Classification And Regression Tree)**	**C4.5**	***CAHID (Chi-squared Automatic Interaction Detection)***
Implementação no Weka	**SimpleCart**	***J48***	**Não Implementado**
Implementação no SPSS	**CRT**	**Não Implementado**	**CAHID e Exhaustive CAHID**
Técnicas de *Pruning*	*Post-Pruning*	*Post-Pruning e Pre-Pruning*	*Pre-Pruning*, via teste de independência do Qui-quadrado
Variáveis Dependentes	Categórica / Continua	Categórica / Continua	Categórica
Variáveis de entrada	Categórica / Continua	Categórica / Continua	Categórica / Continua
Método de divisão de nós	Binária; divisão em combinações lineares	Múltiplas divisões	Múltiplas divisões

Tabela 26: Atributos analisados no SPSS

Atributo	**Tipo**
N_sem	Ordinal
Utilizador	Nominal
Area	Nominal
Tipo_Inc	Nominal
Servico	Nominal
Categoria	Nominal
Situacao	Nominal
Prioridade	Ordinal
Reportado_por	Nominal
Motivo	Nominal
Origem	Nominal
Grupo	Nominal – Variável Dependente

Tabela 27: Árvores de decisão no SPSS (amostra de teste)

ID	Técnica	Método SPSS	*Método Validação*	ICC (%)
S1	Árvore de Decisão	CRT	*Cross-Validation* K=10	**57,0**
S2	Árvore de Decisão	CHAID	*Cross-Validation* K=10	**57,4**
S3	Árvore de Decisão	EXAUSTIVE CHAID	*Cross-Validation* K=10	**55,6**
S4	Redes Neuronais	Multilayer Perceptron	*Default*	**59,0**

ICC – Percentagem de Instâncias Corretamente Classificadas

A primeira execução dos métodos de aprendizagem no SPSS demonstrou que as árvores de decisão obtiveram resultados inferiores aos seus homólogos no Weka, contudo as redes neuronais (ID-S4) demonstraram algum potencial obtendo um eficácia de 59%, o melhor valor obtido até ao momento.

Perante estes resultados optou-se por aprofundar a análise dos dois modelos com o melhor resultados o CHAID e *Multilayer Perceptron*.

A Tabela 28 apresenta alguns dos testes realizados no modelo CHAID, relativamente ao *Overfitting* mediante as parametrizações disponíveis no SPSS para controlo do crescimento das árvores e de *Pruning*. Como limitador de crescimento o SPSS disponibiliza no tabulador *Growth Limits* (Opção *Criteria*) a customização dos parâmetros *Maximum Tree Depth* e *Minimum Number of Cases*. O primeiro (coluna MAXDEPTH) permite limitar o número de níveis na árvore, e os segundos (colunas MINPARENTSIZE e MINCHILDSIZE) permitem controlar o número mínimo de casos por cada nó pai e filho respetivamente.

Uma vez que o algoritmo CHAID e a sua variante Exhaustive CHAID apenas utilizam técnicas de *Pre-Puning*, este é realizado durante a execução do algoritmo, onde o teste do Qui-quadrado é utilizado para limitar a divisão dos nós e, consequentemente limitar a dimensão da árvore de decisão. A customização do parâmetro *Significance Level* (coluna CHAID ALPHASPLIT) no tabulador CHAID, permite definir o critério que controla esta divisão dos nós.

Tabela 28: Customizações dos parâmetros de crescimento no SPSS

ID	Método SPSS	MAXDEPTH	MINPARENTSIZE	MINCHILDSIZE	CHAID ALPHASPLIT	ICC (%)
P1	CHAID	AUTO (3)	Default (100)	Default (50)	Default (0,05)	**57,4**
P2	CHAID	AUTO (3)	50	25	Default (0,05)	**57,4**
P3	CHAID	2	Default (100)	Default (50)	Default (0,05)	**56,7**
P4	CHAID	5	Default (100)	Default (50)	Default (0,05)	**57,7**
P5	CHAID	AUTO	Default (100)	Default (50)	0,001	**57,3**

ICC – Percentagem de Instâncias Corretamente Classificadas; MAXDEPTH - Profundidade máxima da árvore; MINPARENTSIZE - Numero mínimo de casos para nó Pai; MINCHILDSIZE - Numero mínimo de casos para nó Filho; CHAID ALPHASPLIT – Nível de significância

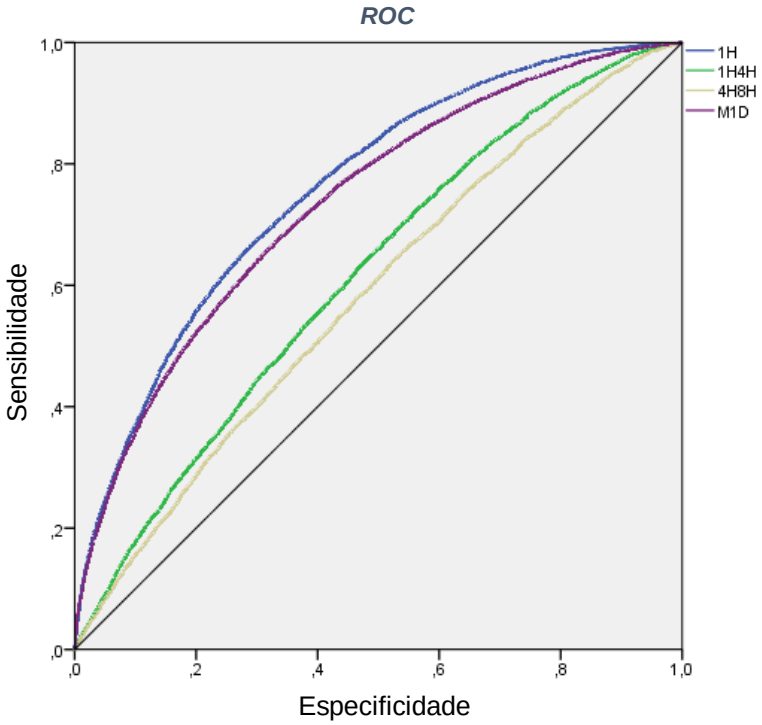
Constata-se por esta análise que técnicas de limitação de crescimento e *Pruning* não representam melhorias significativas na capacidade preditiva do algoritmo. Contudo realça-se o teste ID-P4, que ao permitir o crescimento da árvore (MAXDEPTH = 5) fez aumentar a sua eficácia, contudo acredita-se que continuar a aumentar a dimensão da árvore embora permita melhores resultados, vai levar a uma situação de *Overfitting, o* que se pretende evitar.

Relativamente ao modelo *Multilayer Perceptron,* a Tabela 29 apresenta a matriz de confusão deste modelo e a Figura 25 a curva ROC e a área abaixo da curva ROC (AUC) para cada classe da variável dependente.

Tabela 29: Matriz Confusão algoritmo Multilayer Perceptron da amostra de teste

	Previsto				
Observado	1H	1H4H	4H8H	M1D	**Corretamente Classificado**
1H	1146	0	0	1759	39,4%
1H4H	410	0	0	1410	0%
4H8H	200	0	0	1028	0%
M1D	612	0	0	6652	91,6%
Percentagem global	18,3%	0%	0%	82,1%	**59,0%**

Figura 25: Curva ROC e AUC -Multilayer Perceptron da amostra de teste



AUC

Grupo	Área
1H	0,76
1H4H	0,62
4H8H	0,58
M1D	0,74

ROC - Curva ROC; AUC - Área abaixo da curva ROC

A análise deste resultado comprova à constatação feita na análise à matriz de confusão do classificador SimpleCart do Weka (Tabela 22), ou seja, que os modelos são fortemente influenciado pela distribuição assimétrica entre as classes. Neste sentido e à semelhança do teste realizado com o classificador do Weka, o modelo *Multilayer Perceptron* do SPSS foi executado tendo por com base as subamostras geradas na 3ª iteração de preparação dos dados, de modo a se tentar atenuar a assimetria entre as classes e a baixa capacidade preditiva nas duas classes intermédias (1H4H e 4H8H). A Tabela 30 apresenta estes resultados.

Tabela 30: Comparativo da eficácia de Multilayer Perceptron com 2 subamostras

ID	Amostra	ICC (%)	AUC
S1	Padrão	**59,0**	0,68
S2	Nivelada	**36,4**	0,62
S3	3 Classes	**58,4**	0,67

ICC – Percentagem de Instâncias Corretamente Classificadas; AUC – Área abaixo da curva ROC

Uma vez mais constata-se que as alterações efetuadas na amostra (criação de subamostras) em nada beneficiam a capacidade preditiva do modelo, e que o resultado de 59% é fortemente influenciado pela assimetria da classe de tempo, uma vez que o nivelamento dos registos (ID-

S2) coloca a eficácia do modelo abaixo dos 50%. De igual forma confirma-se que a redução das classes, embora aumente a eficácia na nova classe intermédia (1D), não melhora a capacidade preditiva do modelo (ICC ID-S3 inferior ao padrão).

Os resultados e iterações desta fase de modelação são analisados em detalhe e discutidos na fase seguinte da metodologia CRISP-DM a fase 5: modelação.

3.6. CRISP-DM Fase 5: Avaliação

Nesta fase, pretende-se avaliar os modelos criados e validar se os mesmos cumprem com os objetivos definidos. Deste modo, a Tabela 31 apresenta, por ordem decrescente, as técnicas que apresentaram os melhores resultados e as respetivas capacidades preditivas de cada modelo. A Tabela 32 apresenta as matrizes de confusão dos três modelos que apresentaram os melhores resultados.

Tabela 31: Avaliação dos modelos gerados

ID	Ferramenta	Técnica	Algoritmo	Informação complementar	ICC (%)	AUC
Q1	SPSS	Redes neuronais	Multilayer Perceptron	*Holdout* (70% treino, 30% teste)	59,0	0,68
Q2	Weka	Árvore de decisão	SimpleCart	*K-fold*=10; atributos nominais	58,48	0,67
Q3	Weka	Árvore de decisão	J48	*K-fold*=10; atributos numéricos	58,02	0,66
Q4	SPSS	Árvore de decisão	CAHID	*K-fold*=10; atributos nominais	57,40	--

ICC – Percentagem de Instâncias Corretamente Classificadas; AUC – Área abaixo da curva ROC

Tabela 32: Matriz de confusão Multilayer Perceptron, SimpleCart e J48 (amostras de teste)

Multilayer Perceptron

Observado	Previsto					AUC	
	1H	1H4H	4H8H	M1D	Corretamente Classificado	Grupo	Área
1H	1146	0	0	1759	39,4%	**1H**	0,76
1H4H	410	0	0	1410	0%	**1H4H**	0,62
4H8H	200	0	0	1028	0%	**4H8H**	0,58
M1D	612	0	0	6652	91,6%	**M1D**	0,74
Percentagem global	18,3%	0%	0%	82,1%	**59,0%**	**AUC**	**0,68**

SimpleCart

	Previsto					**AUC**	
Observado	*1H*	*1H4H*	*4H8H*	*M1D*	***Corretamente Classificado***	**Grupo**	**Área**
1H	*4077*	*64*	*5*	*5561*	*42,0%*	**1H**	0,71
1H4H	*1450*	*69*	*3*	*4597*	*1,1%*	**1H4H**	0,58
4H8H	*806*	*37*	*4*	*3291*	*0,1%*	**4H8H**	0,54
M1D	*2321*	*78*	*2*	*21505*	*90,0%*	**M1D**	0,69
Percentagem global	*19,7%*	*0,6%*	*0,0%*	*79,7%*	***58,48%***	**AUC**	**0,67**

J48

	Previsto					**AUC**	
Observado	1H	1H4H	4H8H	M1D	**Corretamente Classificado**	**Grupo**	**Área**
1H	3506	42	0	6159	36,1%	**1H**	0,71
1H4H	1181	49	0	4889	0,8%	**1H4H**	0,57
4H8H	640	28	0	3470	0,0%	**4H8H**	0,54
M1D	1938	67	0	21901	91,6%	**M1D**	0,68
Percentagem global	48,3%	26,3%	0%	60,1%	**58,02%**	**AUC**	**0,66**

Constata-se que, os resultados obtidos com os três modelos de maior sucesso são, na prática, muito semelhantes, quer em termos de instâncias corretamente classificadas (59%, 58,48% e 58,02%) quer em termos de performance do modelo (AUC de 0,69; 0,67 e 0,66). Uma vez que foram testados vários algoritmos de aprendizagem com várias amostras, sem melhoria significativa à capacidade preditiva, conclui-se que a escolha do algoritmo, para efeitos do presente estudo, não constitui um fator diferenciador.

No que respeita aos objetivos de negócio, fixados na obtenção modelo preditivo que ajude a prever o tempo de resolução de um novo incidente, com uma taxa de sucesso superior a 60% e uma vez que apesar dos vários esforços (teste de diferentes algoritmos, softwares e customizações) o modelo com a melhor taxa de sucesso foi de 59%, conclui-se que contrariamente ao esperado os objetivos não foram cumpridos.

A Tabela 33 apresenta um comparativo entre as matrizes de confusão da amostra de treino e da amostra de teste (relativamente ao *Multilayer Perceptron*, dado ter apresentado o melhor resultado), de onde se pode concluir que não existem diferenças significativas entre a eficácia do modelo gerado (com base na amostra de treino) e a sua eficácia quando aplicado à amostra de teste. O que permite concluir que, o fraco desempenho preditivo apresentado, não está

relacionado com *overfitting,* significando que o modelo não detalhou em excesso os dados de treino (o que poderia prejudicar a eficácia do modelo em dados de teste ou reais), razão pela qual obteve praticamente o mesmo resultado em treino e em teste.

Tabela 33: Comparativo entre matriz de confusão da amostra de treino e de teste

		Previsto				
	Observado	1H	1H4H	4H8H	M1D	**Corretamente Classificado**
Treino	1H	2804	0	0	3983	41,3%
	1H4H	983	0	0	3309	0%
	4H8H	468	0	0	2435	0%
	M1D	1331	0	0	15276	92,0%
	Percentagem global	18,3%	0%	0%	81,7%	**59,1%**
Teste	1H	1146	0	0	1759	39,4%
	1H4H	410	0	0	1410	0%
	4H8H	200	0	0	1028	0%
	M1D	612	0	0	6652	91,6%
	Percentagem global	18,3%	0%	0%	82,1%	**59,0%**

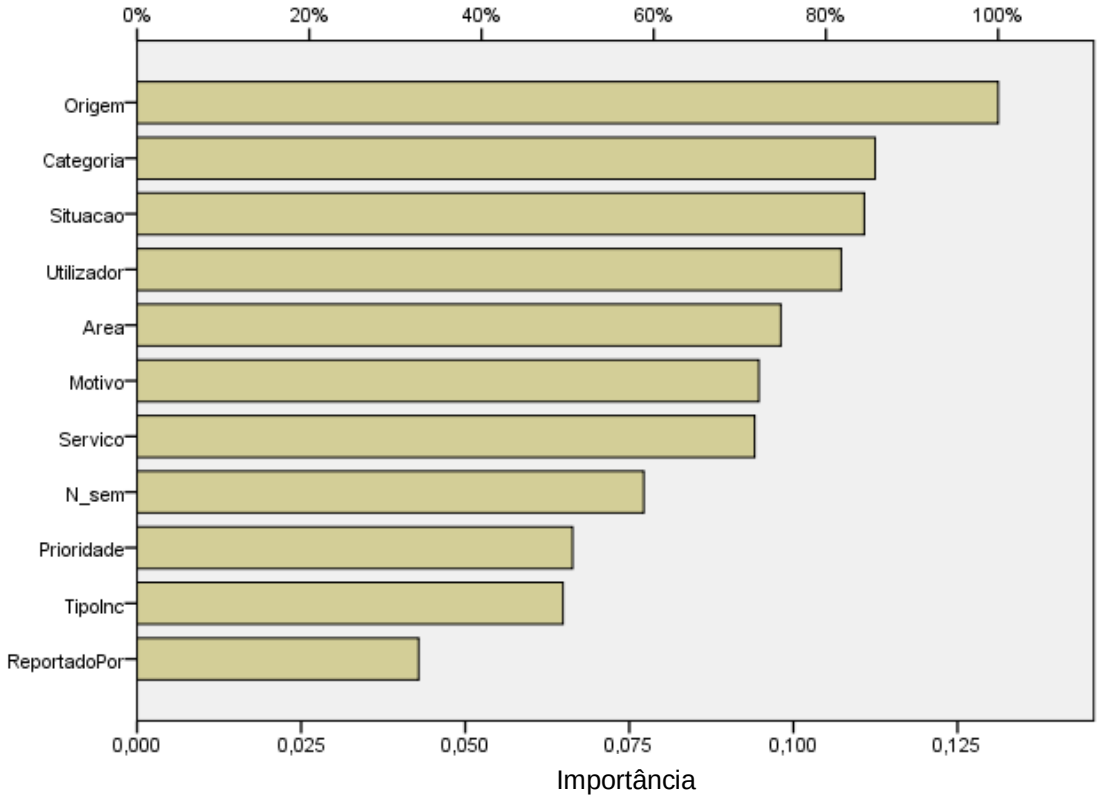
Uma vez que o fator técnica de *data mining*, não aparenta ser responsável pela fraca capacidade preditiva do modelo (dada a multiplicidade de algoritmos e *softwares* testados), aponta-se a ausência de padrões e a aleatoriedade dos dados da amostra, como fator explicativo dos fracos resultados obtidos.

No que respeita à relevância dos atributos para o modelo, a Tabela 34, apresenta a importância normalizada dos atributos para o modelo gerado com *Multilayer Perceptron.*

Da análise da importância dos atributos, conclui-se que os atributos Origem, Categoria e Situação, relacionados com a tipificação da origem da anomalia (conhecimento prévio da anomalia, identificação da funcionalidade afetada e descrição da anomalia, respetivamente) são os que mais influenciam o tempo de resolução, explicando no seu conjunto 35% do modelo (13%, 11% e 11% respetivamente). Os atributos relacionados com o utilizador que reporta a anomalia (Utilizador, Área e Motivo) são os seguintes a influenciar o tempo de resolução, dado que no conjunto conseguiram explicar 31% do modelo (11%, 10% e 10% respetivamente). O atributo Serviço, que identifica a aplicação ou ferramenta afetada, contrariamente ao esperado não apresenta relevância significativa no tempo, esta fator pode ser devido ao fato de haver uma relação entre os atributos Serviço e Categoria (sendo a Categoria um subtipo do Serviço),

logo ter maior relevância o atributo com maior granularidade. O atributo Prioridade, também contrariamente ao esperado, revelou ter baixa influência no tempo de resolução, o que sugere que a prioridade influencia o início da resolução do incidente, contudo não tem impacto aparente no seu tempo de resolução.

Tabela 34: Importância normalizada dos atributos para o modelo



Estas constatações confirmam os atributos referidos no ITIL, como relevantes para a gestão de incidentes (ponto 2.1.4.1, Registo de Incidentes) reforçando que um registo sistematizado da informação relativa à funcionalidade do SI afetado, descrição detalhada dos sintomas, identificação da área de negócio e do utilizador, no momento de registo do incidente, contribuem com 66% para um modelo de previsão do tempo de resolução. Contudo, uma comparação mais detalhada entre os atributos utilizados neste modelo e os atributos referidos no ITIL, revelou que a amostra em estudo, não contêm o atributo "Problema relacionado ou erro já conhecido". O fato deste atributo relacionar incidentes que partilhem a mesma causa ou que já estejam em tratamento, faz realçar que, a grande maioria dos registos da amostra (70%), são relativos a incidentes com as aplicações de negócio. O que, numa organização com forte desenvolvimento interno, sugere uma proporção considerável de correções e evoluções nas suas aplicações de negócio.

Segundo o ITIL é no processo de gestão de problemas (ponto 2.1.1.3) que se regista e gere os incidentes com causa desconhecida (problemas conhecidos), sendo que na organização

em estudo a resolução dos problemas em gestão por este processo, culmina muitas vezes na alteração da aplicação afetada por via de correções/alterações ao código da aplicação. Estas alterações (corretivas e evolutivas) nas aplicações podem provocar efeitos colaterais indesejados, uma vez que pode motivar outras anomalias no ecossistema das aplicações, sendo possível que uma correção provoque novas anomalias. Os testes de regressão (do Inglês *Regression Test*) (MSDN, 2015) são utilizados para minimizar este fator.

Neste contexto, é possível assumir que uma percentagem considerável dos registos da amostra em estudo, têm origem em anomalias provocadas por correções ou evoluções que geraram novos incidentes e não tanto por avarias ou anomalias recorrentes, uma vez que a amostra não contem atributos que permitam relacionar estes dois fatores (fator 1: relação entre novos incidentes e problemas conhecidos, fator 2: identificar se as aplicações ou ferramentas afetadas foram alvo de alterações recentes). Não foi possível identificar quais dos incidentes da amostra são recorrentes (e como tal seguem um padrão) e quais foram originados por fatores externos (logo desprovidos de padrão). Contudo, com base no conhecimento empírico da organização e do processo de gestão de incidentes, confirma-se a existência de uma quantidade considerável de incidentes derivados de correções ou publicação de novas versões das aplicações/ferramentas, o que indicia uma amostra com elevado número de eventos não relacionados e logo sem um padrão.

Neste sentido, formula-se a hipótese de que a amostra em estudo não é representativa, uma vez que, existem fortes indícios da existência de um elevado número de registos influenciados por fatores não representados na amostra e como tal, sem precedente nem padrão apontando-se esta hipótese como justificação para a fraca capacidade preditiva dos modelos gerados. Para estudos futuros sobre previsão de tempo e resolução de incidentes e num contexto semelhante (organizações com quantidades significativas de incidentes aplicacionais, corrigidos internamente) sugere-se a utilização de atributos que: 1) permitam relacionar as causas e sintomas registados, com causas e sintomas já conhecidos (erros/problemas conhecidos). 2) Permitam identificar se o componente de TI afetado (infraestrutura ou *Software*) pelo novo incidente, foi alvo de alterações (evolutivas ou corretivas) recentes.

Embora o modelo criado tenha revelado baixa capacidade preditiva, ajudou a compreender quais os atributos que mais contribuem para a previsão do tempo de resolução de incidentes. Esta análise conclui que, a amostra é fraca em termos de padrões uma vez que existem fortes indícios de que um número considerável de registos, é alvo de fatores externos não conhecidos (como tal desprovido de padrões conhecidos), ficando a sugestão de inclusão na amostra em estudos futuros (sobre previsão de tempos de resolução de incidentes) de atributos que permitam relacionar novos incidentes com problemas já conhecidos e com o registo de alterações recentes nos SI.

4. Conclusões

Apresenta-se de seguida o resumo do trabalho realizado com as respetivas conclusões, o contributo do presente estudo para a aplicação de *data mining* à gestão de incidentes do ITIL, limitações do estudo e sugestões para trabalhos futuros.

4.1. Resumo do trabalho

O presente trabalho de investigação teve como principal objetivo, a criação de um modelo preditivo através da descoberta de comportamentos e padrões para a previsão do tempo de resolução de incidentes ocorridos nos Sistema de Informação de uma grande organização. Para o processo de descoberta foram utilizadas técnicas de *data mining*, aplicadas a processos de tratamento e resolução de incidentes, implementados de acordo com as boas práticas preconizadas pelo ITIL.

No contexto corporativo atual, o alinhamento entre estratégia do negócio e estratégia das Tecnologias de Informação é determinante para o sucesso de uma organização, para potenciar novas estratégias de negócio, mas acima de tudo, para garantir a continuidade das operações vitais ao negócio. Uma organização impedida de explorar na totalidade o seu Sistema de Informação, devido a um incidente, é uma organização que enfrenta perdas financeiras e de oportunidades de negócio, muitas vezes irrecuperáveis.

A utilização de *standards* e normas de Gestão de Serviços de TI, ajuda as organizações a garantir este alinhamento. O ITIL surge como uma compilação de boas práticas, mundialmente aceite, como norma para a implementação de um Sistema de Gestão de Serviços de TI e como base para a certificação ISO/IEC 20000.

A presente investigação procura ajudar neste processo, apoiando as organizações na gestão dos impactos causados por um SI afetado, reforçando a mais-valia da utilização de metodologias como o ITIL e criando um modelo que permita prever a duração de indisponibilidade de um SI face a uma anomalia (tempo de resolução de incidentes).

Aplicando técnicas de *data mining*, a dados de registo de incidentes (de acordo com as boas práticas do ITIL), foi objetivo deste estudo criar um modelo que previsse, com uma taxa de sucesso superior a 60%, qual o tempo previsto de resolução de um incidente, nos primeiros momentos em que este é registado.

Para a obtenção deste objetivo foi utilizada a metodologia CRISP-DM para apoio à descoberta de conhecimento, onde foram testados vários algoritmos de *data mining* em diferentes ferramentas, sobre uma amostra de histórico de incidentes de uma organização, que segue as boas práticas do ITIL no seu processo de gestão de incidentes.

Em termos de ferramenta de *data mining* foi utilizado o *software* Weka (*open source*) e o *software IBM SPSS Statistics* (comercial). Realça-se a performance e a diversidade de algoritmos de *data mining* do Weka, tornando-o uma excelente ferramenta para o utilizador inexperiente devido ao interface gráfico e à variedade de tutoriais e documentação fornecido pelas comunidades *on-line*. Relativamente ao SPSS, há que salientar a robustez desta ferramenta comercial, uma vez que foi a única capaz de processar a amostra considerando atributos como nominais.

Dada a natureza dos atributos a modelar, o objetivo inicial foi utilizar modelos de regressão. Contudo, uma análise inicial demonstrou o fraco resultado desta abordagem, tendo sido alterados os atributos de modo a explorar modelos de classificação.

Os resultados obtidos, contrariamente ao esperado e apesar das várias técnicas, ferramentas e customizações utilizadas, não permitiram atingir o objetivo proposto de criar um modelo preditivo com uma eficácia superior a 60%, tendo o melhor modelo gerado (redes neuronais) conseguido apenas 59% de tempos de resolução corretamente classificados.

No entanto, constatou-se que a técnica de árvores de decisão (para problemas de classificação) produziu resultados muito similares, (algoritmos CART (no Weka SimpleCart) e C4.5 (no Weka J48)), quer em termos de instâncias corretamente classificadas, quer em termos de performance do modelo. De igual forma foi testada a possibilidade de sobreajustamento do modelo (*Overfitting,* o que poderia justificar a baixa capacidade preditiva), mas o teste levou a concluir não haver diferenças no desempenho do modelo entre a fase de criação (sobre a amostra de treino) e fase de teste (sobre a amostra de teste), descartando a hipótese de sobreajustamento.

A amostra revelou sofrer de assimetria na frequência de registos entre classes. Por esta razão foram testadas diferentes dimensões da amostra, entre as quais uma amostra com registos nivelada entre classes, tendo o melhor resultado sido de 36% de instâncias corretamente classificadas, indiciando que o modelo criado (com eficácia de 59%) sofria de uma tendência para prever tempos excessivos, influenciado pela classe com maior número de registos de incidentes.

Uma vez que o teste dos vários algoritmos de aprendizagem, combinados com diferentes dimensões da amostra em diferentes ferramentas não demonstraram variações significativas nos resultados, concluiu-se que a escolha do algoritmo, não constituiu um fator diferenciador.

No que respeita ao segundo objetivo proposto, identificar os fatores explicativos com maior capacidade preditiva do tempo de resolução, obtiveram-se resultados muito interessantes. Neste sentido, foi avaliado cada atributo relativamente à sua importância para o melhor modelo gerado (redes neuronais), realçando-se os atributos relacionados com a tipificação inicial do incidente e os atributos relacionados com a área de negócio que reporta a anomalia.

Confirmaram-se, deste modo, os atributos identificados no ITIL como relevantes para a gestão de incidentes. Esta constatação reforça ainda a importância para a gestão de incidentes, de um registo sistematizado da caracterização do incidente, com maior relevância para a funcionalidade afetada, respetivos sintomas e área afetada.

Já o atributo que identifica a aplicação ou ferramenta afetada (Serviço), contrariamente ao esperado, apresentou baixa relevância, indiciando que quanto maior a granularidade (nível de detalhe) do atributo que identifica a funcionalidade afetada, maior a relevância desse atributo para a previsão do tempo de resolução (uma vez que o atributo Categoria, subtipo do atributo Serviço, revelou ter maior influência no modelo).

De igual forma, o atributo que identifica a prioridade, revelou também uma baixa influência para o modelo, sugerindo que a prioridade influencia apenas o início da resolução do incidente, sem grande impacto no tempo de resolução.

Por fim constatou-se que, atributos que relacionam causas comuns entre incidentes (problemas relacionados) e/ou contexto em que os incidentes ocorrem (alterações nos SI), embora estejam referenciados no ITIL não constam na amostra, logo a sua relevância para ao modelo não pôde ser testada.

Este fato levou à constatação de que, 70% dos registos da amostra utilizados para gerar o modelo, são relativos a incidentes aplicacionais (anomalias nas aplicações de negócio) e uma vez que a organização produtora desta amostra tem uma forte cultura de desenvolvimento interno, fica implícito a existência de uma quantidade considerável de alterações (corretivas e evolutivas) nas suas aplicações de negócio, o que indicia que um elevado número de registos da amostra utilizados para a criação deste modelo, é motivado por alterações ao ecossistema de aplicações (logo novos fatores e sem padrão conhecido) e não por situações recorrentes.

Neste contexto formulou-se a hipótese de que os fracos resultados obtidos pelo modelo, possam ser motivados por uma aleatoriedade e falta de padrões na amostra, devido à ausência de atributos que permitam: 1) Relacionar novos incidentes com problemas já conhecidos e 2) identificar alterações recentes ocorridas no ecossistema de aplicações, na altura de registo do incidente. Ficando a pista de que este tipo de atributos possam ser a chave para a ausência de padrões constatada.

Adicionalmente, esta dificuldade em prever com precisão o tempo de resolução de incidentes pode significar que no momento de registo do incidente, a informação fornecida às equipas de suporte não é a mais correta, podendo levar a uma caracterização errada do incidente. Por exemplo, a um incidente pode ser atribuída prioridade ou impacto errados, ou um incidente ser considerado de aplicação e, na realidade, ser do posto de trabalho. Desta forma, deverá pensar-se em criar um modelo de previsão dos tempos resolução num momento, próximo do

registo, mas após uma primeira avaliação do incidente pelo técnico a quem foi atribuído o incidente.

Embora o modelo criado tenha revelado uma capacidade preditiva inferior ao esperado, ajudou a identificar e compreender os atributos que mais contribuem para a previsão do tempo de resolução de incidentes, o que revela utilidade para a organização, já que foi gerado novo conhecimento relacionado com a gestão de incidentes.

4.2. Contributos

Esta investigação procura ajudar organizações e profissionais de TI a melhor gerir os seus Sistemas de Informação, sugerindo o ITIL como boa prática para esta gestão, em particular da gestão de incidentes, propondo-se a apoiar na difícil tarefa que é estimar o tempo de resolução de um novo incidente.

A utilização da metodologia CRISP-DM (não-proprietária) revelou ser a adequada, fornecendo um fio condutor ao processo de descoberta de conhecimento, pela sua natureza interativa ao longo das várias fases, destacando-se a importância da fase de preparação dos dados, para o sucesso de um estudo de *data mining*.

Convém frisar que esta investigação foi das poucas que se propôs a aplicar técnicas de *data mining* à gestão de incidentes do ITIL. Desta forma, contribui-se para o conhecimento sobre esta temática ao proporcionar mais evidência sobre os resultados da aplicação destas técnicas, neste caso, com o objetivo de prever o tempo de resolução de incidentes, quer tratando-o como um atributo contínuo, quer como discreto (classes de tempo de resolução). Este estudo, evidencia a necessidade de continuar a investigação nesta área, nomeadamente na compreensão dos atributos necessários para a criação de um modelo preditivo de sucesso para o tempo de resolução. De facto, levanta-se a hipótese de não ser possível criar um modelo com uma eficácia relevante, apenas com a informação existente no momento de registo do incidente; e que só após identificada a causa da anomalia (o que requer tempo adicional para um diagnóstico) será possível prever um tempo de resolução.

4.3. Limitações

Como limitação, realça-se o fato desta investigação ter feito uso dos dados de apenas uma organização, que embora implemente as boas práticas no ITIL na gestão de incidentes, pode apresentar características específicas na sua operação que comprometam os resultados. Pelo que, o pouco sucesso do presente estudo, não pode ser generalizado nem conclusivo relativamente à capacidade preditiva do tempo e resolução de incidentes.

4.4. Trabalhos futuros

Esta investigação, dados os seus resultados, proporciona diversas perspetivas de trabalho futuro. Foram identificados os atributos que mais influenciaram o modelo, ficando pistas sobre atributos adicionais necessário a incluir na amostra, para permitir a criação de modelos com melhor qualidade.

Neste sentido sugere-se:

- A realização de um estudo semelhante, tendo por base uma amostra de histórico de incidentes de outra organização que utilize um processo de gestão de incidentes baseado no ITIL, a comparação dos atributos disponíveis e o resultado dos modelos obtidos, ajudará a compreender os fatores que contribuíram para os fracos modelos gerados.

- Enriquecer a amostra disponível para o estudo com atributos que contextualizam o novo incidente, nomeadamente atributos que:

    1. Permitam relacionar novos incidentes com problemas já conhecidos, caso partilhem os mesmos sintomas e/ou afetem o mesmo componente (aplicação, infraestrutura). Podendo ser por via de associação de dois (ou mais) incidentes que partilhem a mesma causa, ou por via do peso dessa relação (o mesmo sintoma no mesmo componente, peso 3; mesmo componente com sintomas diferentes, peso 2; e mesmo sintoma em componentes diferentes, peso 1)

    2. Permitam identificar alterações recentes ocorridas no ecossistema de aplicações no momento de registo do incidente. A título de exemplo, sugere-se a criação de um atributo que identifique se ocorreu recentemente uma correção, uma nova aplicação, etc.

- Recorrer a um painel de especialistas, no sentido de recolher sensibilidades e experiências sobre os atributos importantes para o modelo preditivo.

- Na eventualidade de se comprovar que não é possível criar um modelo suficientemente robusto apenas com a informação existente no momento de registo do incidente, seria interessante realizar novo estudo considerando todos os atributos recolhidos ao longo do processo de resolução de incidentes e identificar qual a combinação necessária de atributos para viabilizar um modelo eficaz. Deste modo, seria possível determinar em que fase do processo de gestão de incidentes é possível prever o seu tempo de resolução.

    .

5. Bibliografia

Azevedo, A. e Santos, M. 2008. KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. *IADIS European Conference Data Mining* (pp. 182-185).

Bang.(2010). A comparison of the business and technical drivers for ISO 27001, ISO 27002, COBIT and ITIL. [Online] Novembro de 2010. http://trongbang86.blogspot.pt/2010/11/comparison-of-business-and-technical.html.

Bon , Jan van. (2007). ITIL® V3: A Pocket Guide. s.l. : ITSM Library.

Cannon, David e Wheeldon, David. (2007). ITIL Service Operation. London : TSO.

Cartlidge, Alson e Hanna, Ashley. (2007). An Introductory Overview of ITIL© V3. UK : itSMF Limited.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al. (2000). CRISP-DM 1.0 - Step-by-step data mining guide. CRISP-DM Consortium.

Cruz, A. J. (2007). Data Mining via Redes Neuronais Artificiais e Máquinas de Vectores de Suporte. Dissertação de Mestrado. Universidade do Minho – Escola de Engenharia, Guimarães.

Drazin, Sam e Montag, Matt. (2012). Decision Tree Analysis using Weka. University of Miami.

Excelos. (2013). ITIL Maturity Level. 2013, p. 4.

Fawcett, Tom. (2005). An introduction to ROC analysis.

Fayyad, Usama, Shapiro, Gregory e Smyth, Padhraic. (1996). From Data Mining to Knowledge Discovery in Databases. AI MAGAZINE.

Han, Kamber. (2006). Data Mining: Concepts and Techniques second editon. s.l. : Morgan Kaufmann Publishers.

HP. (2010). Service Management Services. [Online] 13 de Janeiro de 2010. http://bit.ly/gg0SKq.

IBM. (2009). The IBM Process Reference Model for IT (PRM-IT). [Online] 16 de Janeiro de 2009. http://bit.ly/hfONrD.

Iqbal, Majid e Nieves, Michael. (2007). ITIL Service Strategy. London : TSO.

ISACA. (2010). ISACA. [Online] 1 de Janeiro de 2010. http://www.isaca.org.

ISO/IEC. (2005a). ISO 20000-1 - Information technology - Service management - part1: Specification.

ISO/IEC. (2005b). 20000-2 - Information technology - Service management - part2: Code of practice. 2005.

itSMF. (2013). itSMF 2013 Global Survey. [Online] 2013. http://www.itsmfi.org/files/itSMF%202013%20Survey%20Report_0.pdf.

Lacy. (2014). ITIL V3 support for achieving ISO/IEC 20000. www.bcs.org. [Online] Dezembro de 2014. http://www.bcs.org/content/conWebDoc/15851.

Lacy, MacFarlane. (2007). ITIL Service Transition. London : TSO.

Lacy, Shirley e Macfarlane , Ivor. (2007). ITIL Service Transition. London : TSO.

Laureano, R., Caetano, N. & Cortez, P. (2014). Previsão de tempos de internamento num hospital português: Aplicação da metodologia CRISP-DM. RISTI – Revista Ibérica de Sistemas e Tecnologias de Informação, 13, 83-98. doi: http://dx.doi.org/10.4304/risti.13.83-98

Liao, S.H., Chu, P.H., & Hsiao, P.Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. s.l. : Department of Management Sciences, Tamkang University, 2012.

livetime.com. (2014). itil-v3-service-management-lifecycle. livetime.com. [Online] Dezembro de 2014. http://developer.livetime.com/itil-v3-service-management-lifecycle/.

Loyd, Ruud. (2007). ITIL Service Design. London : TSO.

Michalewicz, Z., Schmidt, M., Michalewicz, M., & Chiriac, C. (2006). Adaptive Business Intelligence (1 ed.). New York : s.n.

Microsoft. (2008). Microsoft Technet Library. [Online] 10 de Outubro de 2008. http://technet.microsoft.com/en-us/library/cc543224.aspx..

MSDN. (2015). MSDN Microsoft. Developing with Visual Studio .NET. [Online] Microsoft, 2015. https://msdn.microsoft.com/en-us/library/aa292167(v=vs.71).aspx.

OGC. (2006). Office of Government Commerce. ITIL IT Service Management: Glossary of Terms, Definitions and Acronyms. 2006.

OGC. (2007). The Official Introduction to the ITIL Service Lifecycle Book. s.l. : TSO (The Stationery Office, 2007.

Rocha, M., Cortez, P. e Neves, J. (2008). Análise Inteligente de Dados - Algoritmos e Implementação em Java. s.l. : FCA.

Sarbanes-Oxley. (2006). A Guide To The Sarbanes-Oxley Act. A Guide To The Sarbanes-Oxley Act. [Online] 2006. http://www.soxlaw.com/.

Song, Yan-yan. (2015). Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. [Online] 2015. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/.

Spalding, George e Case, Spalding. (2007). ITIL Continual Service Improvement. London : TSO.

Tape. (2015). The Area Under an ROC Curve. [Online] Janeiro de 2015. http://gim.unmc.edu/dxtests/roc3.htm.

Tetko, Igor, Livingstone, David e Luik, Alexander. (1995). Neural network studies. 1. Comparison of Overfitting and Overtraining. UK : s.n., 1995.

Wikipedia. (2015). Artificial neural network. [Online] Janeiro de 2015. http://en.wikipedia.org/wiki/Artificial_neural_network.