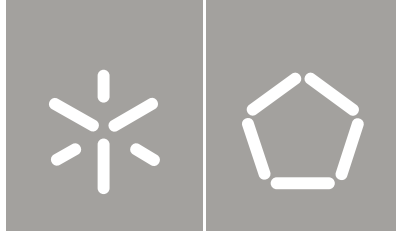




Universidade do Minho
Escola de Engenharia

Pedro Miguel Lopes Bastos

**Sistema de predição de avarias
em máquinas de unidades fabris
globalmente dispersas**



Universidade do Minho
Escola de Engenharia

Pedro Miguel Lopes Bastos

**Sistema de predição de avarias
em máquinas de unidades fabris
globalmente dispersas**

Tese de Doutoramento
Ramo Engenharia Industrial e de Sistemas

Trabalho efectuado sob a orientação dos
Professora Doutora Isabel da Silva Lopes
Professor Doutor Luís Carlos Magalhães Pires

É autorizada a reprodução integral desta tese
apenas para efeitos de investigação,
mediante declaração escrita do interessado,
que a tal se compromete.

Universidade do Minho, Maio de 2015

Assinatura: 

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração da presente tese. Confirmando que em todo o trabalho conducente à sua elaboração não recorri à prática de plágio ou a qualquer forma de falsificação de resultados.

Mais declaro que tomei conhecimento integral do Código de Conduta Ética da Universidade do Minho.

Universidade do Minho, 26 de maio de 2015

Nome completo: Pedro Miguel Lopes Bastos

Assinatura: 

Para a Lillian e Inês
Para os meus Pais

Agradecimentos

Tenho muito que agradecer a várias pessoas pela ajuda e disponibilidade que mostraram ao longo destes quatro anos de realização dos trabalhos de Doutoramento. Obviamente começo por agradecer aos meus orientadores, à professora Isabel Lopes e o professor Luís Pires, pelo apoio científico prestado, pelas críticas construtivas e pelos conselhos preciosos em momentos de dúvida e hesitação. No decorrer de todo este processo houve momentos difíceis em que as suas palavras e recomendações foram decisivas para que tudo se tornasse mais claro, num ambiente de reflexão e satisfação permanentes.

Aproveito para agradecer ao Ministério da Ciência, Tecnologia e Ensino Superior e à Fundação para a Ciência e para a Tecnologia, pelo financiamento deste trabalho através da Bolsa PROTEC (Programa de Apoio à Formação Avançada de Docentes do Ensino Superior Politécnico), pelo apoio concedido à instituição onde exerço funções de docência de forma a suportar o pagamento das propinas de Doutoramento e a dispensa parcial de serviço docente. Gostaria também de aproveitar a ocasião para agradecer ao Departamento de Produção e Sistemas da Universidade do Minho, ao Centro de Investigação Algoritmi da Universidade do Minho e ao Centro de Investigação da Montanha da Escola Superior Agrária de Bragança, pelos subsídios concedidos para deslocações e para conferências onde fui apresentando o trabalho elaborado.

Agradeço também à Empresa Bosch Car Multimédia Portugal S.A. na pessoa do Doutor Carlos Machado pela disponibilidade prestada durante o decorrer dos trabalhos, bem como na disponibilização de dados funcionais da empresa que serviram de suporte à elaboração desta tese.

Os meus agradecimentos ao diretor da Escola Superior Agrária de Bragança, Professor Doutor Albino Bento pelas condições disponibilizadas de forma a proporcionar todas as condições necessárias à elaboração desta tese.

Agradeço também ao Mestre Nuno Carvalho, amigo e técnico informático do centro de informática da Escola Superior Agrária de Bragança, pelo apoio prestado em determinados momentos de indecisões e dúvidas ocorridas durante os trabalhos efetuados.

Aos amigos de longo data, que foram perguntando pelo trabalho e suportaram as minhas ausências.

Não podia deixar de agradecer às pessoas mais importantes da minha vida. À minha família, pais e avó, pelo apoio incondicional que sempre me deram. Em especial tenho de agradecer à

Lillian, esposa e amiga, que sempre me apoiou nos bons e nos maus momentos. Por todo o amor e carinho, por toda a confiança em mim depositada e por toda a ajuda e motivação. Agradeço também à minha filha Inês, que apesar da sua tenra idade sempre me deu incondicionalmente todo o seu amor e carinho em horas em que me sentia um pouco mais em baixo.

Por fim, gostaria de endereçar os meus agradecimentos a todos aqueles de uma forma ou de outra foram ajudando anonimamente nas inúmeras discussões ao longo destes anos.

A todos, o meu sincero obrigado

Resumo

Nos últimos anos, temos assistido a várias e profundas alterações na produção industrial. Muitos processos industriais estão agora automatizados com o objetivo de garantir a qualidade da produção e minimizar os seus custos. Atualmente, as organizações têm vindo a recolher e armazenar quantidades cada vez maiores de dados relevantes e precisos dos seus processos de produção. Este armazenamento de dados oferece um enorme potencial, constituindo uma fonte de novo conhecimento. No entanto, a elevada quantidade de dados e a sua complexidade não se coaduna em muitos casos com a capacidade de analisá-los, e torna-se necessária a utilização de técnicas de análise automatizadas.

O *data mining* emerge como uma importante ferramenta no processo de aquisição de conhecimento a partir de dados provenientes do processo produtivo. Apresenta uma oportunidade de aumento significativo da capacidade de transformação de elevados volumes de dados em informação útil. O uso cumulativo de dados tem sido limitado, o que conduz ao problema de "*fontes de dados ricas, mas pobres em informação*". Neste trabalho, através da utilização da ferramenta de *data mining RapidMiner*, são aplicados diferentes algoritmos a dados de manutenção e de monitorização da condição de determinados equipamentos existentes em diferentes linhas de produção. Os algoritmos aplicados são comparados quanto à exatidão obtida na descoberta de padrões e nas predições efetuadas. A recolha de dados baseia-se num sistema de agentes distribuídos, o que, dada a sua natureza, será responsável pela recolha de dados através de uma arquitetura funcional. O uso de *data mining* está integrado num sistema *on-line* capaz de recolher dados através da utilização de agentes automáticos, apresentando os resultados obtidos às diferentes equipas de manutenção, de forma facilmente compreensível. O objetivo dos algoritmos de predição desenvolvidos é de prever valores futuros com base em registos de valores presentes, a fim de estimar a possibilidade da falha de uma determinada máquina e, desta forma, apoiar as equipas de manutenção no planeamento de medidas adequadas para evitar falhas ou para mitigar os seus efeitos.

As principais contribuições deste trabalho são: (i) definição da arquitetura de um sistema funcional de predição de avarias, (ii) a criação de um protótipo de *data mining* utilizando para tal a ferramenta *RapidMiner v.5.3.15*.

Palavras chave: *Data mining*, e-colaboração, manutenção, gestão de manutenção.

Abstract

In the last years we have assisted to several and deep changes in industrial manufacturing. Many industrial processes are now automated in order to ensure production quality and to minimize costs. Currently, manufacturing companies have been collecting and storing increasingly larger amounts of accurate and relevant production data. The stored data offer enormous potential, providing a source of new knowledge. However, the huge amount of data and its complexity is not consistent in many cases with the ability to analysing ability .

Data mining has emerged as an important tool for knowledge acquisition from manufacturing databases. Data mining technology presents an opportunity to increase significantly the rate at which the volume of data can be turned into useful information. However, the use of accumulated data has been limited, which has led to the “rich data but poor information” problem. In this work, a data mining tool named RapidMiner is used to create and apply different data mining prediction algorithms to maintenance and condition monitoring data. Their accuracy in the discovery of patterns and also the accuracy of predictions is compared. This tool is integrated with an online system which collects data using automatic agents and presents all the results to the maintenance teams in an comprehensible way. The remote data collection is based on an intricate system of distributed agents, which, given its nature, will be responsible for remote data collection through a functional architecture. The purpose of the prediction algorithms applied is to forecast future values based on present records, in order to estimate the possibility of a machine breakdown, and therefore to support maintenance teams in planning appropriate measures to avoid failures or to mitigate their effects.

The main contributions of this work are (i) the definition of a system architecture; (ii) the creation of an data mining prototype system using a data mining tool named RapidMiner v.5.3.15.

Keywords: Data mining, e-collaboration, maintenance, maintenance management.

Índice

Agradecimentos	vii
Resumo	ix
Abstract	xi
Índice	xiii
Índice de figuras	xv
Índice de tabelas	xvii
Lista de siglas e acrónimos	xix
1. Introdução	3
1.1 Motivação e objetivos	4
1.2 Metodologia aplicada	7
1.3 Organização e estrutura da Tese.....	8
2. Estado da arte	13
2.1 Manutenção industrial	13
2.2 Gestão da manutenção.....	16
2.2.1 Sistemas reparáveis e não reparáveis.....	17
2.2.2 Fiabilidade	18
2.2.3 Tipos de manutenção	24
2.2.4 Métodos de gestão da manutenção	31
2.2.5 Modelos de apoio à decisão na manutenção	34
2.2.6 <i>Softwares</i> de manutenção.....	36
2.3 Modelos distribuídos de aquisição de dados	38
2.3.1 Modelo genérico	38
2.3.2 Modelo cliente-servidor	39
2.3.3 Agentes	39
2.4 A descoberta de conhecimento em bases de dados	44
2.4.1 <i>Data Mining</i>	44
2.4.2 Metodologias e especificações.....	49
2.4.3 Objetivos em <i>data mining</i>	56
2.4.4 Modelos, técnicas e algoritmos.....	59
2.4.5 Abordagens para validação de resultados	72
2.5 Algumas aplicações do <i>data mining</i>	77

3. O sistema de predição	83
3.1 Visão global.....	84
3.2 Especificação funcional.....	87
3.2.1 A Metodologia IDEF0	88
3.2.2 Descrição funcional	93
3.3 Recolha de dados.....	115
3.3.1 Tecnologias utilizadas.....	115
3.4 Desenvolvimento do interface para disponibilização de resultados.....	122
3.4.1 Ferramenta de desenvolvimento	123
3.4.2 Linguagens aplicadas	124
3.5 Ferramenta para elaboração de modelos preditivos	126
3.5.1 O RapidMiner	126
3.5.2 Interface gráfico do RapidMiner.....	127
4. Desenvolvimento e implementação do módulo de predição	133
4.1 A Metodologia de conceção	133
4.2 O módulo preditivo	136
4.2.1 Processo 1 – Criação do modelo preditivo	137
4.2.2 Processo 2 – Aplicação do modelo preditivo.....	143
4.3 Caso prático.....	145
4.3.1 Estrutura de dados.....	147
4.3.2 Aplicação do Processo 1	152
4.3.3 Aplicação do Processo 2	157
4.3.4 Validação e interpretação de resultados obtidos pelo modelo.....	159
4.3.5 Disponibilização de resultados por parte do sistema	163
4.4 Resultados e discussão	164
5. Conclusões e perspectivas de trabalho futuro	171
5.1 Contribuições	173
5.2 Desafios e perspectivas de trabalho futuro	175
Bibliografia	179
Anexo A	191
Anexo B	195
Anexo C	197
Anexo D	201

Índice de figuras

FIGURA 2-1 - FUNÇÃO DE PROBABILIDADE DE FALHA ACUMULADA.....	20
FIGURA 2-2- FUNÇÃO DE FIABILIDADE	20
FIGURA 2-3 - CURVA DA BANHEIRA	22
FIGURA 2-4 - TEMPO ENTRE AVARIAS E TEMPO DE FUNCIONAMENTO ATÉ À AVARIA.	23
FIGURA 2-5 - TIPOS DE MANUTENÇÃO.	26
FIGURA 2-6 - CURVA P-F.	29
FIGURA 2-7 - PASSOS RELATIVOS À MANUTENÇÃO PREVENTIVA CONDICIONADA.	30
FIGURA 2-8 - SUBSTITUIÇÃO EM INTERVALOS DE OPERAÇÃO CONSTANTES.	35
FIGURA 2-9 - SUBSTITUIÇÃO EM INTERVALOS DE TEMPO CONSTANTES.	35
FIGURA 2-10 - ÂMBITO DOS AGENTES INTELIGENTES.	41
FIGURA 2-11 - FASES DO PROCESSO DE DCBD.	46
FIGURA 2-12 - CICLO DE VIDA DA METODOLOGIA CRISP-DM.	50
FIGURA 2-13 - ETAPAS DA METODOLOGIA SEMMA.	52
FIGURA 2-14 - FASES DAS METODOLOGIAS CRISP-DM E SEMMA.	54
FIGURA 2-15 - REPRESENTAÇÃO GRÁFICA DA ESPECIFICAÇÃO PMML.....	55
FIGURA 2-16 - EXEMPLO DE FICHEIRO PMML.	56
FIGURA 2-17 - OBJETIVOS E ALGORITMOS DE <i>DATA MINING</i>	56
FIGURA 2-18 - ÁRVORE DE DECISÃO – EXEMPLO: SAIR DE CASA.	61
FIGURA 2-19 - INDUÇÃO DE ÁRVORES DE DECISÃO.	62
FIGURA 2-20 - EXEMPLO DE REGRAS DE INDUÇÃO.....	65
FIGURA 2-21 - REDE DE <i>BAYES</i>	68
FIGURA 2-22 - NEURÓNIO ARTIFICIAL	70
FIGURA 2-23 - ILUSTRAÇÃO DO PRINCÍPIO BÁSICO DA TÉCNICA K-NN.....	71
FIGURA 2-24 - MATRIZ DE CONFUSÃO	74
FIGURA 3-1 - RELAÇÕES ENTRE PROCESSOS, DADOS E INTERVENÇÕES.....	85
FIGURA 3-2 - INFLUÊNCIA DO SISTEMA NA GESTÃO DA MANUTENÇÃO.	86
FIGURA 3-3 - ESQUEMATIZAÇÃO GERAL DO PROJETO.	87
FIGURA 3-4 - NOMENCLATURA DA TÉCNICA IDEF0.	90
FIGURA 3-5 - DECOMPOSIÇÃO PELO MÉTODO IDEF0.....	91
FIGURA 3-6 - RAMIFICAÇÃO DE SETAS.	92
FIGURA 3-7 - JUNÇÃO DE SETAS.	92
FIGURA 3-8 - ENTRADAS, SAÍDAS, CONTROLOS E MECANISMOS MÚLTIPLOS.	93
FIGURA 3-10 - ATIVIDADE AO - SPAMUF.....	97
FIGURA 3-11 - SUBACTIVIDADES A1 – SISTEMATIZAÇÃO DA REDE COLABORATIVA.....	100
FIGURA 3-12 - SUBACTIVIDADES A11 – DEFINIÇÃO E INICIALIZAÇÃO DO SISTEMA.....	102
FIGURA 3-13 - SUBACTIVIDADES A12 – INTEGRAÇÃO DE PARCEIROS.	103
FIGURA 3-14 - SUBACTIVIDADES A2 - GESTÃO DA REDE.	105
FIGURA 3-15 - CRIAÇÃO DA BASE DE DADOS FUNCIONAL DO SISTEMA.	106
FIGURA 3-16 - SUBACTIVIDADES A3 - OPERAÇÃO.	107

FIGURA 3-17 - SUBACTIVIDADES A31 - GESTÃO DE DADOS E COMUNICAÇÃO.....	108
FIGURA 3-18 - SUBATIVIDADES A32 – PREDIÇÃO DE CONHECIMENTO.	109
FIGURA 3-19 - SUBATIVIDADE A322 - GERAÇÃO DE CONHECIMENTO.	110
FIGURA 3-20 - SELEÇÃO DE DADOS CONFORME SOLICITAÇÃO AO SISTEMA.	111
FIGURA 3-21 - FLUXOGRAMA DA METODOLOGIA DESENVOLVIDA PARA IMPLEMENTAÇÃO DA SUBATIVIDADE A3223 - <i>DATA MINING</i>	113
FIGURA 3-22 - SUBATIVIDADE A3223 - <i>DATA MINING</i>	114
FIGURA 3-23 - SISTEMA DE RECOLHA AUTOMÁTICA DE DADOS NÍVEL 1.....	116
FIGURA 3-24 - APLICAÇÃO DO XML.	118
FIGURA 3-25 - ENCAPSULAMENTO DE UM PACOTE.	121
FIGURA 3-26 - PORTAL DO PROTÓTIPO DO SISTEMA DE PREDIÇÃO.....	123
FIGURA 3-27 - INTERFACE DO <i>RAPIDMINER</i> V5.3.015.....	128
FIGURA 4-1 - O MODELO DE PREDIÇÃO E OS PROCESSOS ASSOCIADOS.	137
FIGURA 4-2- CICLOS DE DCBD EM MODELO ESPIRAL EVOLUTIVO (DE BAIXO PARA CIMA CICLOS 1 A N).....	138
FIGURA 4-3 - PROCESSO DE GERAÇÃO DO MODELO DE PREDIÇÃO.	139
FIGURA 4-4 - ESQUEMA RELACIONAL DA BASE DE DADOS FUNCIONAL DO SISTEMA.	140
FIGURA 4-5 - PROCESSO DE PREDIÇÃO.	144
FIGURA 4-6 - ESQUEMA RELACIONAL DA BASE DE DADOS DO CASO PRÁTICO.	148
FIGURA 4-7 - REPRESENTAÇÃO GRÁFICA DA METODOLOGIA SEGUIDA NO CASO DE ESTUDO.....	151
FIGURA 4-8 - SELEÇÃO DE ATRIBUTOS PARA GERAÇÃO DO MODELO PREDITIVO.	152
FIGURA 4-9 - CONFIGURAÇÃO E APLICAÇÃO DO OPERADOR < <i>VALIDATION</i> >.	153
FIGURA 4-10 - MATRIZ DE CONFUSÃO DO MODELO.....	154
FIGURA 4-11 - OUTPUT GERADO PELO 1º PROCESSO.	156
FIGURA 4-12 - EXEMPLO DOS ALERTAS FORNECIDOS PELO SISTEMA.	158
FIGURA 4-13 - OCORRÊNCIAS REAIS EFETUADAS E AS PREVISTAS PELO SISTEMA.....	161
FIGURA 4-14 - VISUALIZAÇÃO GRÁFICA DA PREDIÇÃO EFETUADA EM RELAÇÃO AO PARÂMETRO REAL MONITORIZADO.....	162
FIGURA 4-15 - INTERFACE DE AUTENTICAÇÃO NO SISTEMA.....	163
FIGURA 4-16 - PÁGINA OBTIDA APÓS AUTENTICAÇÃO.	164

Índice de tabelas

TABELA 2-1 - COMPONENTES DO PMML.....	55
TABELA 2-2 - RESUMO DAS ÁRVORES DE DECISÃO.....	65
TABELA 3-1 - FLUXOS NA REPRESENTAÇÃO GLOBAL DO SISTEMA.	95
TABELA 3-2 - MECANISMOS E INFORMAÇÃO DE CONTROLO DO SISTEMA.	96
TABELA 3-3 - FLUXO DE SAÍDA ATIVIDADE A0.....	99
TABELA 3-4 - MECANISMOS E INFORMAÇÃO DE CONTROLO DA ATIVIDADE A0.	99
TABELA 3-5 - FLUXOS DE SAÍDA DA SUBATIVIDADE A1.....	101
TABELA 3-6 - NOVOS MECANISMOS APRESENTADOS NA SUBATIVIDADE A1.....	101
TABELA 3-7 - NOVOS MECANISMOS E INFORMAÇÃO DE CONTROLO DA SUBATIVIDADE A32.....	110
TABELA 3-8 - NOVO FLUXO DE ENTRADA DA SUBATIVIDADE A3221.....	111
TABELA 3-9 - NOVOS FLUXOS DE CONTROLO DA SUBATIVIDADE A322.	112
TABELA 4-1 - ESTRUTURA DA TABELA DE MONITORIZAÇÃO.....	149
TABELA 4-2 - ESTRUTURA DA TABELA DE INTERVENÇÕES.	150
TABELA 4-3 - PARAMETRIZAÇÃO E RESULTADOS OBTIDOS COM DIFERENTES OPERADORES.	155
TABELA 4-4 - CONTAGEM DO NÚMERO DE OCORRÊNCIAS REAIS E AS PREVISTAS PELO SISTEMA.....	160

Lista de siglas e acrónimos

AGPL - *Affero General Public License*

AD - *Árvores de Decisão*

CMMS - *Computerized Maintenance Management System*

CMS - *Content Management System*

CRISP-DM – *Cross-Industry Standard Process for Data Mining*

DCBD - *Descoberta de Conhecimento em Bases de Dados*

DM - *Data Mining*

DOS - *Disk Operating System*

GUI - *Graphic User Interface*

HTML - *HyperText Markup Language*

IA - *Inteligência Artificial*

IAD - *Inteligência Artificial Distribuída*

IR - *Indução de Regras*

IDEFO - *Integration Definition for Function Modeling*

IP - *Internet Protocol*

K-NN - *K-Nearest Neighbours*

MTBF- *Mean Time Between Failures*

OEE - *Overall Equipment Effectiveness*

PERL - *Practical Extraction and Report Language*

PHP - *Hypertext Preprocessor*

PMML - *Predictive Model Markup Language*

RB - *Redes de Bayes*

RBM - *Risk-based Maintenance*

RCM - *Reliability Centered Maintenance*

RNA - *Redes Neurais Artificiais*

ROCOF – *Rate of Occurrence of Failures*

SMA - *Sistemas Multiagente*

SEMMA - *Sample, Explore, Modify, Model, Assessment*

SQL - *Structured Query Language*

TPM - *Total Productive Maintenance*

VPN - *Virtual Private Network*

XML - *eXtensible Markup Language*

Capítulo 1

1. Introdução

Na última década tem-se assistido a alterações radicais nas sociedades onde vivemos. Democracias supostamente consolidadas, indivíduos mais dinâmicos, mais esclarecidos, mais exigentes e mais participativos, conformam uma sociedade de múltiplas oportunidades mas também de múltiplas exigências. Esta dinâmica estende-se ao cenário da vida, concretizando-se na interação social no desempenho de governos e organizações. Uma rápida evolução do mercado, com clientes cada vez mais exigentes, a abertura de fronteiras aos investidores externos como consequência de novos acordos internacionais entre os países e consequente alargamento de mercados, o aparecimento do conceito de globalização da economia juntamente com a introdução do computador exigiu e possibilitou o surgimento de sistemas de produção com elevado nível de automação e flexibilidade, capazes de responder às alterações no ambiente económico e tecnológico.

Este ambiente exerce grande pressão ao nível dos processos de fabrico e funcionamento de unidades industriais que no afã de manterem a sua competitividade e consequente sobrevivência, contemplam alterações mais rápidas nas configurações de fabrico dos equipamentos, possibilitando em simultâneo alta produtividade associada a produtos com grande qualidade. O patamar competitivo situa-se em níveis elevadíssimos, onde a supremacia ou prevalência se define ao nível do detalhe, corolário de uma exasperada competição industrial maioritariamente sustentada na minimização de custos (Machado et al., 2005).

Dentro deste espectro limitado de ações que possibilitam redução de encargos, a distinção dentro da competitividade vincula-se necessariamente à disponibilidade dos equipamentos para concretizar bens e serviços requeridos pelos consumidores.

Em ambientes industriais de produção, caracterizados por serem estocásticos, dinâmicos e por vezes caóticos, a gestão de perturbações assume crucial importância seja numa rápida resposta à ocorrência de perturbações inesperadas seja mediante o impedimento de que estas surjam. Genericamente, a ocorrência de perturbações provoca desvios nos planos de produção e usualmente degrada o desempenho dos sistemas, causando a perda de produtividade e oportunidades de negócio, tidas como fatores cruciais de competitividade (Tsang, 1998).

A função manutenção das unidades industriais apresenta-se como um elemento decisivo no crescimento da disponibilidade funcional dos equipamentos, tendo ao longo dos últimos anos

evoluído consideravelmente na forma como se posiciona perante as exigências que crescentemente lhe são incumbidas. A operação prolongada e eficaz dos sistemas produtivos é uma exigência. Impõe-se a necessidade de conhecer e controlar as possíveis falhas que venham a comprometer a missão produtiva. O objetivo consiste no prolongamento da atividade do sistema a plena carga e de modo contínuo, durante o tempo necessário, sem que seja afetado por defeitos em qualquer uma das suas partes integrantes (Machado et al., 2005).

Tipicamente as ações de manutenção ocorrem dentro das fronteiras da unidade industrial, num ambiente confinado às ocorrências ou histórico de ocorrências da própria organização, não levando em linha de conta a experiência acumulada pela atividade de unidades dispersas pelo mundo que recorrem a equipamentos similares e que experimentaram as suas próprias vicissitudes funcionais.

Uma das características mais salientes da automação industrial, fruto da incorporação de módulos inteligentes nos equipamentos, da capacidade de operação remota, da capacidade de decisão que advém de módulos de inteligência artificial, mas essencialmente da evolução dos sistemas de planeamento e controlo da produção, com ênfase na monitorização, é a produção de grandes quantidades de dados e seu armazenamento, maioritariamente inopinado, que adequadamente trabalhado se perspectiva de elevado valor no auxílio a ações de manutenção, com ênfase na vertente da predição e previsão em detrimento do binómio falha/correção (Lee, 2010).

1.1 Motivação e objetivos

Face ao observado anteriormente, tirando partido da oportunidade de intervenção que resulta do espaço de melhoria que representa a ação da manutenção industrial, aproveitando tecnologias consolidadas e outras emergentes, este trabalho, numa perspetiva mais macro, apresenta e especifica uma orgânica funcional que possibilita a agregação de históricos de manutenção distribuídos, bem como uma recolha dinâmica de novos dados que no seu conjunto, uma vez atuados com ferramentas adequadas, produzirão conhecimento a ser utilizado em ações de planeamento de manutenção ou indicando ações que de outra forma não seriam expectáveis.

As empresas realizam no seu quotidiano ações de manutenção de três tipos i) corretiva ou reativa, ii) preventiva sistemática e iii) preventiva condicionada ou preditiva. No decurso dessas intervenções produzem-se grandes quantidades de dados que são armazenados constituindo-se

como históricos de grande utilidade em ocorrências futuras. O armazenamento dos dados recolhidos é tipicamente realizado localmente, em formato eletrónico ou por vezes ainda em papel. Ao nível da manutenção corretiva ou reativa, a recolha de dados centra-se no registo dos elementos que foram observados antes da falha ocorrer, no diagnóstico realizado, no tipo de intervenção efetuada, no material utilizado, no processo de intervenção e resultados obtidos. O planeamento da manutenção preventiva baseia-se geralmente nas recomendações dos fabricantes das máquinas e registam-se essencialmente elementos que possibilitam aferir se, com base nas indicações do fabricante, o comportamento dos equipamentos converge com os comportamentos indicados nos manuais. A manutenção preditiva recorre a algum do conhecimento oriundo das outras duas formas de intervenção de manutenção, complementando a informação com dados de análises que vão gradualmente sendo realizadas ao equipamento, observando-se conseqüentemente grande criticidade no registo de dados no sentido de estes concorrerem ao estabelecimento de políticas que possibilitam atuar com eficácia em tempo útil.

Genericamente sabe-se que a integração das várias ações de manutenção nem sempre é realizada com a eficácia devida e mesmo quando tal ocorre é sempre limitada pelo facto anteriormente referido de a realidade da atuação se confinar à análise de ocorrências dentro das fronteiras de uma só unidade fabril.

Segundo Gonçalves (2009), os atuais sistemas de manutenção industrial estão a ficar obsoletos, inviabilizando assim uma resposta mais capaz às cada vez maiores exigências do sistema produtivo. Urge o desenvolvimento de sistemas de manutenção capazes de se moldarem aos requisitos dos ambientes produtivos atuais, de se potenciarem, de forma a maximizar a sua capacidade de resposta a falhas e acima de tudo evitar que estas ocorram.

Assim, um dos contributos pretendidos com esta abordagem consiste na especificação funcional de uma arquitetura responsável por, em primeira instância, alimentar todo o sistema com dados de índole diversa, indo desde a perspetiva de dados de controlo do sistema, a dados que resultaram de monitorizações e intervenções locais de carácter corretivo ou preventivo sistemático e que serão utilizados no sistema global de produção de conhecimento. A vertente de controlo possibilita a coordenação das entidades sob as quais recai a responsabilidade de fornecer ao sistema em cada momento os dados necessários à produção de conhecimento, podendo esta ser gerada de forma programada, sob requisição ou de forma endógena face à eminência de um evento pernicioso num parceiro concreto desta rede dinâmica.

A arquitetura distribuída necessária a este tipo de abordagem depara-se necessariamente com dificuldades decorrentes da heterogeneidade dos sistemas individuais, bem como com a necessidade de garantir a integridade, confidencialidade e inviolabilidade dos dados de cada participante na rede, pelo que existe a necessidade de criar um interface ou mecanismo que simultaneamente garanta essas características e que possibilite a transferência dos dados em formato comumente adotado e generalizado ao patamar comum de funcionamento.

Ao nível dos sistemas de controlo de produção, sensíveis às alterações no seu entorno e em consequência reativos e reconfiguráveis, os SMA constituem uma das soluções adotadas mais eficazes, principalmente porque apresentam descentralização de procedimentos sobre estruturas distribuídas e capacidade de adaptação à emergência de forma autónoma, o que possibilita a melhoria da capacidade de resposta rápida à mudança. Esta abordagem assevera-se como adequada numa arquitetura orientada à gestão de manutenção industrial na qual se pretende agregar dados distribuídos, recolhidos sempre que possível autonomamente, e difundidos em formato normalizado ao funcionamento do sistema global. A concretização do elemento agente na arquitetura, não sendo um objetivo deste trabalho, deverá incluir mecanismos ditos inteligentes embebidos nessas unidades de controlo distribuído, capazes de responder a estímulos remotos ou autonomamente produzidos face a alterações no “cais” de dados dos parceiros de forma a suportar as tarefas que lhe foram incumbidas no guião global do sistema, deverá orientar-se através de diretivas XML, que de entre as soluções possíveis, se apresenta como a mais robusta, com provas dadas, tecnologicamente disponível e capaz de convergir com os requisitos especificados nas funcionalidades apresentadas.

A arquitetura proposta, na sua vertente ativa, integra um componente de previsão, complementar às funções tradicionais de gestão de perturbações, permitindo perspetivar em avanço a sua ocorrência futura, minimizando ou eliminando o seu impacto. Esta capacidade preditiva, assente não apenas em dados locais e isolados mas também nos dados coligidos remotamente, transforma as práticas tradicionais de “falha e reage” em práticas “prevê e previne”, introduzindo a capacidade de antecipação de forma a melhorar o desempenho do sistema de intervenção, nomeadamente no que à rapidez na ação e decisão diz respeito.

No decurso da especificação das funcionalidades do sistema e duma coerente articulação e validação de cenários operativos, verificou-se a necessidade de desenvolvimento do sistema SPAMUF (Sistema de Predição de Avarias em Máquinas de Unidades Fabris globalmente dispersas), concretizado sob a forma de protótipo, conceito inovador e estruturante do sistema

que este trabalho propõe. O SPAMUF é um sistema de conhecimento baseado em DM, orientado para a predição de avarias, com incorporação de técnicas automáticas para aquisição de dados e visualização de resultados. Submetido a distintas configurações operativas fornece comportamentos distintos em função dos algoritmos e configurações com que é carregado, determinando ajustes ao processo de formação, formulação e dissolução de parcerias, no sentido de se obter coerência e consistência ao longo de todo o processo de funcionamento do sistema, bem como formulando a adequabilidade dos métodos por comparação a cada configuração.

1.2 Metodologia aplicada

Com base nos indicadores a serem fornecidos pelo sistema será possível estimar e compreender o comportamento dos equipamentos no que diz respeito à possibilidade de ocorrência de falhas. Assim, de forma fundamentada, poder-se-ão redefinir políticas de manutenção para cada equipamento e para os componentes neles inseridos, o que institui os objetivos deste trabalho.

Para a realização deste trabalho foram tidas em consideração as seguintes etapas:

1. Revisão da literatura existente sobre o domínio de aplicação, i.e., sobre manutenção industrial e tipos de manutenção existentes aplicados atualmente no domínio da produção industrial, bem como uma abordagem aos *softwares* de manutenção e a sua implicação na função manutenção. É elaborada uma abordagem aos modelos distribuídos para aquisição de dados, mais propriamente à aplicação de agentes automáticos para recolha de dados, sendo posteriormente abordada toda a temática do DM, bem como as propostas elaboradas por alguns grupos de trabalho nesta mesma área;
2. Revisão de literatura existente sobre os paradigmas a utilizar, com especial realce para aplicações em DM na manutenção. Nesta fase adquiriram-se conhecimentos sobre a utilização de ferramentas, técnicas e paradigmas, e analisaram-se trabalhos anteriores para identificar aspetos ainda não abordados;
3. Conceção do sistema, em que após uma análise profunda dos conhecimentos adquiridos nas fases anteriores, partiu-se para a conceção de um sistema capaz de apoiar pressupostos em aberto na resolução do problema apresentado nesta tese. O sistema gerado surge como forma de auxílio a diferentes organizações na utilização de dados

armazenados referentes a ações de manutenção e dados de monitorização, na perspetiva de indiciar a elaboração de ações de manutenção preventiva;

4. Desenvolvimento e implementação do protótipo através de um caso prático de estudo de forma a validar as soluções apontadas na especificação concetual do sistema. Nesta fase pretendeu-se passar para a prática as ideias concetualizadas na fase anterior. Recorrendo a uma ferramenta de DM foi desenvolvido um protótipo e aplicado num caso de estudo, com o qual se realizaram testes de forma a obter resultados que pudessem suportar todo o trabalho desenvolvido até então.

1.3 Organização e estrutura da Tese

A estrutura do presente trabalho pretende simplificar o acesso rápido por parte do leitor a todas as partes do mesmo. Desta forma o documento foi dividido em partes lógicas para permitir uma melhor compreensão por parte do leitor.

Este documento foi estruturado em 5 capítulos da seguinte forma:

- No capítulo 1 é introduzido o trabalho proposto, efectuando-se em primeiro lugar um enquadramento ao tema e apresentado posteriormente os objectivos e a motivação para a sua realização.
- O capítulo 2 apresenta o estado da arte relativamente a diferentes áreas abrangidas por este projeto. Na primeira secção, são abordados temas relativos à manutenção industrial, sendo introduzidos conceitos associados à manutenção e à fiabilidade. Classificam-se os tipos de manutenção focando a necessidade da sua conjugação na aplicação de uma melhor política de manutenção por parte de uma organização. Seguidamente, são descritos os métodos de apoio à gestão da manutenção que começaram a ser implementados nas últimas décadas, apresentando as suas principais características e últimos desenvolvimentos. Posteriormente, são abordados os sistemas informáticos de gestão da manutenção (CMMS) e os modelos de apoio à decisão na manutenção. Na secção seguinte, são abordados os modelos distribuídos de aquisição de dados, focando o modelo cliente/servidor e os conceitos teóricos acerca das características e funcionamento dos agentes automáticos de recolha de dados. Na terceira secção, são apresentados os conceitos associados ao DM e faz-se

o seu enquadramento no processo de DCBD. São abordadas as metodologias existentes para DCBD e é efetuada uma comparação entre elas. Seguidamente, são abordados os diferentes modelos e técnicas do DM a fim de verificar a sua aplicabilidade na obtenção dos objectivos propostos com este trabalho. Conclui-se este capítulo com a apresentação de alguns trabalhos abordando a aplicação do DM na área da manutenção industrial.

- No capítulo 3 é descrito o sistema funcional do sistema de predição proposto através do uso da metodologia IDEF0. É apresentada a sintaxe e semântica da metodologia e é realçada a mais-valia da sua utilização. São abordadas as tecnologias relativas ao XML e VPN realçando os benefícios da sua utilização na normalização dos *inputs* e *outputs* do sistema proposto. Seguidamente é apresentado o *interface Web* da aplicação desenvolvida neste trabalho bem como a descrição das linguagens de programação utilizadas na sua implementação. Este capítulo é concluído com a apresentação da ferramenta de DM onde a parte fulcral do sistema é gerada.
- O capítulo 4 descreve o protótipo construído, que servirá de demonstrador da conjectura considerada ao longo da Tese. É efetuada uma visualização da estrutura de dados que serve de *input* a esse mesmo protótipo sendo de seguida apresentados os processos que o constituem. Neste capítulo é também apresentado o caso prático efetuado de forma a validar as conceções apresentadas aquando do desenvolvimento do protótipo.
- No capítulo 5 são apresentadas as conclusões e contribuições do trabalho desenvolvido, bem como uma visão das perspectivas de trabalho a desenvolver futuramente de forma a abranger um maior número de empresas, com todas as condicionantes e possíveis melhorias que isso poderá envolver.
- No final do documento são apresentadas as referências bibliográficas que foram citadas ao longo do texto, bem como apêndices com informação adicional relativa a partes constituintes da tese e acerca do código XML gerado pela aplicação de DM utilizada no desenvolvimento do protótipo apresentado nesta tese.

Capítulo 2

2. Estado da arte

Neste capítulo são abordadas as matérias imprescindíveis para compreender o sistema gerado. Salienta-se, em primeiro lugar e de forma breve, a importância da função manutenção para as organizações. Neste âmbito, são descritos aspetos que permitem compreender a evolução da manutenção através da perspectiva histórica e expõe-se uma revisão bibliográfica acerca das várias metodologias de manutenção existentes. É analisado o estado da arte relativo aos modelos distribuídos de aquisição de dados, bem como todos os conceitos e fundamentos teóricos referentes ao objetivos, metodologias, técnicas e algoritmos de DM que aplicados a dados funcionais de equipamentos justificarão o trabalho e conclusões apresentados nas secções subsequentes.

2.1 Manutenção industrial

A manutenção é definida como a combinação de todas as ações técnicas, administrativas e de gestão, durante o ciclo de vida de um bem, destinadas a mantê-lo ou repô-lo num estado em que ele possa desempenhar a ação requerida (NP-EN-13306, 2007).

A função manutenção tem-se tornado aos olhos dos gestores cada vez mais importante devido ao seu papel de manter e melhorar a disponibilidade dos sistemas, garantir a segurança dos operadores e a qualidade do produto (Muller et al., 2008).

Genericamente, assume-se que o papel da manutenção industrial é o de preservar o funcionamento adequado de um sistema físico, para que este continue a desempenhar a função para a qual foi projetado. Ou seja, a manutenção industrial pode ser definida como o conjunto de atividades cuja realização pretende preservar ou melhorar a segurança, desempenho, fiabilidade e disponibilidade das estruturas da unidade fabril, dos sistemas e componentes no sentido de assegurar um desempenho adequado quando tal for solicitado (Weinstein e Chung, 1999). Esta preocupação funcional tem implicações com os produtos gerados, custo unitário e racionalidade no consumo de energia do sistema em questão, expandindo a sua influência a questões como sejam a qualidade do produto final, controlo do processo ou segurança do/dos operários afetos ao sistema físico, conformidade com normas de proteção ambiental e integridade estrutural, contemplando inclusive a aparência física do sistema.

Nos ambientes produtivos modernos exigem-se elevados níveis de fiabilidade dos equipamentos envolvidos no processo produtivo. Por vezes torna-se importante a substituição de um item antes de uma possível falha, sendo que a possibilidade dessa mesma falha aumenta com a idade e utilização desse mesmo equipamento (Nakagawa e Mizutani, 2009). Os avanços tecnológicos permitiram, por um lado, um incremento dos níveis de eficiência, e por outro mudaram a forma como estas instalações e equipamentos têm sido tradicionalmente mantidos. Esta perspetiva estende-se à ação de manutenção, possibilitando novas formas de despontar uma determinada ação de manutenção (Haider e Koronios, 2006). As modernas ferramentas de manutenção disponíveis não sendo “a panaceia”, auxiliam de sobremaneira a ação do operário, reduzindo o esforço físico necessário e diminuindo os tempos de intervenção, bem como incrementam a precisão dessa mesma intervenção. Por outro lado, esta evolução exige maiores níveis de qualificação dos operadores.

Numa outra perspetiva, a redução dos custos de produção com base na redução de *stocks* provocou um incremento da vulnerabilidade ao risco, pelo que a redução das avarias nas máquinas se tornou uma premência. Em conformidade com este sentimento a manutenção viu o seu papel reconhecido, tornando-se num fator chave da produção (Luce, 1999).

A necessidade de convergir com os requisitos explicitados por parte das empresas, provocou grande pressão nos sistemas de manutenção das fábricas. A função manutenção, tida como não agregadora de valor, é cada vez mais solicitada no sentido de contribuir para a redução de custos, mantendo, simultaneamente, as máquinas em bom estado de funcionamento (Bansal et al., 2005a). Desta forma, a atividade de manutenção é cada vez mais reconhecida como decisiva em termos de competitividade dos negócios. A manutenção deve pois estar devidamente integrada no plano estratégico da empresa. Os seus objetivos deverão estar enquadrados nos objetivos, de carácter mais geral, definidos para toda a empresa bem como em sintonia com os objetivos da produção.

Tendencialmente, as empresas de produção reconhecem os elevados custos associados às avarias que ocorrem nas linhas de produção. A ocorrência desses eventos implica custos com inspeções, reparações, bem como custos inerentes ao tempo de não produção/utilização dos equipamentos. Nesta situação existe ainda a questão da segurança ou saúde, no caso de a avaria ser potencial provocadora de lesões nos operários (Edwards et al., 1998).

A estratégia da manutenção, que é influenciada por fatores internos e externos, é concretizada por políticas de manutenção adequadas. Esta estratégia deverá depender da estratégia global

definida para a empresa e será particularmente relevante a sua ligação e interdependência com a que for estabelecida para a produção (Edwards et al., 1998).

A qualidade da manutenção desenvolvida afeta diretamente o desempenho de um determinado negócio. Os fatores que sustentam esta premissa passam pela resolução de avarias, segurança, disponibilidade operacional dos equipamentos, melhor qualidade dos produtos, entre outros. O mundo atual caracteriza-se pela existência de grandes expectativas, pelo aparecimento de constrangimentos normativos cada vez mais caros, pelas mudanças constantes ao nível dos paradigmas tecnológicos e pelas reorganizações estratégicas e/ou funcionais constantes e urgentes com consequências potencialmente gravosas para as equipas de manutenção (Moubray, 1997).

As tecnologias associadas à função manutenção encontram-se num percurso de descoberta e aplicação de formas financeiramente adequadas e eficazes para evitar a deterioração do desempenho. As técnicas de gestão de avarias incluem ações preditivas e preventivas, procura de avarias, funcionar até à falha e alterações ao projeto do sistema produtivo no local, ou na forma como este é operado. Cada categoria inclui um leque de opções, onde algumas são claramente mais eficazes do que outras (Siqueira, 2005).

A equipa de manutenção não tem apenas que estar atenta ao leque dessas opções, tendo também que decidir quais as mais adequadas em cada situação. Fazer uma escolha sensata, deverá melhorar o desempenho do sistema, bem como reduzir o custo global. No entanto, fazer uma escolha errada pode contribuir para o surgimento de novos problemas, enquanto os existentes podem piorar. Assim, deve-se sensibilizar as equipas para que, de entre as possibilidades de atuação, tomem as decisões mais sensatas do ponto de vista custo/eficácia (Moubray, 1997).

As avarias normalmente atraem a atenção porque podem afetar negativamente a produtividade, segurança, saúde ambiental, qualidade, serviço ao cliente, competitividade ou custo unitário. A severidade ou frequência com que os responsáveis pela manutenção lidam com as referidas consequências ditará qual a técnica de gestão de avarias a aplicar. Assim, a missão poderá apontar para a definição de qual a consequência a evitar. A política a aplicar deverá ser eficaz na utilização dos recursos (pessoas, materiais, peças de reposição, ferramentas, etc). O custo de manutenção depende não apenas da equipa de manutenção, mas também dos projetistas e operadores de um dado equipamento/sistema.

Atualmente, devido às crescentes exigências de melhoria de desempenho e menor preço, as empresas têm enveredado pela tomada de diferentes atitudes de forma a se manterem competitivas, atendendo a estas exigências e reduzindo os seus custos.

2.2 Gestão da manutenção

A enorme competitividade verificada hoje em dia em empresas concorrentes é muito elevada. Só empresas bem preparadas e com um bom programa de manutenção de forma a evitar prejuízos inerentes a falhas ou avarias de máquinas ou componentes e ao mesmo tempo acrescentar valor ao produto, se tornam competitivas e dessa forma conseguirão assegurar o seu futuro. A maioria das empresas pode reduzir os seus custos com a manutenção, melhorando assim os níveis de produtividade

A norma NP-EN-13306 (2007) define a gestão da manutenção como consistindo em “todas as atividades de gestão que determinam os objetivos, a estratégia e as responsabilidades referentes à manutenção e que os implementam por diversos meios tais como o planeamento, o controlo e supervisão da manutenção e a melhoria de métodos na organização, incluindo os aspetos económicos”.

Segundo Wireman (2004), a gestão da manutenção tem como objetivos:

- Assegurar a produção com o menor custo possível, com a qualidade desejável e dentro dos padrões de segurança exigíveis;
- Identificar e reduzir custos referentes aos processos de manutenção;
- Elaborar registos relativos a ações de manutenção;
- Otimizar os recursos disponíveis para ações de manutenção (humanos e materiais);
- Otimizar a vida útil do equipamento;
- Minimizar o consumo de energia;
- Minimizar o stock de consumíveis

A gestão da manutenção consiste numa série de atividades integradas entre si, que devem ser vistas como um todo e não de forma individualizada, já que são impossíveis de se desagregar (Kans, 2008).

No entanto o desgaste de um equipamento é um processo normal e inevitável. Cada equipamento tem o seu próprio grau de uso permissível no que respeita à totalidade das funções

que executa que, sob condições específicas, está inerente àquele equipamento, quer devido ao seu fabrico quer devido à sua recuperação depois de ações de manutenção.

Desta forma é necessário associar a temática da gestão da manutenção à gestão de ativos físicos, mais propriamente na norma ISO-55001 (2014), que fornece uma visão geral da gestão de ativos, princípios, terminologia e benefícios esperados com a adoção da gestão de ativos. Especifica requisitos para um sistema de gestão de ativos dentro do contexto da organização podendo ser aplicada a todos os tipos de ativos, e por todos os tipos e tamanhos de organizações.

A gestão de ativos surge como algo mais abrangente do que a gestão da manutenção porque envolve todo o ciclo de vida do equipamento desde a aquisição do mesmo e até à sua desativação.

2.2.1 Sistemas reparáveis e não reparáveis

Para melhor compreensão dos conceitos explanados nas seguintes subseções é importante começar por distinguir os conceitos de falha e de avaria. Define-se falha como o fim da capacidade do elemento para executar a função requerida, depois de se dar a falha, o elemento encontra-se avariado, assim sendo, pode-se dizer que a avaria é um acontecimento e em falha é um estado (NP-EN-13306, 2007).

É importante também salientar a necessidade de distinção e esclarecimento dos conceitos de sistemas reparáveis e não reparáveis devido ao facto de a previsão e quantificação da fiabilidade se efetuar de forma distinta, conforme o tipo de sistema.

Um sistema reparável, normalmente associado ao conceito de equipamento, a ocorrência de uma ou mais avarias não significa o fim da sua operacionalidade, mas apenas de determinada descontinuação enquanto são repostas as condições nas quais possa, novamente e de forma favorável, desempenhar as funções entretanto interrompidas. Contrariamente, num sistema não reparável, normalmente associado ao conceito de componente, não sendo económica ou tecnologicamente passível de reparação, só ocorre uma única falha, após a qual irá ser substituído por outro do mesmo género em virtude de ter ficado inutilizado (Rigdon, 2008).

Como exemplos de sistemas reparáveis poder-se-ão referir todos os géneros de eletrodomésticos, como por exemplo um secador ou uma máquina de lavar roupa. Por sua vez, como exemplos de sistemas não reparáveis poder-se-ão citar bens, tais como, lâmpadas, rolamentos, filtros de óleo e correias de distribuição de um automóvel.

A necessidade de diferenciar e esclarecer os conceitos de sistemas reparáveis e não reparáveis deve-se ao facto de a previsão e quantificação da fiabilidade se efetuar distintamente, consoante se trate de um, ou de outro, dos sistemas.

2.2.2 Fiabilidade

As normas internacionais definem fiabilidade como a probabilidade de um sistema desempenhar corretamente, sem falhas, a função requerida durante um dado período de tempo e em condições específicas (Teixeira, 2008).

Trata-se evidentemente de um conceito que relaciona o estado de funcionamento com o tempo, podendo esta métrica ser substituída por outra, por exemplo, horas de trabalho, quilómetros percorridos, ciclos de operação executados, etc. Assim sendo, é necessário definir claramente o critério de falha, tornando possível determinar de forma clara o momento a partir do qual se considera o desempenho abaixo do aceitável, ou seja, uma falha.

A manutenção é essencial na preservação da fiabilidade dos equipamentos e do ambiente produtivo. De modo geral, a fiabilidade consiste no grau de confiança ou probabilidade atribuída ao funcionamento sem falhas por parte de um sistema, em determinado ambiente e durante um determinado período de tempo (Morais, 2007).

A fiabilidade é definida como a probabilidade de um equipamento desempenhar adequadamente as funções para as quais é requerido, durante um período especificado de tempo e nas condições normais de funcionamento (Lopes, 2007).

A definição da fiabilidade caracteriza-se por dois aspetos importantes. Por um lado, o tempo durante o qual o equipamento é solicitado e, por outro lado, as condições operacionais e ambientais de funcionamento. Os equipamentos são projetados tendo em conta determinadas condições ambientais e de funcionamento, sendo que qualquer alteração dessas condições pode provocar a falha prematura desses mesmos equipamentos. Assim sendo, o comportamento do equipamento deixa de ser previsível e a previsão da fiabilidade deixa de ser válida (Lewis, 1996).

A função fiabilidade $R(t)$ é definida matematicamente como a probabilidade de um sistema sobreviver pelo menos durante um período de tempo t (Lewis, 1996) (2-1).

$$R(t) = \{\text{Probabilidade de um item operar sem falhas durante um tempo } t\} \quad (2-1)$$

De forma a relacionar matematicamente o conceito de estado de funcionamento com o tempo considere-se a variável T como o tempo de vida de determinado sistema não reparável ou tempo até à próxima falha no caso de equipamentos reparáveis. A fiabilidade para o tempo t será expressa como a probabilidade de $T > t$, sendo a função fiabilidade, designada por $R(t)$, expressa por (2-2):

$$R(t) = P_r(T > t), \quad (2-2)$$

Com o $R(t) \geq 0$.

A cessação de funcionamento de um equipamento provocará a avaria, sendo a sua probabilidade nesse instante t expressa pela função probabilidade de falha. Demonstra-se por $F(t)$ (função acumulada de falhas) e consiste numa função complementar de $R(t)$, assim (2-3):

$$F(t) + R(t) = 1 \quad (2-3)$$

Desta equação é possível deduzir a seguinte expressão para a função $F(t)$ (2-4):

$$F(t) = 1 - R(t) \quad (2-4)$$

Ou seja, $F(t)$ expressa a probabilidade da falha ocorrer até um tempo t , enquanto $f(t)$ representa a função densidade de probabilidade de falha.

$F(t)$ é crescente no tempo, atingindo o valor unitário quando t tende para infinito, tal é representado na Figura 2-1.

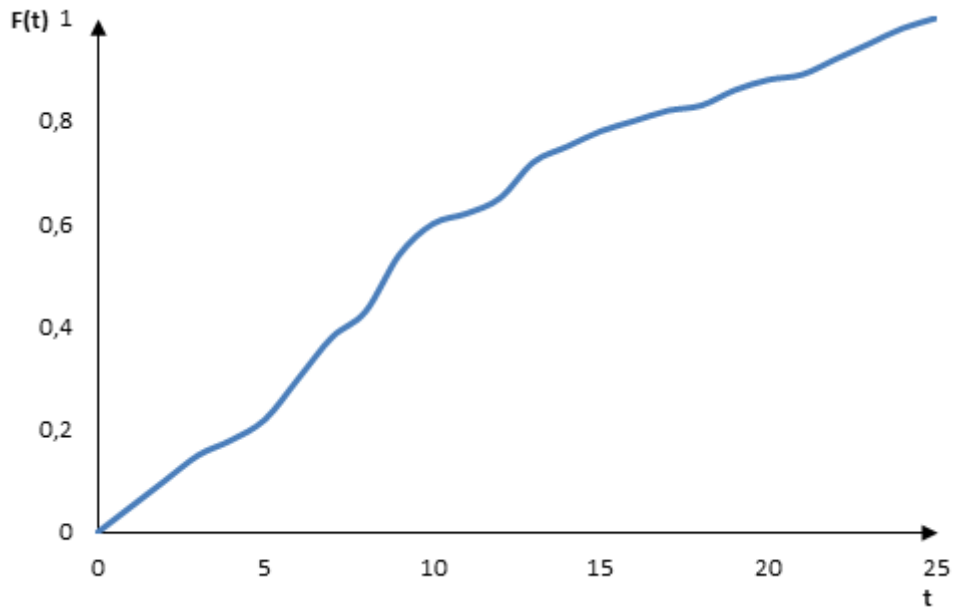


Figura 2-1 - Função de probabilidade de falha acumulada. Adaptado de Teixeira (2008).

Uma vez conhecido um critério objetivo para definição da falha, considera-se que o estado de falha e o estado de operação adequada são mutuamente excludentes.

A Figura 2-2 mostra a função fiabilidade.

Como corolário, tem-se que (2-5) (2-6):

$$R(0) = 1 \quad (2-5)$$

$$R(\infty) = 0 \quad (2-6)$$

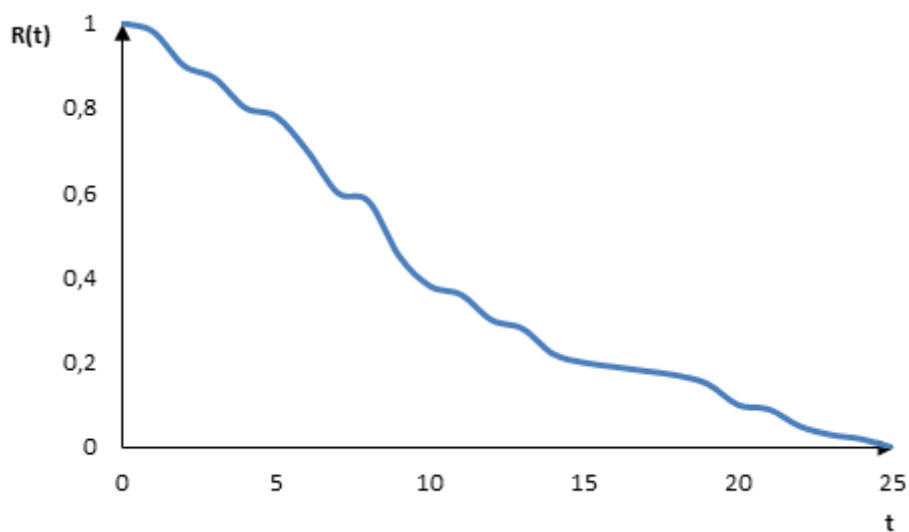


Figura 2-2- Função de fiabilidade. Adaptado de Teixeira (2008).

$R(t)$ e $f(t)$ são relacionadas pela seguinte expressão (2-7):

$$f(t) = -\frac{d}{dt}R(t) \quad (2-7)$$

Analisando o gráfico da Figura 2-2, verifica-se que o grau de fiabilidade de um item decai ao longo do tempo, ou seja, quanto maior o tempo de operação, maior será a probabilidade de ocorrer a falha. Da mesma forma, apenas no instante em que o equipamento é colocado em operação é que a sua fiabilidade é de 100% e a sua degradação procede ao longo do tempo.

Concluindo, a função $R(t)$ é associada à fiabilidade ou probabilidade de funcionamento, a função $F(t)$ está relacionada com probabilidades acumuladas de falha e a função $f(t)$ é usada para caracterizar a distribuição de falhas. Considerando os conceitos de sistemas reparáveis e não reparáveis anteriormente descritos, facilmente se compreende que o conceito de fiabilidade não poderá ser interpretado literalmente da mesma forma em ambos os casos. A devida distinção e a forma de quantificação, consoante se trate de um dos dois géneros de sistemas, são assuntos abordados nas próximas subsecções.

2.2.2.1 Fiabilidade de sistemas reparáveis

A fiabilidade de um sistema reparável, representa a probabilidade de não se verificar qualquer género de avaria durante um dado período de tempo, no qual se considera possível a ocorrência de mais que uma avaria (O'Connor e Kleyner, 2012).

A avaliação da fiabilidade de um sistema reparável pode ser feita através da taxa de avarias ou de ocorrência de falhas de um sistema. Esta taxa de avarias $\lambda(t)$ (ou ROCOF), pode ser definida como sendo derivada em ordem ao tempo do número esperado de falhas até ao instante t . Segundo Lewis (1996) também pode ser definida como a “variação do número esperado de avarias relativas a um determinado instante de tempo”, e expressa na seguinte expressão matemática (2-8):

$$\lambda(t) = \frac{d E[N(t)]}{dt} \quad (2-8)$$

Ou seja, a derivada em ordem ao tempo da esperança matemática do número de avarias no instante t .

A taxa de avarias, é um conceito que deve ser relacionado a acontecimentos repetitivos e observados por unidade de tempo, constituindo-se conseqüentemente, como um indicador utilizado, exclusivamente, na quantificação da fiabilidade de sistemas reparáveis.

O gráfico que representa a variação da taxa de avarias ao longo do tempo é designado por Curva da Banheira (Figura 2-3).

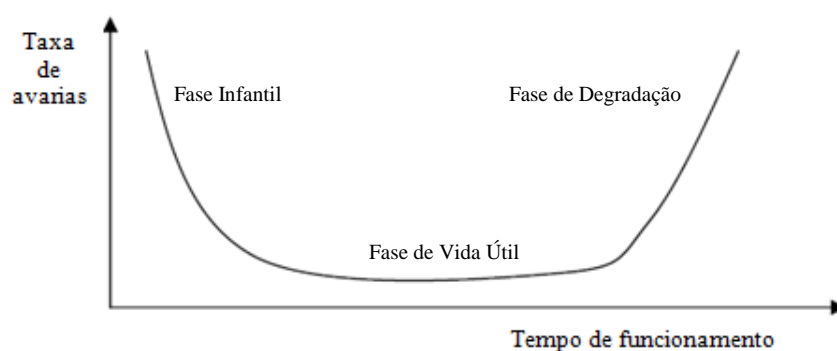


Figura 2-3 - Curva da banheira. Adaptado de Torell e Avelar (2004).

A Curva da Banheira representa o comportamento de um sistema perante a falha. Podem-se identificar três fases que se designam por fase infantil, fase de vida útil e fase de degradação.

Na fase infantil poderá existir grande incidência de componentes ou equipamentos com defeito de fabrico, de projeto, erro de aplicação ou erro de instalação. A taxa de avarias decai com o tempo sendo este um período de tempo curto em que a taxa de avarias é elevada mas decrescente.

Na fase de vida útil, a taxa de avarias é sensivelmente menor e relativamente constante ao longo do tempo. A ocorrência de falhas advém de fatores menos controláveis, como fadiga por sobrecarga ou desgaste acelerado devido à interação dos materiais com o meio. Este tipo de falha caracteriza-se pela dificuldade de predição/prevenção.

Na fase de degradação, verifica-se um aumento da taxa de avarias decorrente do desgaste natural (atrato, corrosão) originada por efeitos cumulativos tais como a fadiga, a corrosão ou o desgaste.

2.2.2.1.1 Tempo médio entre falhas

O tempo médio entre falhas, MTBF, é uma medida da fiabilidade de sistemas reparáveis. O MTBF mede o tempo médio durante o qual o equipamento permanece em funcionamento até ocorrer uma avaria. O valor instantâneo ou pontual do MTBF é obtido pelo inverso da taxa de avarias. De seguida é apresentada a fórmula de cálculo do MTBF para sistemas na fase de vida útil (2-9):

$$\emptyset = \frac{T}{R} \quad (2-9)$$

\emptyset = MTBF

T = Tempo total

R = Número de avarias

O tempo de funcionamento até à avaria T_i tem especial interesse nos testes de tendência, que procuram determinar se as avarias ocorrem ou não de uma forma aleatória (Figura 2-4). Nos casos em que existe tendência na ocorrência de avarias (taxa de avaria variável com o tempo), a taxa de avarias pode ser modelada pelos modelos de Crow ou de Cox & Lewis. O respetivo MTBF também será variável e dado pelo inverso da taxa de avarias.

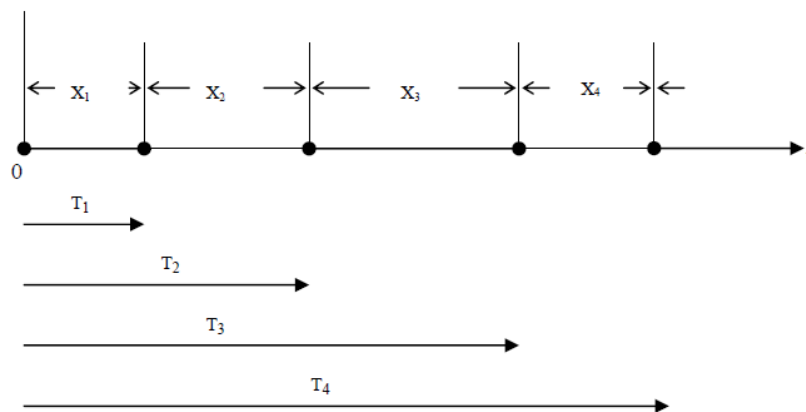


Figura 2-4 - Tempo entre avarias e tempo de funcionamento até à avaria.

2.2.2.2 Fiabilidade de sistemas não reparáveis

Foi visto que a fiabilidade é descrita como a probabilidade de um sistema funcionar sem avarias durante um certo período de tempo. Descreveu-se, igualmente, que num sistema não reparável,

só acontece uma e apenas uma falha. Por conseguinte, a fiabilidade de um sistema não reparável, normalmente redutível ao componente ou item, traduz a probabilidade desse sistema novo operar sem incidentes durante um certo período T .

Segundo O'Connor e Kleyner (2012), a fiabilidade de um sistema não reparável é caracterizada pela probabilidade de funcionamento, para além de um dado período de tempo, onde apenas uma única avaria pode ocorrer. Perante sistemas não reparáveis, a fiabilidade pode ser avaliada e quantificada através da análise estatística de tempos de sobrevivência de vários componentes do mesmo género presentes, em um, ou mais equipamentos. Estes tempos que se caracterizam pelo tempo que decorre desde a entrada em serviço dos componentes novos e a sua falha são representados por *time to failure* (*tff*'s).

Através da informação oriunda dessa análise estatística é possível obter uma estimativa para o tempo médio de falha, MTTF - *Mean Time to Failure*, que ao indicar o tempo médio de sobrevivência até à ocorrência da falha, proporciona indicações acerca da performance do sistema não reparável em estudo.

O MTTF pode, alternativamente, ser obtido recorrendo à expressão matemática que define o valor esperado para uma variável contínua (2-10):

$$MTTF = \int_0^{\infty} tf(t)dt \quad (2-10)$$

2.2.3 Tipos de manutenção

Em termos gerais, a engenharia tem tido uma grande preocupação com o alargamento do tempo de funcionamento dos equipamentos e redução da ocorrências de falhas, traduzindo-se numa maior disponibilidade destes. Tem havido uma preocupação com o aumento da fiabilidade dos equipamentos (Pereira e Sena, 2012).

Tal como descrito anteriormente, a fiabilidade consiste na capacidade do sistema em atingir os seus objetivos ou cumprir as funções para o qual foi projetado. Num contexto de mercado internacional, as organizações necessitam de melhorar as suas estratégias de manutenção de forma a melhorar a sua produtividade. Neste contexto as estratégias de manutenção são empregues para melhoria da fiabilidade. Diferentes tipos de políticas de manutenção podem ser

aplicadas: preventivas ou corretivas de acordo com o responsável de manutenção (Vrignat et al., 2012).

Uma manutenção eficaz passa pela integração entre a função manutenção e a engenharia do processo na fase de seleção e instalação de máquinas e equipamentos, e estendendo-se a ações pró-ativas sobre essas mesmas máquinas e equipamentos quando em funcionamento, passando necessariamente por intervenções de manutenção preventiva, preditivas e modificações ao nível do projeto (Palmer, 1999). A manutenção é assim tida como uma função vital de suporte ao negócio, especialmente motivada pela crescente necessidade de grandes investimentos ao nível dos recursos físicos e a necessidade de os preservar nas melhores condições possíveis.

A manutenção tem um papel fundamental na obtenção dos objetivos da organização, por contribuir para o alcance dos objetivos produtivos, quer em termos de prazos, quer em qualidade. Existindo diferentes tipos de manutenção, importa definir, para cada equipamento, qual a política mais adequada face à importância desse equipamento para o processo produtivo e de acordo com as necessidades de fiabilidade e disponibilidade que o caracterizam. Numa mesma organização, poderão estar presentes os diferentes tipos de manutenção, aplicados a diferentes equipamentos, face às suas características próprias e à importância para o processo produtivo. A manutenção deverá, também, intervir no processo de seleção e aquisição de novos equipamentos, por forma a contribuir, com a sua experiência, para a escolha do equipamento que melhor venha a responder às exigências e constrangimentos do contexto em que vão laborar.

Um planeamento de manutenção eficaz envolve o desenvolvimento de um conjunto de ações que incluem ações de manutenção preventiva, reparação, resolução de problemas, etc. É exigido a identificação de um grande número de fatores, tais como, ordens de serviço, planeamento de trabalho, planeamento de materiais, agendamentos, etc. Quase a totalidade destas ações dependem do processamento de dados convencionais. A intervenção humana ou avaliação é necessária aquando da resolução de problemas e do planeamento de ações de manutenção preventiva.

A literatura (Chu et al., 1998; Pinjala et al., 2006) aponta de uma forma genérica três tipos de manutenção (Figura 2-5) que são descritos nas subsecções seguintes.

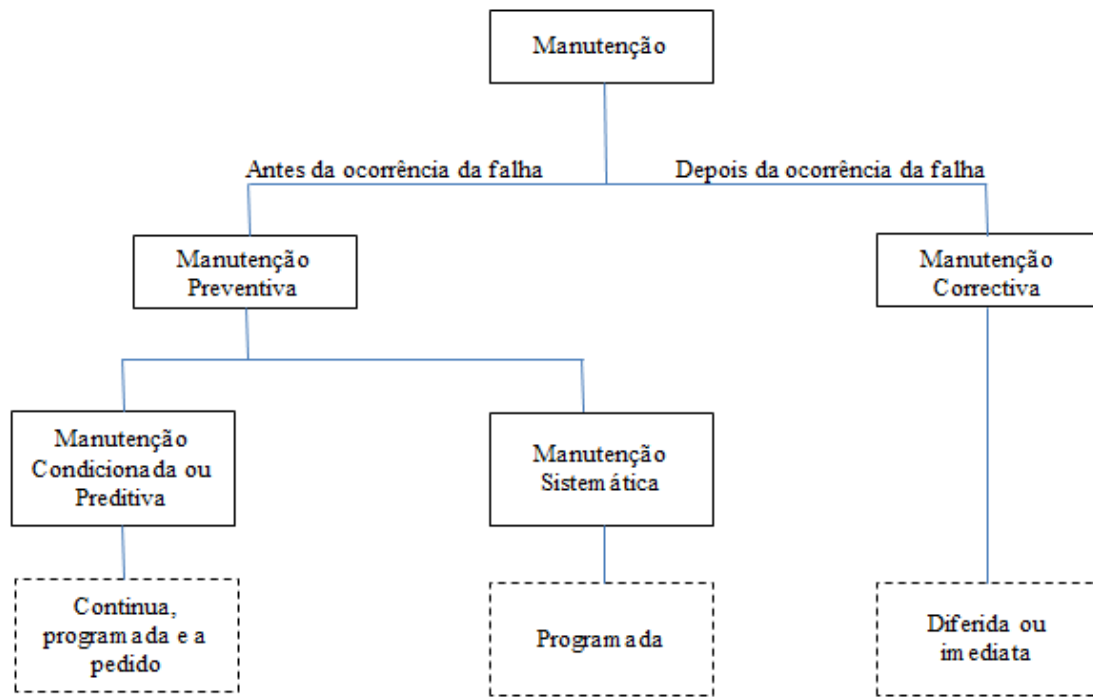


Figura 2-5 - Tipos de manutenção. Adaptado de NP-EN-13306 (2007).

Uma boa política de manutenção requer uma conjugação destes três tipos de manutenção (Figura 2-5) de forma a serem aplicadas onde e quando necessárias, eliminando assim ações de manutenção fora de tempo ou desnecessárias. Apesar dos avanços nas tecnologias de manutenção preditiva, a aplicação de operações de manutenção correctiva e preventiva sistemática ainda são norma em muitos processos produtivos. Hoje em dia, quase 30% dos equipamentos industriais não beneficiam de tecnologias de manutenção preditiva (Hashemian e Bean, 2011).

2.2.3.1 A manutenção correctiva

A manutenção correctiva ou reativa, que consiste basicamente em intervir no equipamento apenas mediante a ocorrência de uma falha. O equipamento está em funcionamento até que a falha ocorra, momento em que é reparado ou substituído. As desvantagens desta abordagem incluem capacidade de produção imprevisível e flutuante, níveis elevados de produtos não conformes e sucata, custos elevados de manutenção devido ao custo de reparação de falhas catastróficas (Bateman, 1995), bem como custos associados a *stocks* de peças de substituição e ao trabalho extra. A manutenção correctiva não procura evitar falhas, no entanto tenta antecipar os requisitos dessas mesmas ações de manutenção de forma a poder saber qual é o número de

peçoal necessário e os *stocks* de peças de substituição necessários. Uma unidade fabril que utilize somente este tipo de manutenção deverá ser capaz de reagir a todas as possíveis falhas dentro da fábrica. Este método reativo de gestão leva a que o departamento de manutenção mantenha *stocks* de peças suplentes podendo incluir equipamentos de reserva ou, pelo menos, os componentes principais dos equipamentos críticos da unidade fabril. De forma a minimizar o impacto na produção criada por falhas inesperadas dos equipamentos, as equipas de manutenção devem estar aptas a reagir imediatamente às falhas. O resultado de uma abordagem somente reativa traduz-se num maior custo de manutenção e menor disponibilidade dos equipamentos no processo produtivo (Swanson, 2001).

2.2.3.2 A manutenção preventiva

A manutenção preventiva consiste em realizar operações de manutenção periódicas aos sistemas de forma a prevenir a ocorrência de avarias. Consiste em intervenções efetuadas a intervalos de tempo pré-defenidos ou de acordo com critérios prescritos, com a finalidade de reduzir a probabilidade de avaria ou a degradação de um equipamento. Distinguem-se dois tipos de manutenção preventiva, a sistemática onde as revisões ou substituições de itens são efetuadas segundo um programa preestabelecido ou baseado no tempo ou no número de unidades de utilização (ex.: quilómetros, ciclos, etc.); e a condicionada, onde a manutenção é executada se um diagnóstico revelar um determinado grau de degradação (Lopes, 2007).

2.2.3.2.1 Manutenção preventiva sistemática

Neste caso particular os dados de funcionamento, manuais de fabricantes e algumas heurísticas e modelos matemáticos são a base para a definição dos períodos de manutenção (um exemplo trivial reside nas revisões automóveis, realizadas a cada 30.000Km, ou em máquinas, por exemplo após 100 horas de trabalho). A ocorrência de avarias em máquinas depende essencialmente da idade e utilização destas, mas não se deve descurar a aplicação de técnicas de manutenção preventiva no sentido de minimizar essas ocorrências durante a sua vida útil (Gharbi e Kenné, 2005).

A aplicação da manutenção sistemática não evita completamente a rutura. Ela pode apresentar um custo elevado, pois por vezes são substituídos componentes que não atingiram nem a metade da sua vida potencial. Este custo assume uma maior expressão quando a

substituição de peças implica a desmontagem parcial ou total do equipamento, com a consequente possibilidade de paragem da produção (Yao et al., 2004).

2.2.3.2 Manutenção preventiva condicionada

É sabido que cerca de 99% das avarias é precedido de algum indicador (Bloch e Geitner, 1997). Desta forma a manutenção preditiva poderá provavelmente ser vista como uma forma económica de manter uma máquina em produção. A ideia principal consiste em permitir a determinação da “saúde” de uma máquina em tempo real (*online*). Como tal, as falhas podem ser previstas antes de ocorrerem. A manutenção poderá ser programada conforme as necessidades. Os benefícios pretendidos através da manutenção condicionada incluem a redução do tempo de inatividade, menores custos de manutenção e redução de situações inesperadas de falhas catastróficas (Lin e Tseng, 2005).

Desta forma, a manutenção preventiva condicionada ou preditiva baseia-se na vigilância do funcionamento do equipamento/sistema através de parâmetros significativos do seu funcionamento, ou seja, trata de estimar o estado de funcionamento do equipamento/sistema através da medição de parâmetros vitais de operação, isto é, este tipo de manutenção é baseada nos valores observados para estes parâmetros. Este tipo de manutenção estima o estado de funcionamento de um sistema através da medição de parâmetros pré-determinados que revelam o estado de degradação de um determinado equipamento. Poder-se-à dizer que acompanha o desenvolvimento da falha, apontando o momento mais adequado da intervenção. Isto pode significar o aumento consciente da utilização de determinado item, traduzindo-se em economia quer de recursos, quer financeira (Teixeira, 2008).

Os parâmetros de controlo mais comuns são (Lu et al., 2007):

- Espectro em frequência do nível de vibrações do equipamento (aplicável principalmente a equipamentos com órgãos rotativos, ou outros que sejam sujeitos a esforços dinâmicos);
- Análise de óleos. Permite testar a aptidão do lubrificante e estimar o estado de degradação dos órgãos lubrificados mediante a análise das partículas em suspensão (aplicável principalmente a equipamentos transmissores de potência mecânica: redutores, diferenciais, etc.);
- Exame do estado de superfícies;
- Exame da estrutura de materiais;
- Exames de dissipação de energia;
- Etc.

A manutenção preventiva condicionada consiste numa estratégia tendo em vista a gestão da falha, baseia-se na capacidade das equipas de manutenção para detetar possíveis falhas com antecedência, a fim de tomar as medidas apropriadas. A curva comum que ilustra o comportamento do equipamento à medida que se aproxima de uma falha é a denominada curva de falha PF (Moubray, 1997).

A curva mostra que à medida que uma falha começa a manifestar-se, o equipamento deteriora-se até ao ponto em que, eventualmente, pode ser detetada (P). Se a falha não é detetada e mitigada, ela continua a degradar-se até que uma falha “mais grave” ocorra (F) (Figura 2-6).

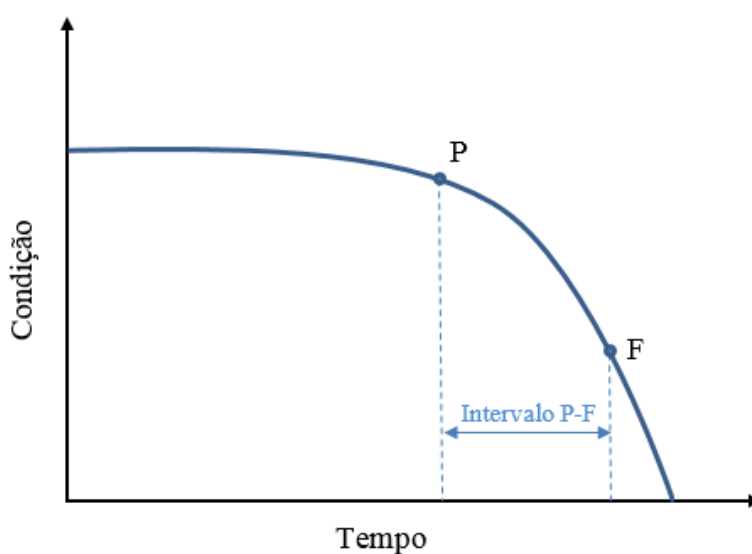


Figura 2-6 - Curva P-F. Adaptado de Moubray (1997).

O intervalo de tempo entre P e F, comumente designado de intervalo P-F, consiste na janela de oportunidade durante o qual uma inspeção pode, eventualmente, detetar a falha iminente e fazer-lhe frente (Moubray, 1997).

Este tipo de manutenção tem por objetivo apoiar as organizações no desenvolvimento estratégico através da aplicação de técnicas de inspeção e análise de condição não intrusiva, melhorando a disponibilidade dos equipamentos e reduzindo as paragens não programadas.

A manutenção preventiva condicionada é constituída por três processos chave (Figura 2-7) (Jardine et al., 2006):

- A fase de aquisição de dados, responsável pela aquisição e incremento de dados em sistemas informáticos;
- O processamento de dados onde é efetuado o manuseamento e tratamento dos dados recolhidos na fase anterior, para melhor entendimento e interpretação destes;

- Tomada de decisão de manutenção de forma a recomendar medidas eficientes de manutenção.

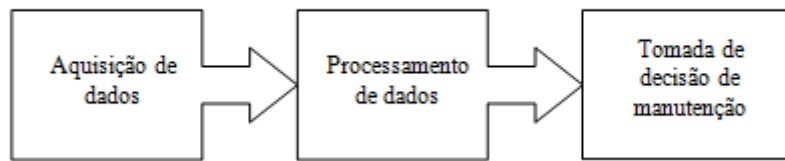


Figura 2-7 - Passos relativos à manutenção preventiva condicionada.

Esta metodologia é mais eficaz quando existem indicadores precisos e confiáveis da "saúde" dos sistemas e seus sub-sistemas (Banks et al., 2008).

Na manutenção condicionada, a intervenção é realizada quando indicadores permitem inferir o funcionamento no limite das condições normais mas ainda não indutoras de avaria. A observação visual consiste no método mais antigo e ainda assim um dos mais poderosos e amplamente utilizados na monitorização da condição do equipamento. Posteriormente, a monitorização da condição passou a utilizar métodos automatizados, recorrendo a técnicas avançadas de processamento de sinal com base no reconhecimento de padrões, incluindo redes neuronais, lógica fuzzy, e modelação de dados empírica e física orientada (Lu et al., 2007). Assim que um equipamento começa a falhar, poderá emitir sinais capazes de serem percebidos pelos diferentes sentidos humanos, podendo desta forma perceber os percursos da falha. Felizmente, hoje em dia existem sensores capazes de substituir esses sentidos e assim identificar o aparecimento de falhas e degradações dos equipamentos. A integração destes sensores com as técnicas de manutenção condicionada podem evitar a substituição desnecessária de equipamentos e permitem reduzir os custos de manutenção e melhorar a segurança, a disponibilidade e a eficiência dos processos (Hashemian e Bean, 2011). A premissa comum da manutenção condicionada é que a monitorização regular da condição mecânica real, o rendimento operacional, e outros indicadores da condição operacional das máquinas e dos sistemas produtivos fornecem os dados necessários para assegurar o intervalo máximo entre reparações. Ela também minimiza o número e os custos de paragens não-programadas criadas pelas falhas das máquinas. Existem referências a este tipo de manutenção como "profética" uma vez que o que se pretende é adivinhar que uma falha poderá estar para acontecer antes de estar num estágio iminente.

Estudos recentes mostram que a manutenção preditiva consegue assegurar níveis elevados de desempenho e fiabilidade (Bansal et al., 2005b). Um programa de manutenção preditiva abrangente pode fornecer dados sobre a condição mecânica real de cada máquina e o rendimento

operacional de cada sistema produtivo. Estes dados habilitarão o gestor e equipas de manutenção a programar atividades de manutenção muito mais efetivas em termos de custo e operacionalidade. A manutenção preditiva para além de poder minimizar o número de quebras de todos os equipamentos da unidade fabril também pode assegurar que o equipamento reparado esteja em condições aceitáveis. Torna possível identificar problemas na máquina antes que se tornem sérios, já que a maioria dos problemas podem ser minimizados se forem detetados e reparados com antecedência.

2.2.4 Métodos de gestão da manutenção

A manutenção sempre procurou ir ao encontro dos interesses e prioridades da produção, evoluindo da simples manutenção correctiva para a manutenção preventiva. Finalmente, evoluiu para a manutenção preventiva condicionada, pela qual procurou juntar aos benefícios do planeamento e da redução de falhas inoportunas conseguidas com a manutenção preventiva, os benefícios da medição de parâmetros caracterizadores do grau de desgaste.

Novos métodos de apoio à gestão da manutenção têm sido empregues nos últimos anos. Os métodos Manutenção Produtiva Total ou TPM, Manutenção centrada na fiabilidade ou RCM e mais recentemente a Manutenção Baseada no Risco ou RBM, passaram a ser utilizadas com grandes perspectivas de retorno para as organizações (Cooke, 2003).

A seguir são apresentadas as principais características destas novas metodologias, que estão presentes nas estruturas da manutenção das empresas modernas, com diferentes graus de intensidade e diferentes aspetos de utilização (Zio e Compare, 2013).

2.2.4.1 TPM

O TPM consiste num método de apoio à gestão da manutenção que tem sido empregue com sucesso, a nível mundial, nas últimas três décadas (Jeon et al., 2011). Como resposta aos problemas relativos à manutenção encontrados em ambientes produtivos, Nakajima (1998) desenvolveu e introduziu o conceito de TPM. O TPM implica uma atitude de gestão da manutenção integrada com as restantes funções da empresa, nomeadamente com a produção e a qualidade, estabelecendo um triângulo coeso e, sempre que possível, consensual, que deverá ser o orientador das atividades da manutenção em prol de um interesse comum, que é o da organização (Chan et al., 2005).

Este método consiste numa abordagem de manutenção preventiva e produtiva fundamentado em atividades baseadas em equipas visando melhorar o desempenho do equipamento e simultaneamente evitando falhas nesses mesmos equipamentos (Swanson, 2001). Estas atividades baseadas em equipas têm por objetivo melhorar a eficiência dos equipamentos sendo fundamentadas no estabelecimento de comunicação entre os engenheiros e as equipas de manutenção acerca de potenciais problemas desses mesmos equipamentos (Adair-Heeley, 1989). Simultaneamente, é estabelecido um envolvimento dos operadores nas atividades de rotina removendo-se as fronteiras entre manutenção e a operação com o objetivo de atingir o aumento de disponibilidade dos equipamentos. Assim, o objetivo é alcançar zero falhas e zero avarias, fatores que permitem zero perdas de produção e ao mesmo tempo zero defeitos nos produtos.

Quando se aborda a avaliação do desempenho no âmbito do TPM, utiliza-se o indicador OEE (*Overall Equipment Effectiveness*) que é composto por três fatores, nomeadamente os fatores “Disponibilidade”, “Desempenho” e “Qualidade” já que o TPM visa maximizar a eficiência do equipamento e tem impacto nestes três fatores (Schippers, 2001).

2.2.4.2 O RCM

Segundo Fore e Msipha (2010), a história do RCM está bem documentada e é bem conhecida pela maioria das equipas e responsáveis de manutenção. A maioria tem plena consciência da metodologia RCM para o desenvolvimento de planos de manutenção, apesar de diferentes versões da mesma metodologia possuírem pequenas nuances nos seus processos padrão.

A metodologia RCM é utilizada de forma a determinar o que deve ser realizado de forma a assegurar que qualquer ativo físico continue a operar tal como o esperado num contexto operacional presente. Para ser desenvolvida, a metodologia implica sete questões acerca do ativo ou sistema em análise, para que assim seja preservada a função do sistema produtivo (Moubray, 1997):

1. Quais são as funções e padrões de desempenho do ativo no seu contexto de operação atual?
2. De que forma ele falha na execução da sua função?
3. O que causa cada falha funcional?
4. O que acontece quando ocorre cada falha?
5. Qual a importância de cada falha?
6. O que pode ser efetuado para vaticinar ou prevenir cada falha?

7. O que pode ser realizado se não for encontrada uma tarefa pró-ativa apropriada?

Segundo Fore e Msipha (2010), Moubray define o RCM de duas formas:

- i. Um processo utilizado para determinar os requisitos de manutenção de qualquer ativo físico no seu contexto operacional;
- ii. Um processo utilizado para determinar o que deve ser feito de forma a assegurar que qualquer ativo físico continue a fazer o que os seus utilizadores querem que faça, dentro do seu contexto operacional.

No entanto, a norma IEC-60300-3-11 (2009) surge com uma nova definição baseada nas definições de Moubray. Segundo esta norma, o RCM consiste numa forma sistemática de identificar tarefas eficazes e eficientes de manutenção preventivas, para itens de um sistema ou equipamento, e respetivos intervalos entre intervenções de acordo com um conjunto específico de procedimentos (Fore e Msipha, 2010).

Esta metodologia foi amplamente utilizada no setor aeronáutico durante muitos anos. Dentro de uma sistemática bastante regulamentada, a manutenção das aeronaves obedecia a um rígido calendário de tarefas de inspeção e revisões (Rausand, 1998).

A implementação do RCM tem como objetivo alcançar a fiabilidade e a segurança inerentes aos equipamentos, com o mínimo custo, identificando as tarefas de manutenção tecnicamente aplicáveis e adequadas para detetar e evitar as falhas das funções do equipamento, ou para reduzir as suas consequências. Se o método for bem aplicado, é capaz de reduzir em grande margem os trabalhos de manutenção corretiva (Campbell, 1999).

Fatores como o desgaste, corrosão, fadiga, fenómenos físico-químicos e acidentes que ocorrem nos componentes de qualquer equipamento alteram as suas condições normais. Esses fenómenos e eventos que ocorrem durante a sua utilização podem degradar essas condições o suficiente para que os componentes e equipamentos não mais apresentem o desempenho requerido, levando à ocorrência de falhas. Segundo Smith (1993), os benefícios ou vantagens da utilização de RCM bem como as desvantagens são várias, e têm impacto nas operações, segurança, logística, configuração e administração.

Vantagens:

1. Pode ser definido um programa de manutenção mais eficiente;
2. Reduz custos, eliminando a manutenção desnecessária de equipamentos;
3. Minimiza a frequência de revisões;
4. Reduz a probabilidade de falhas súbitas de equipamentos;

5. Foca-se em atividades de manutenção em componentes críticos do sistema;
6. Aumenta a fiabilidade dos equipamentos;
7. Analisa a raiz da causa da ocorrência.

Desvantagens:

1. Os custos de arranque associados à formação de pessoal e as necessidades de equipamentos podem ser significativos;
2. O seu potencial de poupança não é facilmente visto pela administração.

2.2.4.3 O RBM

O conceito de Manutenção Baseada no Risco foi desenvolvido com o objetivo de inspecionar os componentes de alto risco com maior frequência e rigor de forma a alcançar níveis de risco toleráveis. O RBM consiste numa metodologia para a tomada de decisão relativamente às tarefas de manutenção de forma a reduzir a probabilidade de falha de um equipamento e as consequências que daí advêm (Arunraj e Maiti, 2006).

Esta metodologia divide-se em duas fases:

- i. A avaliação do risco;
- ii. O planeamento da manutenção baseada no risco.

O seu objetivo principal consiste em reduzir o risco global que pode resultar como consequência do surgimento de avarias inesperadas nos equipamentos. As atividades de inspeção e manutenção são tornadas prioritárias, tendo como base a quantificação do risco decorrente de uma falha de um equipamento ou componente, de modo a minimizar o risco global (Khan e Haddara, 2004).

2.2.5 Modelos de apoio à decisão na manutenção

Segundo Lopes (2007), encontram-se na literatura vários modelos de manutenção preventiva; uns consideram a substituição preventiva do item, outros consideram, para além da substituição, a realização de revisões. São estudadas várias políticas de manutenção, sendo estas: substituição após um número especificado de revisões, substituição ou revisão em intervalos constantes, etc.

A minimização do custo total de manutenção é a premissa da maioria dos modelos. No entanto, existem alguns que consideram a minimização do tempo de paragem do equipamento,

maximizando desta forma a disponibilidade, o que posteriormente se traduz nos custos de operação. Outras medidas, para além das anteriores, apenas se encontram pontualmente.

Na literatura existem diversos modelos de manutenção preventiva que propõem diferentes políticas de substituição. Utilizando a classificação elaborada por Ascher e Feingold (1984) existem (Lopes, 2007):

- políticas que têm como premissa que a substituição dos itens é realizada depois de concluído um intervalo de tempo fixo de operação (T) do sistema sem avarias, ou então aquando da ocorrência de uma avaria (Figura 2-8). Ascher e Feingold (1984) caracterizam-nas como política ou modelo do tipo 1 e destacam os trabalhos de Barlow e Hunter (1960), de Barlow e Proschan (1965), de Makabe e Morimura (1963), de Glasser (1967) e de Fox (1966). O modelo de tipo 1 entende que os tempos entre avarias são independentes e igualmente distribuídos ou que a função de risco dos itens é crescente.

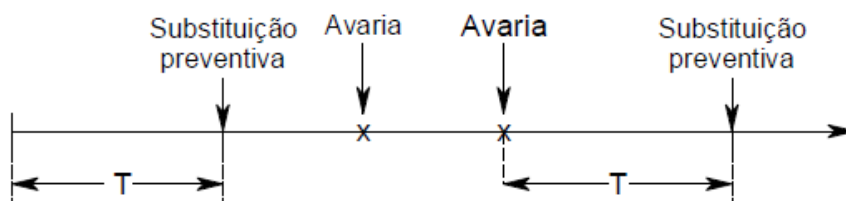


Figura 2-8 - Substituição em intervalos de operação constantes. Adaptado de Lopes (2007).

- políticas que consideram a substituição em intervalos de tempo de operação constantes, não considerando o número de falhas ocorridas (Figura 2-9). Estas são as políticas que Ascher e Feingold (1984) chamaram de políticas do tipo 2. Ascher e Feingold (1984) realçam os trabalhos de Barlow e Hunter (1960), Makabe e Morimura (1963) e de Sivazlian (1973).

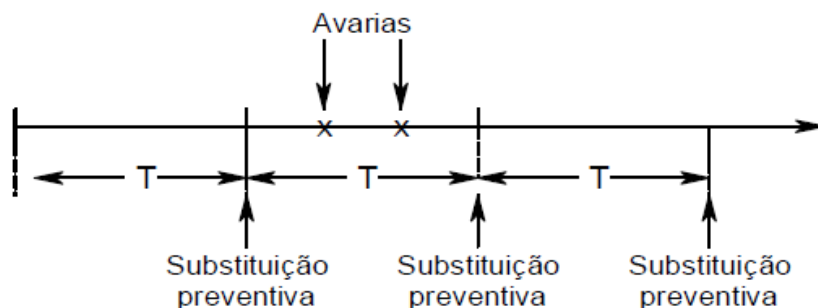


Figura 2-9 - Substituição em intervalos de tempo constantes. Adaptado de Lopes (2007).

- para outras políticas, a substituição ocorre após um número pré-especificado de falhas. Depois da ocorrência de cada falha, o sistema sofre uma reparação mínima e na próxima falha, o sistema é substituído. Ascher e Feingold (1984) designam-nas de políticas do tipo 3 e citam os trabalhos de Makabe e Morimura (1963) e Park (1979).
- existem outras políticas que introduzem algumas alterações às políticas anteriores e ainda outras que são combinações dessas políticas (as políticas do tipo 1 modificadas de Schaeffer (1971), Cléroux e Hanscom (1974), Ran e Rosenlund (1976), Cléroux et al. (1979) e de Nakagawa (1980), as políticas do tipo 4 de Makabe e Morimura (1963), do tipo 2 e 3 de Morimura (1970) e Muth (1977), apresentadas por Ascher e Feingold (1984)).

As políticas até aqui apresentadas, consistem em políticas que consideram um intervalo de tempo para a substituição planeada dos sistemas. Ascher e Feingold (1984) apresentam outra classe de políticas de substituição que se fundamentam na limitação dos custos de reparação – políticas estudadas por Hastings (1969) e por Drinkwater e Hastings (1967). Se o custo de reparação exceder o custo limite imposto, procede-se à substituição do equipamento e se o custo de reparação for inferior ao custo limite, procede-se à reparação mínima do equipamento.

2.2.6 Softwares de manutenção

A gestão da manutenção deve ser entendida como uma ferramenta estratégica, essencial para assegurar uma elevada disponibilidade de uma unidade fabril, que envolve o desenvolvimento e execução de um planeamento referentes às tarefas de manutenção. Toda a organização deve contribuir no sentido de ajudar o departamento de manutenção a melhorar a *performance* dos equipamentos (Wilder e Cannon, 1993). A função do departamento de manutenção inclui a realização de todas as tarefas associadas à monitorização de equipamentos, a realização de relatórios de trabalho, a aquisição de materiais e equipamentos, a contratação de recursos humanos, entre outras (Huo e Zhang, 2003).

Os sistemas informáticos de gestão de manutenção (*Computerized Maintenance Management Systems* – CMMS) evoluíram do elementar controlo de ativos e ações de manutenção preventiva para sistemas de informação de manutenção. A manutenção industrial é uma das componentes do processo produtivo de uma organização que gera e usa diariamente *gigabytes* de informação técnica e operacional. Um CMMS tem como principal objetivo alcançar um equilíbrio adequado entre a disponibilidade das instalações, a eficiência dos processos e a otimização do funcionamento dos equipamentos no seu processo produtivo. A utilização de um sistema

informático de gestão de manutenção surge como uma forma ideal de atingir este objetivo (Sahoo e Liyanage, 2008).

Atualmente os CMMS estão a tornar-se cada vez mais importantes nas organizações. Segundo Kans (2008), as tecnologias de informação consistem numa ferramenta importante de forma a alcançar a eficiência e eficácia dos processos de manutenção, quando corretamente aplicados. Estes sistemas residem em ferramentas de forma a gerir todo o sistema de manutenção podendo mesmo ser integrados com outros sistemas existentes, onde toda a informação é agregada, organizada e tratada estando sempre acessível a todos, permitindo assim planear e acompanhar todas as atividades de manutenção.

Com este tipo de sistemas torna-se possível diminuir o tempo de resposta das equipas de manutenção, reduzir o tempo de inatividade de um dado equipamento, diminuindo assim a sua frequência de falhas (Carnero e Novés, 2006).

Segundo Huo e Zhang (2003), as vantagens de um CMMS são óbvias e importantes, podendo ser sumarizadas da seguinte forma. Primeiro, a capacidade de armazenamento e procura de informação num sistema deste tipo reduz de forma drástica o tempo despendido em busca de informação relevante. Em segundo, a capacidade de resposta de um CMMS traduz-se em melhores *outputs* e melhor capacidade de análise dos mesmos. Por fim e não menos importante, estes tipos de sistemas permitem a geração automática de trabalhos de rotina, bem como a gestão da disponibilidade de peças de reposição.

Desta forma, a implementação de um CMMS por parte de uma organização traduz-se numa gestão da manutenção de maior qualidade, num melhor suporte à tomada de decisão e conseqüentemente no aumento de eficiência (Wilder e Cannon, 1993).

Um CMMS deve albergar um registo de intervenções quer correctivas, quer planeadas. Estas intervenções quando finalizadas deverão dar entrada no sistema de forma a gerar um historial de avarias e intervenções referentes a um determinado equipamento. Deve de igual forma gerir todas as tarefas de manutenção preventiva (Blackmore, 2001). Frequentemente muitos equipamentos não são sujeitos a ações de manutenção preventiva, desta forma e através da introdução no sistema de orientações dos fabricantes e conhecimento adquirido, algumas tarefas específicas de manutenção poderão ser efetuadas baseando-se em horas de operação. O sistema poderá produzir alertas acerca da altura adequada para se efetuar uma intervenção sobre um determinado equipamento.

Assim sendo, a utilização de um CMMS tornar-se uma mais-valia na implementação de estratégias de manutenção por parte de uma organização.

2.3 Modelos distribuídos de aquisição de dados

A evolução na área da Inteligência Artificial (IA), conjugada com a expansão das redes informáticas, possibilitou a difusão e exploração do processamento distribuído. Na área da IA surgem trabalhos sobre conjuntos de processos com capacidade de processamento, comunicando entre eles, com potencial de aplicação na resolução de problemas. É o início da tecnologia de agentes.

Segundo Schmidt (2003), as primeiras pesquisas na área da Inteligência Artificial exploravam processos computacionais centralizados, compostos por um único centro de controlo, um foco de atenção e uma base de conhecimento (base de dados). Isto devia-se às técnicas e tecnologias disponíveis na época, tanto no respeitante ao *hardware* quanto às próprias técnicas de IA. Esta característica começa a mudar com a possibilidade de processamento paralelo e/ou distribuído conduzido pela evolução do *hardware*, dos sistemas operativos e redes e ainda, pelo aparecimento na área da IA de técnicas capazes de realizar o processamento distribuído e paralelo (ex.: Redes Neurais), assim como o controlo distribuído (ex.: Sistema *Blackboard*).

Diversos trabalhos começam a ser desenvolvidos tirando partido das características dos sistemas distribuídos para a resolução de potenciais problemas. Esta nova área de aplicação de IA designa-se de Inteligência Artificial Distribuída. A Inteligência Artificial Distribuída (IAD) preocupa-se principalmente com a criação de agentes inteligentes autónomos e com a comunicação de conhecimentos entre sistemas inteligentes (Schmidt, 2003).

Esta secção aborda os conceitos de modelo cliente-servidor, agente e sistema multiagente. Relativamente ao conceito de agente, foca-se principalmente as suas características e a distinção existente entre os tipos de agentes. Ainda nesta secção é abordada a importância da implementação de um sistema multiagente.

2.3.1 Modelo genérico

Atualmente o computador consiste na plataforma mais utilizada para efetuar a aquisição, processamento e tratamento de dados bem como o controlo de sistemas. Entre as principais

razões para a sua popularidade pode destacar-se o baixo custo, a flexibilidade, a facilidade de utilização e o seu desempenho (Serrano et al., 2004).

Os sistemas informáticos permitem a visualização e armazenamento de dados e o controlo de diversas grandezas relevantes em equipamentos tais como a pressão, o caudal, a temperatura, etc.

Hoje em dia, através do recurso a diversos *softwares* é possível criar ambientes de trabalho bastante atrativos em diversos tipos de aplicações, tais como: aplicações laboratoriais de aquisição de dados e controlo de processos, controlo de processos industriais, monitorização de condições ambientais, medições em geral, etc. Assim, é possível construir sistemas de aquisição de dados adaptados especificamente às necessidades, permitindo obter um sistema personalizado.

2.3.2 Modelo cliente-servidor

O modelo cliente-servidor é um modelo computacional que diferencia clientes e servidores, sendo interligados entre si geralmente através de uma rede de computadores. Cada instância de um cliente pode enviar pedidos de dados para algum dos servidores conectados e esperar pela resposta. Por sua vez, algum dos servidores disponíveis pode aceitar tais pedidos, processá-los e devolver o resultado ao cliente. Apesar do conceito ser aplicado em diversas campos de aplicação, a arquitetura é praticamente a mesma.

Muitas vezes os clientes e servidores comunicam através de uma rede de computadores constituída por diferentes sistemas informáticos (*hardware*), no entanto o cliente e o servidor podem por vezes residir no mesmo sistema informático. A máquina servidor é um *host*¹ que executa um ou mais programas partilhando os seus recursos com os clientes.

Um cliente não partilha os seus recursos, mas solicita um conteúdo ou um serviço a um servidor.

2.3.3 Agentes

Devido à multiplicidade de papéis que um agente pode desempenhar, é muito difícil formular em poucas palavras uma definição para agentes inteligentes. Podem ser definidas como entidades autónomas, dotadas de uma base de conhecimento e capazes de interagir com o meio em que

¹ Consiste numa qualquer máquina ou computador conectado a uma rede, podendo oferecer informações, recursos, serviços e aplicações aos diferentes utilizadores ou a outros nós existentes na mesma rede.

estão inseridos, tomando assim, decisões que irão auxiliar ou até mesmo substituir o trabalho humano (Luck et al., 2003).

Um agente é visto como uma entidade cognitiva, ativa e autónoma, ou seja, que possui um sistema interno de tomada de decisão, agindo sobre o mundo e sobre os outros agentes que o rodeiam, e, por fim, capaz de funcionar sem necessitar de algo ou de alguém para o guiar (tem mecanismos próprios de percepção do exterior) (Luck et al., 1997).

Para clarificar o que é um agente inteligente é conveniente apresentar os elementos que o caracterizam, podendo ser revelada na sua forma mais genérica por uma entidade baseada em *hardware* ou, mais frequentemente, em *software* (Bradshaw, 1997):

- Autonomia: funcionam sem a intervenção direta de operadores e possuem algum tipo de controlo sobre as suas ações e sobre o seu estado interno;
- Aptidão Social: interagem com outros agentes e, possivelmente, com seres humanos através de algum tipo de linguagem de comunicação;
- Reação: percebem o seu ambiente e respondem aos estímulos dele recebidos;
- Iniciativa: não reagem somente ao seu ambiente, mas também devem exibir um comportamento orientado à satisfação dos seus objetivos;
- Continuidade temporal: são processos em execução contínua, que tanto podem estar ativos, em *foreground*, como “adormecidos”, em *background*;
- Orientação a objetivos: devem ser capazes de lidar com tarefas complexas de alto nível. A decisão relativa à segmentação de uma tarefa em subtarefas menores e a ordem ou forma de execução das subtarefas deve ser tomada pelo próprio agente;
- Mobilidade: é a habilidade que um agente possui de movimentar-se numa rede, ocupando diferentes nodos e recursos ao longo do tempo;
- Benevolência: é a ideia de que o agente não possui objetivos conflituosos e que cada agente irá sempre tentar fazer o que lhe for pedido;
- Racionalidade: é a hipótese de que os agentes irão agir de forma a atingir os seus objetivos e não afastar-se deles;
- Adaptabilidade: deve ser capaz de adaptar-se aos hábitos, métodos de trabalho e preferências dos seus utilizadores;
- Colaboração: não deve aceitar (e executar) instruções não pensadas, mas deve levar em conta que os seres humanos cometem erros. Assim, deve também ser capaz de recusar ordens

que, por exemplo, produziriam uma sobrecarga inaceitável na rede ou que ocasionariam danos a outros utilizadores.

Estas características dão origem a diferentes tipos de agentes, conforme se indica na subsecção seguinte.

2.3.3.1 Tipos de agentes

Os domínios de aplicação de agentes incluem a filtragem de informação, o trabalho colaborativo, a gestão de agenda, o uso de interfaces, o comércio electrónico e a gestão de redes. A classificação do tipo de agentes depende essencialmente das suas características e do tipo de ambiente sob o qual vão operar. Gilbert et al. (1995), definiu um método de classificação de agentes baseado na inteligência, ação e mobilidade (Figura 2-10). Através desta classificação é possível verificar que os agentes inteligentes possuem um nível de ação que pode incluir interação com serviços, aplicações, dados ou utilizadores. Possuem simultaneamente inteligência ao nível das capacidades de raciocínio, preferências, planeamento e aprendizagem.

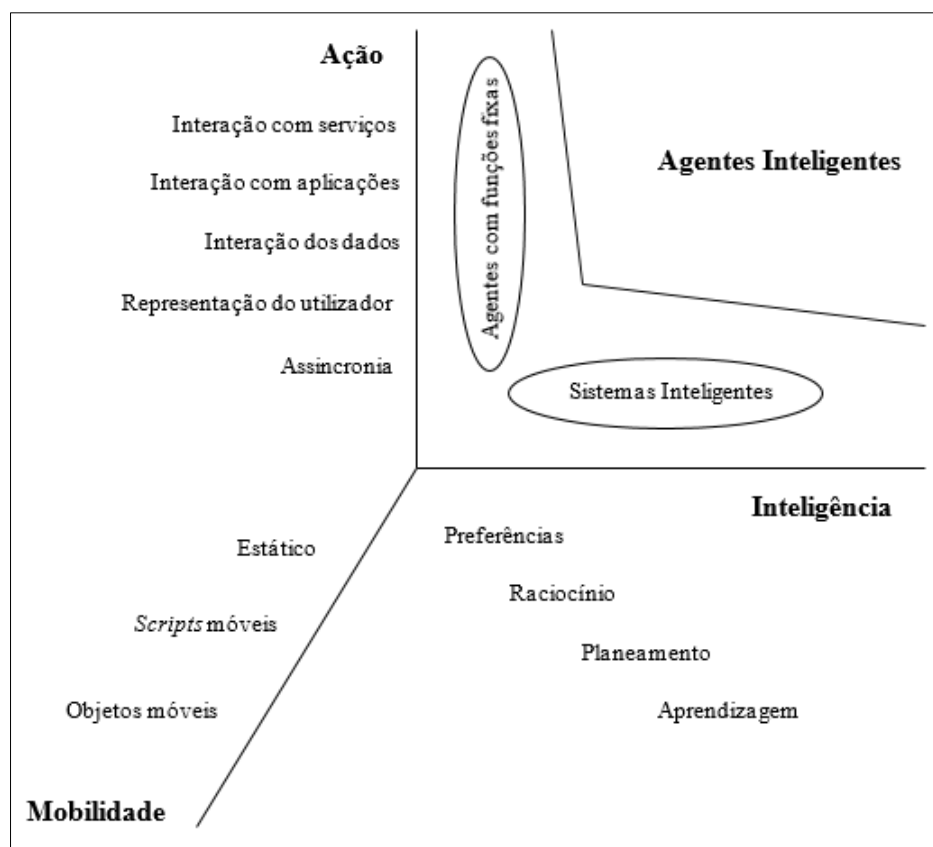


Figura 2-10 - Âmbito dos agentes inteligentes. Adaptado de Gilbert et al. (1995).

Nwana (1996) classificou os agentes de forma mais abrangente e de acordo com três atributos fundamentais: autonomia, cooperação e aprendizagem. A autonomia refere-se à aptidão de o agente conseguir agir sem intervenção humana. A cooperação refere-se à capacidade de interagir com outros agentes ou humanos. A aprendizagem diz respeito à aptidão do agente poder melhorar o seu desempenho adquirindo conhecimento.

2.3.3.2 Sistema multiagente

Os SMA formam uma subárea da Inteligência Artificial distribuída e concentram-se no estudo de agentes autônomos num universo multiagente. Para os SMAs, o termo autônomo designa o facto de que os agentes têm uma existência própria, independente da existência de outros agentes. Normalmente, cada agente possui um conjunto de capacidades comportamentais que definem a sua competência, um conjunto de objetivos, e a autonomia necessária para utilizar as suas capacidades comportamentais a fim de alcançar os seus objetivos.

Os SMAs dividem-se essencialmente em dois tipos: os SMA cognitivos e os reativos (Alves, 2007).

Os SMA cognitivos são baseados em organizações humanas como grupos, hierarquias e mercados. Segundo Ferber e Gasser (1991), as principais características dos SMA Cognitivos são as seguintes:

- representação explícita do ambiente e dos outros agentes da sociedade;
- capacidade de manter um histórico das interações e ações passadas e serem capazes de planear ações futuras;
- o seu sistema de compreensão, que permite examinar o ambiente, e o de comunicação, que permite a troca de mensagens entre agentes, são distintos. A comunicação entre agentes é realizada de modo direto, através do envio e receção de mensagens;
- o seu mecanismo de controle é deliberativo. Os agentes cognitivos raciocinam e decidem em conjunto sobre quais as ações que deverão ser executadas, que planos devem ser seguidos e que objetivos devem ser alcançados;
- considerando a última proposição, os modelos de organização dos SMAs cognitivos são modelos sociológicos, como as organizações humanas;
- um SMA cognitivo contém, normalmente, poucos agentes, na ordem de algumas dezenas, no máximo.

Os SMA reativos são baseados em modelos de organização biológica ou etológica (formigas, abelhas, etc.). O modelo de funcionamento de um agente reativo é constituído pelo par Estimulo-Resposta (Ação-Reação). As principais características dos agentes e dos SMA reativos são (Ferber e Gasser, 1991):

- não existe representação explícita do conhecimento: o conhecimento dos agentes é implícito (as suas regras de comportamento) e a sua manifestação realiza-se através do seu comportamento e do comportamento dos restantes agentes;
- não existe representação do ambiente: o comportamento (resposta) de cada agente é baseado no que ele compreende (estímulo) a cada instante. Não há uma representação interna explícita do ambiente;
- não existe memória das suas ações: os agentes reativos não mantêm nenhum tipo de histórico de suas ações, ou seja, o resultado de uma determinada ação passada não influencia directamente a sua decisão numa ação futura;
- organização etológica: a forma de organização dos SMAs reativos é similar à observada em animais que vivem em grandes comunidades;
- grande número de membros: em geral, os SMA reativos possuem um grande número de agentes, com populações que podem chegar à ordem de milhares de membros.

Em suma, os agentes reativos são muito simples e não possuem representação do seu ambiente. As suas reações dependem unicamente de sua percepção do ambiente em que está inserido.

Desenvolver um sistema multiagente consiste num processo bastante complexo. De forma a diminuir o tempo de desenvolvimento existem diversas plataformas (*frameworks*) para o seu desenvolvimento, de entre as quais se salientam a seguir as principais (Alves, 2007):

- JATLite - Foi desenvolvida na Universidade de Stanford e consiste num pacote de programas escrito na linguagem *Java*² que é capaz de criar novos agentes de software com a capacidade de comunicar através da Internet. As comunicações de baixo nível são baseadas em TCP/IP e a comunicação entre agentes é baseada em mensagens usando a linguagem KQML (*Knowledge Query and Manipulation Language*) (Jeon et al., 2000);

² Linguagem concebida para ser uma linguagem *script* com orientação a objetos. Consiste numa linguagem de *script cliente-side*, orientada a objetos e tem como principal aplicação as páginas *web*.

- Jade - A plataforma Jade (*Java Agent Development Framework*), completamente programada em linguagem *Java*, simplifica a implementação de um sistema multiagente. Foi desenhada de forma a suportar escalabilidade (Bellifemine et al., 2007).
- Jason - A plataforma Jason (*Java-based interpreter for an extended version of AgentSpeak*) consiste numa plataforma *open source* que permite o desenvolvimento de SMA com a possibilidade do utilizador final personalizar muitas das suas funcionalidades (Bordini et al., 2007).

A utilização de uma plataforma para o desenvolvimento de SMA permite uma abstracção da própria arquitectura multiagente, tornando mais célere o desenvolvimento de sistemas.

2.4 A descoberta de conhecimento em bases de dados

Seguidamente são apresentados os objetivos, conceitos, abordagens, metodologias e especificações relacionadas com a temática da DCBD, bem como a apresentação de algumas abordagens de referência na aplicação desta temática na área da manutenção industrial.

2.4.1 Data Mining

O avanço da tecnologia tem proporcionado novas formas de criar e armazenar dados (Santos e Azevedo, 2005). Hoje em dia, a quantidade de dados armazenados excede a capacidade de análise desses mesmos dados sem o uso de técnicas de análise automatizadas. Como consequência deste aumento efetivo de informação, o processamento através de métodos tradicionais tornou-se mais difícil e complexo (Goebel e Gruenwald, 1999). As ferramentas convencionais de análises de dados têm capacidades limitadas em detetar padrões e descobrir o conhecimento inserido nos dados, pois só utilizam métodos estatísticos (Santos e Azevedo, 2005). A extração de conhecimento nas bases de dados é um campo que está em evolução de forma a fornecer soluções de análise automatizadas.

A necessidade de desenvolver um processo automático ou semiautomático para a extração de conhecimento a partir de bases de dados consiste numa resposta a cinco grandes desafios, referidos por Tan et al. (2006), que motivaram o desenvolvimento do DM em detrimento da utilização das metodologias de análise de dados tradicionais:

i) Dimensionalidade: o crescimento exponencial da capacidade de processamento e armazenagem dos sistemas informáticos e consequente decréscimo dos custos de geração, recolha e armazenamento de informação resultam em novas e multifacetadas estruturas de bases de dados com uma vastidão de atributos associáveis a cada objeto de estudo;

ii) Escalabilidade: as técnicas e algoritmos desenvolvidos devem ser altamente escaláveis, ou seja, devem funcionar com grandes volumes de dados (*terabytes* de informação). Como norma, as técnicas de análise de dados tradicionais podem funcionar bem com pequenas amostras mas não ser suficientemente escalável para tratar bases de dados com maior dimensão;

iii) Heterogeneidade e complexidade: as bases de dados armazenam novas tipologias de atributos diferentes dos atributos tradicionais (categóricos e numéricos) comumente utilizados em análise de dados. Por outro lado, têm surgido novas e complexas fontes de dados provenientes de páginas *web*, redes sociais, fluxo constante de dados (*data streams*), etc;

iv) Distribuição da informação: a circulação da informação é um aspeto fulcral na capacidade de aceder, armazenar e disponibilizar informação. Torna-se essencial responder a desafios como: reduzir a quantidade de informação, consolidar os resultados da análise de dados provenientes de múltiplas fontes, e garantir a segurança da informação do ponto de vista da sua qualidade e também da privacidade;

v) Paradigma de análise: o princípio tradicional das técnicas de análise de dados baseia-se na utilização de testes de hipótese, a partir do qual se desenha um dispositivo experimental que permite recolher os dados necessários para o teste e posterior análise da hipótese colocada. A dimensionalidade, heterogeneidade e complexidade da informação disponível torna extremamente complexo este processo, que obrigaria, por exemplo, à colocação de centenas de hipóteses.

Tornou-se necessário uma nova geração de técnicas e ferramentas computacionais de forma a assistir o ser humano na extracção de informação útil, isto é, conhecimento. O DM surge da convergência de várias disciplinas, nomeadamente, a estatística clássica, a IA e a aprendizagem automática, entre outras, para responder a todos estes desafios colocados pela rápida evolução dos recursos e tecnologias computacionais verificados nas últimas décadas.

Por tudo isto, emergiu a área da DCBD (*Knowledge Discovery in Databases – KDD*), recorrendo a modelos e técnicas de DM para extrair conhecimento útil, padrões e tendências, previamente desconhecidas, de forma autónoma e semi-automática (Apte et al., 2002).

O processo de descoberta de conhecimento em bases de dados é iterativo e interativo, envolvendo diversas etapas, como é demonstrado na Figura 2-11.

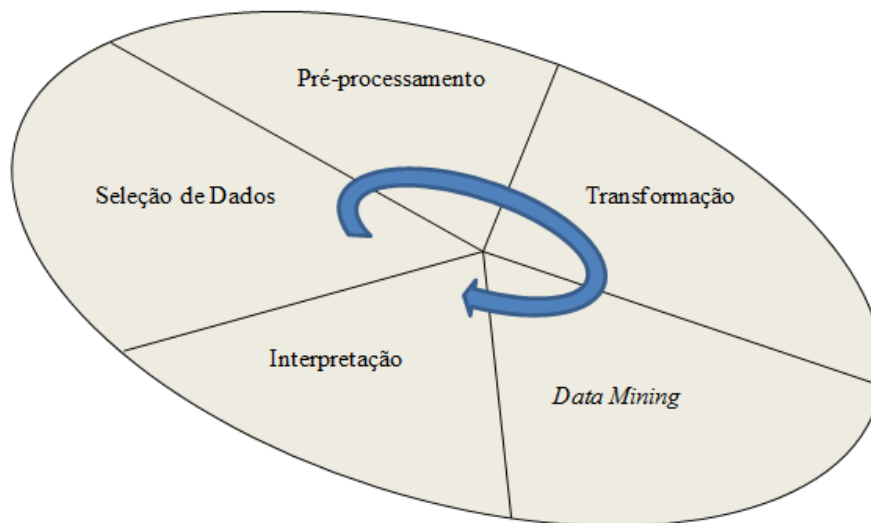


Figura 2-11 - Fases do processo de DCBD.

Desta forma, pode-se considerar os seguintes passos no processo de extração de conhecimento (Fayyad et al., 1996a):

- i) Seleção de dados, consiste na recolha e criação de um conjunto de dados que servirão de base de estudo. Com o domínio de estudo especificado, deve-se localizar e escolher quais as fontes de dados que servirão de base ao estudo a efetuar, para que um conjunto de dados apropriado possa ser selecionado e retirado das mesmas. As fontes de dados utilizadas para realizar o processo de DCBD podem ser classificadas como internas e externas. As fontes internas normalmente são repositórios de dados que já estão incorporados no sistema de aplicação do domínio em questão, e normalmente, são constituídas por *data warehouses*³, bases de dados operacionais e dados históricos. As fontes externas são compostas por todos os outros tipos de dados que habitualmente não estão incorporados no sistema de aplicação. Os dados existentes em fontes internas normalmente já estão armazenados em algum tipo de repositório estruturado (i.e. bases de dados relacionais) e por isso, requerem menos processamento, já que eles podem ser recuperados utilizando-se, por exemplo, consultas SQL⁴. No entanto, os dados contidos em fontes externas necessitam de

³ Ou base de dados. Consiste num sistema de computação utilizado para armazenar informações relativas às atividades de uma organização em bases de dados.

⁴ *Structured Query Language*, linguagem de pesquisa declarativa padrão de uma base de dados relacional.

formatação e seguidamente serem inseridos no sistema para que possam ser recuperados e preparados para análise. As fontes de dados podem possuir quantidades na ordem dos *gigabytes* ou *terabytes* e, além disso, estes dados podem ser provenientes de distintas fontes, necessitando que estes sejam reduzidos e colocados num conjunto de dados com estrutura homogénea (i.e. tabelas) de forma a permitir a sua análise conjunta. A realização de amostragem é crítica dentro do processo de DCBD, pois se não forem utilizadas técnicas estatísticas adequadas, pode-se gerar um subconjunto de dados não representativo, resultando análises que não demonstram a verdadeira situação dos factos registados na base de dados ou distorcendo os possíveis resultados;

- ii) Pré-processamento dos dados: decide-se as estratégias e realizam-se as devidas limpezas dos dados, a fim de remover ruídos e tratar possíveis inconsistências. Seguidamente, é efetuada uma análise dos dados armazenados, verificando a sua topologia e a identificação de valores ausentes e/ou desconhecidos;
- iii) Transformação dos dados: como as técnicas (algoritmos) de DM poderão não conseguir aceder e processar os dados no seu formato original, devido à forma de armazenamento ou normalização adotada na modelação da base de dados de suporte ao estudo, torna-se necessária a conversão desses mesmos dados para um formato mais apropriado, sendo possível ainda a sua sumarização, a fim de reduzir o número de variáveis consideradas ou criar novos atributos que possam ser agregadores de valor à base de dados;
- iv) *Data mining*: etapa onde é seleccionada e aplicada uma determinada técnica de acordo com o objetivo do processo de DCBD, isto é, classificação, previsão, descrição, segmentação ou visualização. Nesta fase é escolhida a técnica (algoritmo) de DM e sua metodologia de implementação na busca de novos padrões de conhecimento. Esta escolha dependerá do objetivo do DM, que poderá passar pela criação de modelos de predição ou apenas a obtenção de uma melhor compreensão da base de dados. Posteriormente, são aplicadas as técnicas seleccionadas sobre os dados analisados com o objetivo de localizar os padrões desejados. A qualidade dos resultados desta etapa depende diretamente da correta execução das etapas anteriores. Esta é a fase em que através da utilização de algoritmos de DM, os dados previamente seleccionados, pré-processados e transformados são utilizados pelo algoritmo de DM

escolhido com base no objectivo do processo de DCBD: regressão, classificação, previsão, etc. As tarefas e os algoritmos de DM são objecto de uma descrição mais detalhada nas subsecções 2.4.3 e 2.4.4;

- v) Interpretação dos resultados: os novos padrões de conhecimento encontrados devem ser validados a partir da interpretação e avaliação destes. Para interpretação dos padrões extraídos, é possível lançar mão a ferramentas estatísticas e de visualização que permitam efetuar uma “leitura” precisa dos resultados obtidos, de forma a possibilitar a verificação da validade e novidade, ou mesmo, a relevância dos padrões encontrados. Caso o conhecimento não seja validado, então, pode-se retornar às etapas anteriores e tentar refazê-las ou melhorá-las. Esta iteração pode ocorrer até que se obtenha resultados aceitáveis ou concluir-se que não seja possível extrair conhecimento relevante dos dados em análise.

Desta forma, a DCBD, tecnicamente, deslumbra-se como um desafio de carácter significativo, bem como difícil. De forma global, o processo de DCBD é considerado “a extracção não trivial de informação implícita, previamente desconhecida e potencialmente útil, a partir dos dados”(Fayyad et al., 1996a). Assim, faz-se uma distinção clara entre a extração dos dados e a descoberta de conhecimento. A descoberta de conhecimento examina os resultados tal como provêm da análise dos dados (processo de extrair tendências ou modelos dos dados), e após este processo, de forma cuidada e com precisão, transforma-os em informação útil e compreensível. Esta informação não é tipicamente recuperável pelas técnicas normais, mas sim através do uso de técnicas de IA (Fayyad e Piatetsky-Shapiro, 1996b).

O processo de DCBD depende de uma nova geração de ferramentas e técnicas de análise de dados, que envolve diversas etapas. A principal, que forma o núcleo do processo e que muitas vezes se confunde com ele, tal como referido anteriormente, chama-se DM.

O DM e a DCBD tornaram-se proeminentes campos de pesquisa em constante e rápido crescimento. Muitas técnicas e algoritmos têm sido desenvolvidos e utilizados em DM. Se os dados armazenados estão disponibilizados em diferentes formatos e/ou tabelas, têm que ser integrados numa etapa de preparação, pré-etapa de DM. Esta etapa de preparação dos dados não é automatizada através de nenhuma ferramenta de DM.

O DM, por outro lado, através do uso de algoritmos específicos ou de mecanismos de pesquisa, tenta descobrir padrões discerníveis e tendências nos dados, inferindo regras para os mesmos (Han e Kamber, 2001).

Quando, ao longo do processo de DCBD, ocorrem falhas originadas por decisões que se revelam inapropriadas, estas vão refletir-se na validade e utilidade dos modelos obtidos, não satisfazendo os objetivos delineados, ou retratando apenas o comportamento dos dados analisados, não podendo ser aplicados a dados desconhecidos por uma dificuldade de generalização. Quando se verificam estes casos, retrocede-se no processo, de modo a alterar as decisões tomadas, sendo o processo posteriormente retomado, permitindo identificar novos modelos que resultam das alterações efetuadas, verificando o grau de validade e utilidade dos mesmos e analisando o incremento ou decréscimo relativamente aos modelos anteriormente gerados (Teixeira, 2005).

Apesar dos algoritmos disponíveis possuírem critérios objetivos de avaliação da qualidade das regras, a introdução de medidas de interesse subjetivas tem como propósito limitar o conjunto de resultados a apresentar ao utilizador. A definição de medidas de interesse subjetivas, e que dependem de utilizador para utilizador, tendem a aumentar o grau de envolvimento do utilizador no processo de descoberta de conhecimento, tendo como contrapartida o aumento do interesse das diversas regras encontradas. Duas medidas de interesse subjetivas são o grau de surpresa, salientando que um padrão é interessante se ele é inesperado pelo utilizador, e a utilidade do padrão, sendo este interessante se o utilizador ou a organização poder usufruir do mesmo em seu benefício (Teixeira, 2005).

2.4.2 Metodologias e especificações

O processo de DCBD torna-se mais fácil de compreender, implementar e desenvolver se for enquadrado no contexto de uma metodologia (Groth, 2000).

As metodologias mais conhecidas são a CRISP-DM e SEMMA assim como a especificação PMML (Santos e Azevedo, 2005). Estas metodologias serão descritas nas subsecções seguintes.

2.4.2.1 Metodologia CRISP-DM

Esta metodologia foi concebida em finais de 1996, sendo os seus fundamentos construídos com base na teoria, nos princípios académicos, mas também baseados na prática, na experiência daqueles que desenvolvem projetos de DM (Santos e Azevedo, 2005). Foi incorporado conhecimento prático, utilizado na indústria, de forma a dar resposta aos requisitos dos

utilizadores, não se centrando unicamente na tecnologia, mas também nas necessidades dos utilizadores para resolver problemas de negócio (Han e Kamber, 2001).

A metodologia CRISP-DM é descrita como um processo hierárquico, com um ciclo de vida desenvolvido em seis fases: Estudo do negócio, Exploração dos dados, Preparação dos dados, Modelação, Avaliação e Aplicação. Esta metodologia é dinâmica, não tendo uma sequência fixa entre todas as fases, dependendo estas do desempenho das fases anteriores ou das tarefas particulares de determinada fase (Chapman et al., 2000).

A Figura 2-12 apresenta o ciclo de vida da metodologia CRISP-DM. Cada fase do ciclo é composta por diversas tarefas, descritas de forma hierárquica por níveis de abstração, que permitem tornar o processo perceptível a investigadores com *backgrounds* distintos.

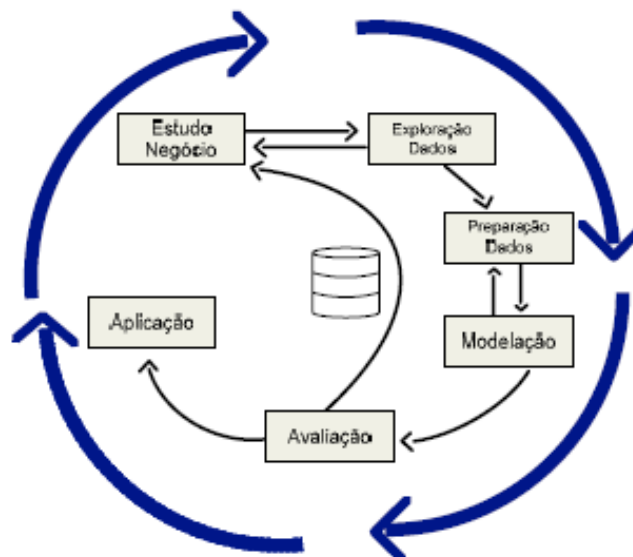


Figura 2-12 - Ciclo de vida da metodologia CRISP-DM. Adaptado de Chapman et al. (2000).

As fases da metodologia apresentadas na Figura 2-12 referem-se a (Chapman et al., 2000):

- Estudo do negócio: envolve a definição de objetivos e requisitos de um projeto e da necessidade de implementar um processo de DM dentro de uma determinada organização. Baseado no conhecimento adquirido, esta fase abrange ainda o desenho do tipo de tarefas a implementar;
- Exploração dos dados: a recolha dos dados envolve não só os mecanismos de acesso como uma avaliação descritiva das suas características e dos níveis de qualidade (despistando os problemas associados a erros e omissões, por exemplo). Assim é possível estabelecer um primeiro conjunto de hipóteses bem como detetar conjuntos de variáveis com especial interesse para o tema em estudo;

- c) Preparação dos dados: inclui a execução de todas as tarefas e respetivas ferramentas que permitem construir um conjunto de dados que respondam aos requisitos da fase seguinte, de modelação. Incluem-se tarefas como a limpeza, a transformação dos dados e a seleção (de subconjuntos de variáveis, de variáveis individuais e de objetos). O processo é repetido múltiplas vezes e integrado de forma iterativa com a fase de modelação;
- d) Modelação: um modelo, em DM, consiste numa representação computadorizada de observações do mundo real. Os modelos consistem na aplicação de algoritmos de forma a procurar, identificar e exibir quaisquer padrões relativamente aos dados. Existem dois tipos básicos ou tipos de modelos de DM: aqueles que classificam e aqueles que preveem. Para um dado problema ou hipótese são seleccionadas diversas ferramentas e modelos, das variadas opções existentes. Embora dedicadas à resolução do mesmo tipo de problemas, cada técnica encerra em si requisitos próprios que obrigam, por um lado, o investigador a otimizar os seus parâmetros de forma a atingir os objetivos estabelecidos e, por outro, a garantir requisitos específicos nos dados de entrada;
- e) Avaliação: o(s) modelo(s) construídos devem ser avaliados com medidas de qualidade adequadas e também com a revisão de todos os passos e opções que foram sendo tomadas durante a execução de todo o processo. O investigador deverá certificar-se que o modelo representa uma boa solução para os desafios que tinham sido colocados ao processo de DM e deverá estar em condições de decidir sobre a viabilidade da utilização dos resultados;
- f) Aplicação: consideradas boas soluções, os modelos obtidos são incorporados em processos de tomada de decisão e em novos estudos analíticos. Esta fase representa a integração dos resultados nas atividades da organização.

2.4.2.2 Metodologia SEMMA

A metodologia SEMMA foi desenvolvida pelo Instituto SAS⁵, dedicado ao desenvolvimento de soluções estatísticas, análise de dados, *business intelligence*, DM e suporte à decisão (Santos e Azevedo, 2005).

⁵ Construtor privado de *software* a nível mundial e líder no mercado de *Business Analytics*. O *software* SAS é utilizado em várias organizações (empresas privadas, sector público, governo e universidades).

O instituto SAS define DM como o processo de extrair informação valiosa de relações complexas de um grande volume de dados. Assim, o processo de DM é dividido em cinco etapas, que compõem o acrónimo SEMMA: *Sample* (Amostragem), *Explore* (Exploração), *Modify* (Modificação), *Model* (Modelação) e *Assessment* (Avaliação) (Cohen et al., 1999; Berson et al., 2000; Groth, 2000).

Esta metodologia é constituída por diferentes etapas, correspondendo a um ciclo, podendo estas ser executadas repetidamente quando necessário.

A Figura 2-13 ilustra as tarefas subjacentes à metodologia SEMMA ao longo das suas cinco etapas.

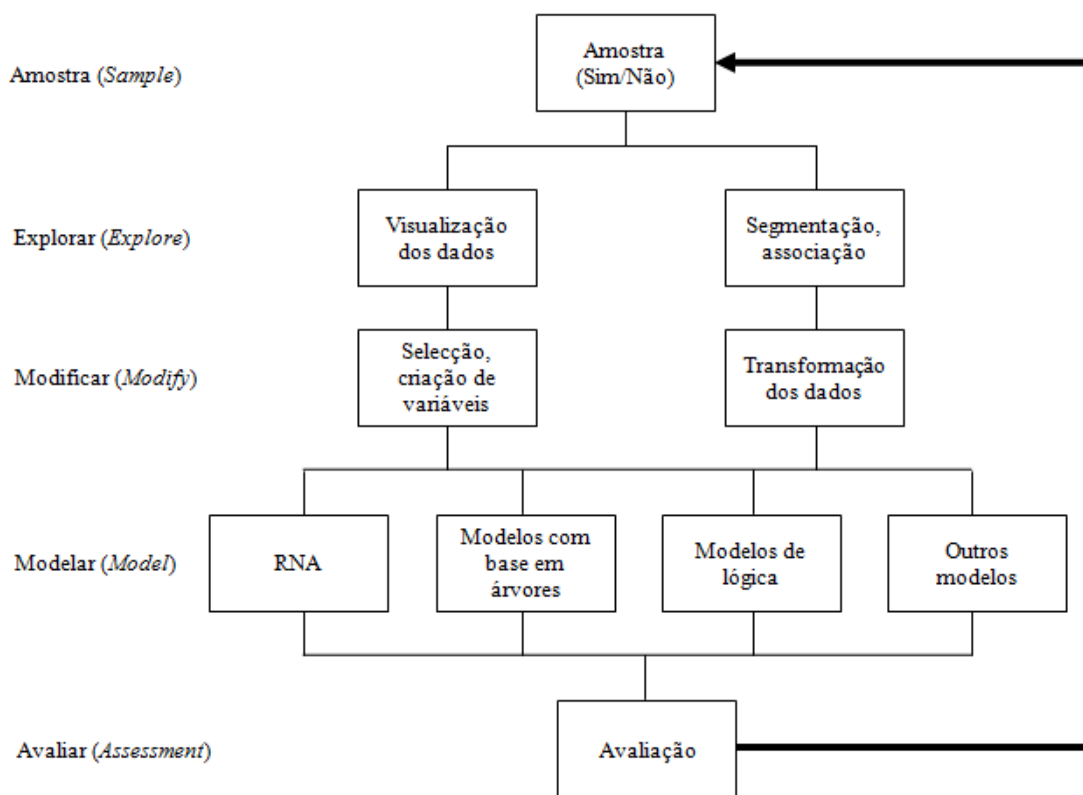


Figura 2-13 - Etapas da Metodologia SEMMA (Santos e Azevedo, 2005).

- a) Amostragem: o início desta etapa dá-se com a extração de dados do universo em estudo. A amostra deve ser significativa, no entanto deverá também ser suficientemente pequena de forma a tornar fácil e rápida a sua manipulação. Basear o processo de DM numa amostra representativa reduz drasticamente o volume e o tempo de processamento necessários de forma a retirar informação crucial para o negócio (Cohen et al., 1999);

- b) Exploração: nesta etapa são exploradas visualmente ou numericamente (através de gráficos de distribuição, histogramas, tabelas, etc.) as tendências ou agrupamentos inerentes aos dados. A exploração ajuda a refinar o processo de descoberta (Santos e Azevedo, 2005). Nesta etapa poderá recorrer-se ao uso de técnicas avançadas de estatísticas, a distribuição de poisson, os mínimos quadrados, o qui-quadrado e a regressão linear;
- c) Modificação: nesta etapa, concentram-se todas as transformações necessárias com base nos resultados obtidos na etapa anterior. Pode haver a inclusão de informação, seleção ou introdução de novas variáveis, de forma a se conseguir obter variáveis mais significativas;
- d) Modelação: nesta etapa, definem-se técnicas de construção de modelos de DM que incluem as técnicas de aprendizagem automática, RNA, IR, AD, RB e técnicas estatísticas (regressão linear, indução de probabilidades). Esta etapa tem como objetivo selecionar as técnicas de construção de modelos, de forma a prever com confiança os resultados propostos (Cohen et al., 1999);
- e) Avaliação: etapa final onde é avaliado o desempenho do modelo, geralmente através da aplicação deste, numa amostra de dados selecionada para este fim (conjunto de teste). Esta etapa tem como objetivo aplicar o modelo à amostra de dados e verificar o seu desempenho, podendo assim proceder a alguns ajustes, caso necessário (Cohen et al., 1999).

Ambas as metodologias, CRISP-DM e SEMMA, foram desenvolvidas com a finalidade de ajudar no processo de DM e de resolver os problemas de negócio de forma mais rápida e exequível.

Da Figura 2-14 depreende-se que a metodologia CRISP-DM é mais completa e por conseguinte mais vantajosa, uma vez que todo o processo se inicia com o estudo do negócio e que as suas etapas estão orientadas para os objetivos especificados do negócio, traduzindo-se numa forma mais direta de condução do projetos de DM.

FASES DA METODOLOGIA CRISP-DM

- Estudo do negócio;
- Estudo dos dados;
- Preparação dos dados;
- Modelação;
- Avaliação;
- Aplicação;

FASES DA METODOLOGIA SEMMA

- Amostra;
- Explorar;
- Modificar;
- Modelar;
- Avaliar;

Figura 2-14 - Fases das Metodologias CRISP-DM e SEMMA. Adaptado de Santos e Azevedo (2005).

A par destas metodologias mais conhecidas, que definem todo o processo de DCBD, outras metodologias foram desenvolvidas de forma a definir tarefas mais específicas. O desenvolvimento de *API's*⁶ encontra-se em ascensão devido à capacidade de um programador poder aceder aos recursos de *data mining* de um determinado *software* e processar os resultados através de outro distinto (François, 2008).

2.4.2.3 Especificação PMML

Esta especificação desenvolvida pelo *Data Mining Group* (DMG) proporciona um padrão de interpretação *standard* para representação de inúmeros modelos de DM tais como: as regras de associação, os modelos de *clustering*, as RNA e as AD. Desta forma, os modelos podem ser facilmente compartilhados entre diferentes ferramentas de DM, evitando problemas de incompatibilidade (Guazzelli, 2010).

A PMML (Figura 2-15) é capaz de representar uma ampla gama de técnicas estatísticas e de IA, podendo ser utilizada também para representar os respetivos dados de entrada, bem como todas as transformações necessárias de forma a tornar esses mesmos dados em novas características significativas (Guazzelli et al., 2009).

A PMML incorpora pré-processamento e pós-processamento de dados, bem como o próprio modelo preditivo. Esta especificação (atualmente na versão 4.2⁷) é baseada na linguagem XML⁸, podendo proporcionar a diferentes aplicações uma forma de definir modelos relacionados com análises preditivas e DM, e serem compartilhados por diferentes aplicações PMML.

⁶ API, de *Application Programming Interface* (ou Interface de Programação de Aplicativos) é um conjunto de rotinas e padrões estabelecidos por um *software* para a utilização das suas funcionalidades por aplicativos que não pretendem envolver-se em detalhes da implementação do *software*, mas apenas usar os seus serviços.

⁷ Lançada em 28 de fevereiro de 2014.

⁸ XML (*eXtensible Markup Language*). O XML é um formato para a criação de documentos com dados organizados de forma hierárquica.

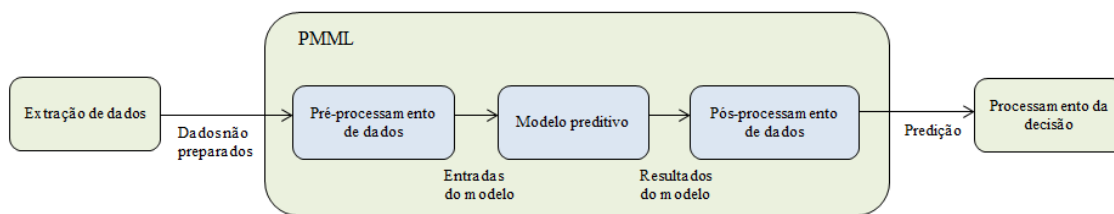


Figura 2-15 - Representação gráfica da especificação PMML. Adaptado de Guazzelli et al. (2009).

Resumidamente, é permitida a construção de um modelo de previsão usando uma determinada aplicação (por exemplo, um *software* de desenvolvimento) e usá-lo noutro (por exemplo, um sistema *on-line* em tempo real) para visualizar, analisar e avaliar resultados ou até utilizar esses mesmos modelos (Ferguson e Huston, 1998).

A PMML pode ser sumarizada como é indicado na Tabela 2-1:

Tabela 2-1 - Componentes do PMML. Adaptado de Kadav et al. (2003).

Componente	Descrição
Bases de dados	Contém os dados e suas especificações que não variam com o modelo
Esquema de DM	Contém informação específica para um determinado modelo e varia com o mesmo. Cada modelo contém um esquema de DM que lista os campos utilizados nesse mesmo modelo. Estes campos consistem num subconjunto dos campos da base de dados.
Transformação de dados	Define campos derivados. Campos derivados podem ser definidos por: normalização, agregação, etc.
Modelo estatístico	Aplicação de funções básicas de estatística aos dados numéricos existentes nos modelos de DM gerados, tais como: máximo, mínimo, desvio padrão, média, etc.
Parâmetros do modelo	Especifica os parâmetros reais que definem os dados estatísticos e os modelos de DM por si só. Os diferentes modelos suportados consistem em regras de associação, modelos de clustering, RNA e AD.
Objetivos do DM	A utilização de diferentes modelos para atingir diferentes objetivos, sendo estes o de classificação, previsão, descrição, segmentação e visualização.

A Figura 2-16 exemplifica a sintaxe de um documento PMML.

```
<?xml version="1.0"?>
<PMML xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns="http://www.dmg.org/PMML-2_1" version="2.1">
  - <Header copyright="Copyright (c) Microsoft Corporation, All Rights Reserved">
    <Application version="9.0" name="Microsoft Analysis Services"/>
  </Header>
  - <DataDictionary numberOfFields="2">
    <DataField name="InputStat" dataType="double" isCyclic="0" optype="continuous" displayName="InputStat"/>
    <DataField name="SecondInputStat" dataType="double" isCyclic="0" optype="continuous" displayName="SecondInputStat"/>
  </DataDictionary>
  <TransformationDictionary/>
  - <NaiveBayesModel threshold="1" algorithmName="Internal_Name_For_My_Algorithm" functionName="classification" modelName="PluginModel">
    - <Extension name="MiningAlgorithmFlags" extender="Microsoft Analysis Services">
      <key-val name="MyFeatureSelectionParameter" value="1"/>
    </Extension>
    - <MiningSchema>
      <MiningField name="InputStat" usageType="active"/>
      <MiningField name="InputStat" usageType="predicted"/>
      <MiningField name="SecondInputStat" usageType="active"/>
      <MiningField name="SecondInputStat" usageType="predicted"/>
    </MiningSchema>
    - <ModelStats>
      - <UnivariateStats field="InputStat">
        <Counts missingFreq="0" totalFreq="14"/>
        <NumericInfo standardDeviation="7.86363368258451" mean="11.1428571428571" maximum="20" minimum="1"/>
      </UnivariateStats>
      - <UnivariateStats field="SecondInputStat">
        <Counts missingFreq="0" totalFreq="14"/>
        <NumericInfo standardDeviation="7.38172039015075" mean="-1.71428571428571" maximum="8" minimum="-10"/>
      </UnivariateStats>
    </ModelStats>
  </NaiveBayesModel>
</PMML>
```

Figura 2-16 - Exemplo de ficheiro PMML.

2.4.3 Objetivos em *data mining*

Recorrendo à análise da Figura 2-17 é possível verificar os tipos de objetivos mais frequentes no contexto da DCBD, bem como os algoritmos mais adequados.

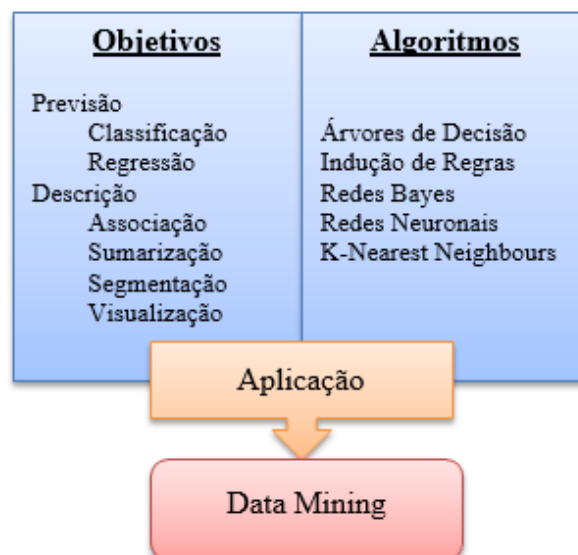


Figura 2-17 - Objetivos e algoritmos de *data mining*.

Através da análise da figura verifica-se que existem vários objetivos de DM, sendo normalmente enquadrados nas categorias de Previsão e Descrição.

2.4.3.1 Previsão

A previsão envolve o uso de algumas variáveis ou campos numa Base de Dados no sentido de prever valores desconhecidos ou valores para outras variáveis de interesse. O objetivo principal do DM preditivo consiste em automatizar o processo de tomada de decisão, criando um modelo capaz de prever ou estimar um valor. Por outras palavras, a previsão prevê valores futuros de variáveis de interesse com base em algumas variáveis, bem como possibilita a descoberta de padrões de comportamento dessas mesmas variáveis a partir de exemplos. Normalmente, os resultados do modelo são utilizados diretamente sobre as próprias variáveis de interesse, tornando a exatidão do modelo como a medida de desempenho mais importante para a sua avaliação, tanto nos casos de classificação como nos de regressão.

2.4.3.1.1 Classificação

A classificação corresponde à possibilidade de encontrar uma função passível de associar um caso a uma classe dentro de diversas classes discretas⁹ de classificação, podendo assim classificar um novo objeto de acordo com um modelo de classificações. As diferentes técnicas utilizadas recorrem a conjuntos de treino com exemplos pré-classificados tendo como finalidade a construção de modelos adequados à descrição das classes, sendo posteriormente aplicados a dados não classificados.

2.4.3.1.2 Regressão

A regressão, também frequentemente referida como previsão, tem como objetivo prever valores futuros ou desconhecidos de uma variável dependente, a partir de exemplos. Funcionalmente é caracterizada pela atribuição de um peso relativo a cada variável (i.é., valor numérico), permitindo valorizar cada registo, em função do seu conteúdo e relativizá-lo face aos outros.

⁹ Variável discreta: quando os elementos são numeráveis, números inteiros, i.e., conjunto finito.

2.4.3.2 Descrição

O objetivo primário do DM descritivo consiste em aumentar o conhecimento e a compreensão sobre os dados, focando a procura de padrões descritivos reconhecidos ou interpretáveis pelos humanos. Os resultados obtidos nem sempre se traduzem em ações com aplicação direta. As novas perspectivas e os conhecimentos adquiridos sobre os dados durante a construção do modelo, são o aspeto mais importante do processo.

2.4.3.2.1 Associação

Esta técnica permite definir um modelo que descreva possíveis dependências significativas entre variáveis, através da identificação de grupos de dados fortemente correlacionados. As associações detectam-se quando se verificam várias ocorrências num único evento, podendo surgir a nível estrutural (i.é., o modelo é representado de uma forma gráfica e com variáveis localmente dependentes em relação a outras) ou quantitativo (i.é., o modelo especifica o peso das dependências segundo uma escala numérica).

2.4.3.2.2 Sumarização

O seu objetivo consiste em determinar uma descrição compacta para um subconjunto de dados, através da aplicação de métodos próprios. Os métodos de sumarização mais sofisticados derivam de regras de resumo e descobertas de relações funcionais entre variáveis (i.é., relação entre a naturalidade e a preferência clubística). As técnicas de sumariação são sempre aplicadas à análise exploratória de dados e à geração automática de relatórios.

2.4.3.2.3 Segmentação

A segmentação ou *clustering* permite identificar um conjunto finito de categorias para descrever dados. Permite identificar grupos homogêneos de objetos em que cada grupo consiste numa classe. Dentro da mesma classe os objectos são semelhantes e entre classes discrepantes. A segmentação assegura que a similaridade inter-segmentos é baixa e a similaridade intra-segmentos é alta.

2.4.3.2.4 Visualização

A visualização trata a representação de resultados finais ou intermédios de DM, através de formatos visuais, geralmente gráficos. Pretende descrever informações complexas através de diagramas, permitindo uma melhor representação de padrões e tendências.

2.4.4 Modelos, técnicas e algoritmos

Por “modelo” entende-se a representação matemática de um sistema com o objetivo de o estudar, sendo a descrição formal e simplificada de um sistema, ou seja, a estruturação da representação do conhecimento com vista a atingir um resultado (Rud, 2001).

Com “técnica” compreende-se o conjunto de processos baseados em conhecimentos científicos, de cálculo ou experimentação, utilizados para a obtenção de um resultado (Teixeira, 2005).

Por “algoritmo” entende-se um qualquer procedimento computacional bem definido que toma algum valor ou conjunto de valores como entrada produzindo algum valor ou conjunto de valores como saída. Desta forma, um algoritmo consiste numa sequência de passos computacionais que transformam a entrada em saída. Um algoritmo também pode ser visto como uma ferramenta para resolver um problema computacional bem especificado. O enunciado do problema especifica em termos gerais o relacionamento entre a entrada e a saída desejada. O algoritmo descreve um procedimento computacional específico de forma a alcançar esse relacionamento da entrada com a saída (Cormen et al., 2009).

Os modelos podem ser estáticos, quando representam um sistema num determinado momento, ou dinâmicos, quando o representam ao longo do tempo, acompanhando todas as alterações consequentes do seu funcionamento.

Deve destacar-se que não existe um modelo universal de DM que resolva, de forma eficiente, todos os problemas. A escolha de um determinado algoritmo é de certa forma uma arte (Fayyad et al., 1996a), uma vez que existem diferentes modelos para as mesmas tarefas de DM com vantagens e desvantagens intrínsecas.

Em relação ao tipo de aprendizagem, ou seja, ao modo como os algoritmos aprendem, estes podem ser classificados como sendo de aprendizagem supervisionada e não supervisionada. Na aprendizagem supervisionada, os algoritmos utilizam dados para treino e dados para validação e posterior teste dos modelos gerados. Por exemplo podem-se considerar 2/3 dos dados para treino

e 1/3 para validação (teste). Este tipo de aprendizagem é bastante popular envolvendo a presença de um “professor”, sendo fornecidas respostas corretas à rede. Perante uma configuração que é apresentada um algoritmo produz uma resposta, que é comparada com a resposta correta.

Na aprendizagem não supervisionada não existe divisão entre dados de treino e validação, assim, o algoritmo deve tentar encontrar os padrões através da caracterização e/ou segmentação destes dados. Este tipo de aprendizagem segue uma abordagem diferente, onde não é fornecida ao sistema uma indicação externa acerca da resposta correta. A aprendizagem é realizada através da identificação de características nos dados de entrada, adaptando-se a regularidades estatísticas ou agrupamentos de padrões dos exemplos de treino.

Neste campo existe um vasto volume de algoritmos, dos quais se podem destacar, devido à sua maior utilização, as AD, as RNA, os algoritmos genéticos, a IR, as RB e os sistemas de classificação (Santos e Azevedo, 2005).

Segundo Rezende (2003), as várias técnicas de aprendizagem automática (AA) utilizáveis no contexto do DM, agrupam-se nos cinco seguintes paradigmas:

- i) Simbólico – Procuram aprender através de exemplos e contra-exemplos e representações simbólicas de conceitos (AD e regras de indução);
- ii) Estatístico – Utilização de modelos estatísticos, destacando-se os métodos paramétricos e os modelos de Bayes (Mitchel, 1998). As RNAs podem ser consideradas como métodos estatísticos paramétricos;
- iii) Baseado em exemplos – Classificam exemplos nunca antes analisados através de exemplos similares conhecidos. As técnicas que se enquadram nesta classe são: o raciocínio baseado em casos e a técnica designada por *Nearest Neighbours*;
- iv) Conexionista – As RNAs correspondem a construções matemáticas simplificadas inspiradas no modelo biológico do sistema nervoso;
- v) Evolutivo – Paradigma que deriva do modelo biológico associado à evolução. Podemos incluir nesta classe os algoritmos genéticos e os sistemas de classificação.

De seguida é efetuada uma descrição dos algoritmos utilizados em DM e anteriormente citados.

2.4.4.1 Árvores de decisão

A árvore de decisão é um dos algoritmos mais popular e mais utilizado em DM. A sua popularidade deve-se principalmente à fácil compreensão da sua representação gráfica, que pode ser indubitavelmente transformada em regras facilmente compreensíveis (Su et al., 2009).

Neste algoritmo, é escolhida a variável a avaliar e o *software* procura nas restantes variáveis as mais correlacionadas, elaborando uma árvore com várias ramificações. A árvore criada consiste num conjunto de regras que seguem uma hierarquia de classes e valores (Figura 2-18).

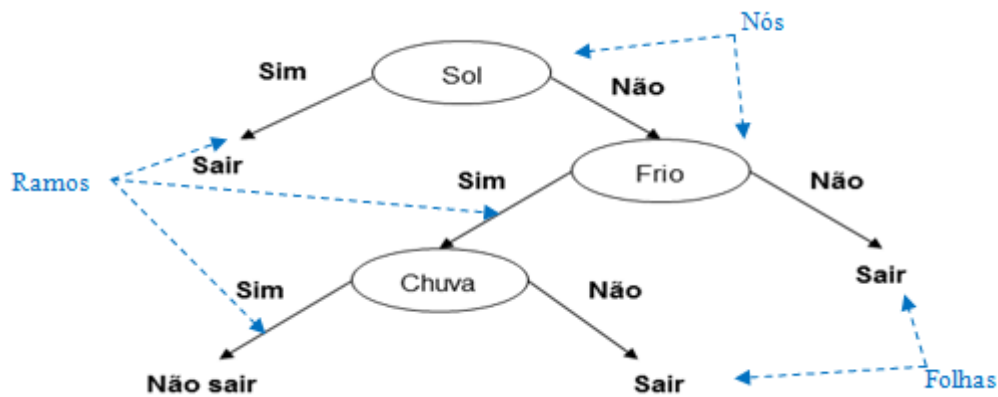


Figura 2-18 - Árvore de Decisão – Exemplo: Sair de casa.

Dado um determinado conjunto de dados, torna-se necessário selecionar o campo objeto de estudo (*output*). Posteriormente, com o primeiro ramo (nó) da árvore de decisão, é apresentado o único e mais importante fator correlacionado com o objeto de saída. Desta forma, é possível visualizar rapidamente qual o factor que mais influencia o objecto de estudo.

Uma árvore de decisão representa um conjunto de regras de classificação desde o nó de raiz até aos nós terminais (i.e. folhas). A cada nó da árvore é especificado um conjunto de testes para as variáveis em estudo e cada nó descendente corresponde a valores possíveis dessas variáveis. Todo este processo é repetido para os novos nós, tomando estes como nova raiz. O objetivo é gerar a árvore que melhor se adequa ao problema.

Apresenta-se a seguir um exemplo de um algoritmo de indução de AD (Figura 2-19). O algoritmo constrói árvores partindo de dados de treino, de forma recursiva, subdividindo este conjunto de dados até que seja formada apenas por nós “puros”, em que cada nó represente apenas uma única classe, ou satisfação de um critério. A estrutura das árvores geradas são constituídas por folhas (nós puros) que representem a classes/objetos, os nós internos correspondem aos atributos que especificam um teste efetuado num único atributo com duas ou mais saídas possíveis (subárvores) e por fim os ramos que correspondem aos valores dos atributos (Michalsky et al., 1998).

Seja T a árvore de decisão a induzir

Seja S o conjunto de exemplos para aprendizagem

Se todos os exemplos de S pertencem à mesma classe C

Então a árvore tem um nó C

Senão seleccionar o atributo A mais “esclarecedor” cujos valores são v_1, \dots, v_n

particionar S em n subconjuntos S_1, \dots, S_n para cada valor v_i de A

construir (recursivamente) sub árvores T_1, \dots, T_n para cada S_1, \dots, S_n .

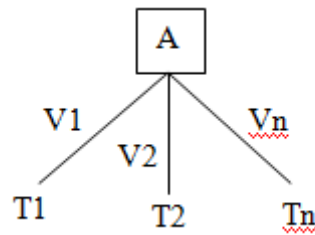


Figura 2-19 - Indução de Árvores de Decisão. Adaptado de Rich e Knight (1991).

Durante o processo de treino, as AD devem, repetidamente, encontrar a forma mais eficiente de separar os registos em dois nós. O objectivo é encontrar o atributo mais útil para a classificação dos exemplos. As medidas que quantificam a utilidade de um atributo avaliam como é feita a separação dos exemplos de treino de acordo com o valor da classe alvo.

Existem duas métricas de homogeneidade, *Gini* (medida de impureza) e *Entropy*, para avaliação da separação dos registos. O algoritmo usa estas métricas para avaliar a qualidade das várias alternativas das condições de separação e selecciona a mais adequada, consoante a métrica escolhida (Quinlan, 1998).

A métrica de *Gini* atribui a um dos ramos a percentagem mais elevada de uma classe e a entropia tenta balancear e separar as classes tanto quanto possível (Quinlan, 1986).

Para um conjunto de dados S contendo exemplos de n classes, $Gini(A)$ é definido pela expressão (2-11).

$$Gini(A) = \sum p(v) \sum_{i \neq j} p(i|v)p(j|v) \quad (2-11)$$

Onde:

i, j – classes;

A – atributo;

v – valores do atributo A ;

$p(v)$ – probabilidade do valor v ocorrer no conjunto S ;

$p(i|v)$ – probabilidade de na classe i o atributo A ter o valor v .

Consiste num índice computacional gerado por Ron Bryman que oferece uma enorme velocidade de processamento e capacidade de trabalhar quer com números, quer com texto (Berry e Linoff, 1993).

A vantagem deste método é de envolver um cálculo relativamente simples efetuando a distribuição de classes em cada nó. O índice de Gini é calculado para cada atributo, utilizando para o nó o atributo com menor índice, isto é, o atributo mais informativo é aquele que tem menor índice de impureza.

Após a construção da árvore, a “poda” passa a ter um papel importantíssimo produzindo árvores menores e com maior potencial e exatidão (Quinlan, 1998).

Existem dois tipos de poda (Cios et al., 1998), durante a aprendizagem (*forward pruning* ou *pre-pruning*) e após aprendizagem (*post-pruning* ou *backward pruning*). A maior parte dos algoritmos utiliza a técnica de poda após aprendizagem. A poda durante a aprendizagem é feita em simultâneo com a construção da árvore o que acarreta a tomada de decisões complexas em tempo real. No entanto, neste último caso a vantagem persiste em que na maior parte das vezes só é possível determinar se uma subárvore deve ser podada depois da construção da árvore inteira (Santos e Azevedo, 2005). Porém e mesmo após a poda, uma árvore pode apresentar uma estrutura bastante complexa e de difícil compreensão, levando ao uso de IR para extrair regras menos complexas e de maior precisão (Cios et al., 1998).

Apresenta-se a seguir um algoritmo de poda após aprendizagem de uma árvore de decisão.

Seja T uma árvore de decisão completa não podada

 calcular o erro estático para cada folha da árvore

para todo o nó e folha no sentido das folhas para a raiz

 calcular: erro estático e o erro propagado

 se o erro estático \leq erro propagado

 então podar a árvore passando esse nó a ser uma folha

Sendo que o erro estático é o erro de classificação de um nó (2-12):

$$E_{Est}(nó) = \frac{N - n + k - 1}{N + k} \quad (2-12)$$

Onde:

N – número de exemplos classificados no nó

n – número de exemplos pertencentes à classe maioritária (com mais exemplos)

k – número de classes total do problema

O erro propagado é o erro de classificação de um nó devido à propagação de erros das subárvores (2-13):

$$E_{Prop}(nó) = \sum_i P_i * E_{Est}(nó_i) \quad (2-13)$$

Onde:

i – ramo do nó

P_i – frequência relativa dos exemplos no ramo i

$nó_i$ – sub nó de nó

Porém na fase de poda surge o problema relacionado com a forma efetuar uma previsão das taxas de erro. De acordo com Quinlan (1998), existem dois tipos de técnicas para previsão da taxa de erro:

- a) A primeira efetua uma previsão da taxa de erro com base em dois conjuntos de dados, teste e treino. Na construção da árvore em que se recorre aos dados de teste, os dados de treino não foram examinados, assim sendo, as estimativas obtidas a partir dos mesmos não são tendenciosas. Porém, a desvantagem desta técnica está associada à necessidade de um grande volume de dados por forma a que se possa separar um conjunto para construção da árvore e outro para a poda. Quando os dados são escassos, poderá ser utilizada a técnica de validação cruzada (*Cross-Validation*), onde os dados de treino são divididos em n conjuntos de igual tamanho e correspondentemente são feitos n pares de treino e teste: no primeiro par o subconjunto 1 é usado para teste e o restante para treino; no segundo par o subconjunto 2 é usado para teste e o restante para treino; e assim por diante.

- b) A segunda utiliza apenas os dados de treino a partir dos quais a árvore foi construída. A principal desvantagem prende-se com a necessidade de uma quantidade significativa de dados para descobrir estruturas complexas. No entanto, as vantagens residem no facto de levarem em consideração as regras mais relevantes, permitindo determinar quais os fatores principais de forma mais rápida.

Resumidamente, as AD (Tabela 2-2) consistem num dos métodos mais simples e adequados para problemas com muitas dimensões, tratando-se de um método preditivo não linear rápido.

Tabela 2-2 - Resumo das Árvores de Decisão. Adaptado de Santos e Azevedo (2005).

Objectivos do DM	Vantagens	Desvantagens
- Classificação	- Construção simples e rápida	- Necessário muitos dados para descobrir estruturas complexas
- Segmentação	- Adequado a problemas com muitas dimensões	
- Previsão		
- Visualização	- Fácil representação e visualização	

2.4.4.2 Indução de regras

A partir de uma árvore de decisão é possível extrair um conjunto de regras representativas do modelo através da técnica de IR. Tal é mostrado na Figura 2-20 através da apresentação das regras associadas ao modelo apresentado na Figura 2-18.

<p>Regra #1</p> <p><i>If Sol = 'Sim'</i></p> <p><i>then Sair.</i></p>	<p>Regra #3</p> <p><i>If Sol = 'Não'</i></p> <p><i>and Frio = 'Sim'</i></p> <p><i>and Chuva = 'Não'</i></p> <p><i>then Sair.</i></p>
<p>Regra #2</p> <p><i>If Sol = 'Não'</i></p> <p><i>and Frio = 'Sim'</i></p> <p><i>and Chuva = 'Sim'</i></p> <p><i>then Não Sair.</i></p>	<p>Regra #4</p> <p><i>If Sol = 'Não'</i></p> <p><i>and Frio = 'Não'</i></p> <p><i>then Sair.</i></p>

Figura 2-20 - Exemplo de regras de indução.

A IR consiste na detecção de tendências e padrões em grupos de dados sendo uma das técnicas de DM mais conhecida (Berson et al., 2000).

O principal objetivo deste algoritmo consiste na obtenção de dependências entre os valores ou atributos. Geralmente os resultados são apresentados em forma de regras $X \rightarrow Y$, significando que “se X está presente, então Y também está presente com uma determinada probabilidade” (Toloo et al., 2009).

O elemento X pode ser uma combinação de atributos e valores, formando assim regras mais complexas. Estas regras têm dois graus associados: o grau de confiança e o nível de suporte. A primeira é a probabilidade da condição da regra se verificar, e a segunda, o número de casos em que a regra se verifica.

As regras devem satisfazer três propriedades (Quinlan, 1998; Kufirin, 1997):

- Previsões corretas, ou seja, na maioria das vezes se o SE da regra é verdadeiro, o ENTÃO da regra também é verdadeiro;
- Compreensíveis, ou seja, as regras representam conhecimento a um alto nível de abstração, devendo ser de fácil compreensão para o utilizador, em contrapartida das equações matemáticas complexas;
- Úteis para a tomada de decisão que está relacionada com o facto da regra expressar conhecimento novo ou surpreendente para o utilizador.

As vantagens deste algoritmo consistem:

- No modo direto de lidar com os dados;
- No seu desempenho;
- Na facilidade de explicação e compreensão das regras;
- E os passos para a solução do problema serem facilmente observáveis.

A IR surge associada às AD. De uma forma geral, utiliza-se a IR de forma a representar o conhecimento associado às AD.

2.4.4.3 Redes de Bayes

O teorema de *Bayes* é fundamentado na teoria das probabilidades, permitindo representar de forma numérica o grau de certeza de um determinado evento, podendo ser manipulado de acordo com as regras definidas na teoria da probabilidade (Heckerman e Wellman, 1995).

As *redes Bayesianas* consistem em mecanismos eficientes para análise de dados que apresentam relacionamentos de precedência temporal. Uma RB é constituída por dois componentes: uma estrutura gráfica e parâmetros numéricos.

Os sistemas baseados em RB são capazes de gerar automaticamente predições ou decisões mesmo na situação de inexistência de alguns dados utilizados na fase de modelação (Heckerman et al., 1995).

O teorema de *Bayes* é definido como (Zerbinatti e Júnior, 2006) (2-14):

$$P(H/X) = \frac{P(X/H)P(H)}{P(X)} \quad (2-14)$$

Onde:

- $P(H)$ e $P(X)$ são as probabilidades à priori de H e X
- $P(X/H)$ e $P(H/X)$ são as probabilidades à posteriori de X condicional a H e de H condicional a X respetivamente.

A regra de *Bayes* (Figura 2-21) mostra como alterar as probabilidades à priori, tendo em conta novas evidências, de forma a obter probabilidades à posteriori. O classificador que maximiza $P(H/X)$ é conhecido como classificador MAP (*maximum a posteriori*). As RB baseiam-se na teoria de *Bayes* e consistem numa técnica fundamental no reconhecimento de padrões e classificação.

Este algoritmo assume que os atributos são independentes, mas tem demonstrado um bom desempenho mesmo em situações onde são encontradas claras dependências entre atributos. A força da dependência é representada por probabilidades condicionais que são associadas a cada grupo de nós pais-filhos existentes na rede. Dada a distribuição de probabilidade conjunta $P(x_1, x_2, x_3, \dots, x_n)$ e uma determinada ordenação d destas variáveis, inicia-se a construção do grafo, escolhendo o nó raiz X_1 e atribuindo-lhe a probabilidade marginal $P(x_1)$. Em seguida, acrescenta-se mais um nó X_2 ao grafo. Se X_2 for dependente de X_1 então é traçado um arco ligando ambas as variáveis através de uma seta apontando para X_2 e quantifica-se este arco com $P(x_2/x_1)$, caso contrário, mantém-se as variáveis desconectadas e atribui-se uma probabilidade à priori $P(x_2)$ a X_2 . A operação é repetida em relação às restantes variáveis, após o qual é obtida uma RB (Heckerman e Wellman, 1995).

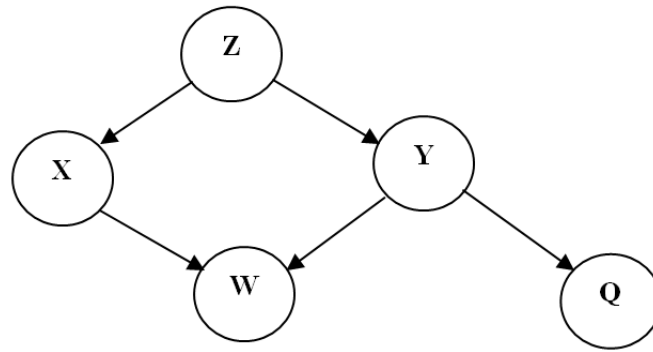


Figura 2-21 - Rede de Bayes.

A RB tem um comportamento bastante robusto ao ruído e a atributos irrelevantes. Toda a informação necessária à construção de um modelo é adquirida com uma única passagem pelos dados o que o torna um dos algoritmos de classificação e predição mais rápidos.

Para a representação de conhecimento, as RB consistem em grafos acíclicos com as seguintes características: os nós correspondem a variáveis aleatórias e uma ligação direcionada ou arco liga pares de variáveis. O significado de um arco dirigido do nó Z ao nó Y é que Z tem uma influência direta sobre Y. Cada nó tem associado os estados da variável e uma tabela de probabilidades condicionadas que quantifica os efeitos que os pais exercem sobre um nó (probabilidade do nó estar num estado específico dado os estados dos seus pais) (Lampinen e Vehtari, 2001).

Segundo (Lemm, 1999), a topologia da rede pode ser vista como uma base de conhecimento abstrata, representando a estrutura dos processos causais no domínio.

As RB podem ser consideradas diagramas que organizam o conhecimento através de um mapeamento entre causas e efeitos, podendo então calcular a probabilidade de um evento ocorrer, condicionado pela ocorrência de outro (Heckerman e Wellman, 1995). Estas redes têm a capacidade de gerir conjuntos de dados incompletos, devido ao facto de proporcionar um modo natural para codificar a dependência entre variáveis de entrada.

Uma variante das RB consiste nas redes *Naive Bayes* também utilizadas para criação de modelos de previsão. A designação “*naive*” em *Naive Bayes* é-lhe atribuída pelo facto de o algoritmo usar técnicas *Bayesianas*, mas não considerar dependências que possam existir. Portanto, as suas suposições são consideradas ingênuas por se sustentarem em dois importantes pressupostos (Buntine, 1994):

- i) A suposição que os atributos preditivos são condicionalmente independentes dada a classe.

- ii) Postula-se que nenhum atributo oculto ou subentendido influencia o processo de predição.

O classificador *Naive Bayes*, fornece uma abordagem simples e com semânticas claras para representação, uso e aprendizagem de conhecimento probabilístico. É indicado para o contexto de aprendizagem supervisionada, no qual a meta de desempenho é a predição precisa das instâncias de teste e o conhecimento gerado sobre os dados na fase de treino (John e Langley, 1995).

O algoritmo *Naive Bayes* é um classificador probabilístico, que utiliza cálculos de probabilidade baseando no Teorema de *Bayes* associado a fórmulas estatísticas para encontrar modelos. Esses modelos são gerados utilizando uma fórmula onde os resultados dos cálculos probabilísticos são considerados conhecimento prévio.

2.4.4.4 Redes neuronais artificiais

As RNAs são um dos algoritmos mais conhecidos e usados em DM (Berson et al., 2000).

Uma RNA consiste num conjunto de elementos de processamento simples (nós), com um grande número de interligações. Toda a sua estrutura é baseada na recriação do cérebro humano, isto é, na capacidade de aprender e auto corrigir-se.

Uma RNA consiste numa estrutura computacional baseada em unidades de processamento (neurónios) ligadas entre si e organizadas em grupos (camadas) (Santos e Azevedo, 2005). Todos os neurónios comunicam entre si enviando sinais através das ligações e todo o processamento da informação é realizado como se ocorresse simultaneamente nos vários neurónios. Cada um possui um conjunto de ligações de entrada e de saída e cada ligação tem um peso associado.

Para a construção de uma RNA, tem de se determinar o número de neurónios (nodos), o seu tipo, a forma como estes vão estar ligados, definir a sua codificação de entrada e descodificação de saída, iniciar os pesos da rede e proceder ao treino da rede aplicando um algoritmo apropriado (Groth, 2000).

Cada nodo recebe uma série de valores e, em função deles, determina um valor a apresentar como saída. Os valores de saída de alguns nodos, são em alguns casos valores de entrada de outros, dependendo da interligação existente entre eles. A forma como esta interligação se encontra estabelecida é importante para o tipo de resultados obtidos (Gago, 2001).

A função básica de cada neurónio consiste em:

- i) avaliar os valores de entrada;

- ii) calcular o total para valores de entrada combinados;
- iii) comparar o total com um valor limiar;
- iv) determinar o que será a saída.

Enquanto a operação de cada neurónio for razoavelmente simples, procedimentos complexos podem ser criados através da conexão de um conjunto de neurónios. Tipicamente, as entradas dos neurónios são ligadas a uma camada intermédia (ou várias camadas intermédias) que é então conectada a uma camada de saída.

Segundo Cortez (2002), no neurónio artificial existem três elementos fundamentais, o conjunto de sinapses ou elos de conexão, o combinador linear (Σ) e a função de ativação (Figura 2-22):

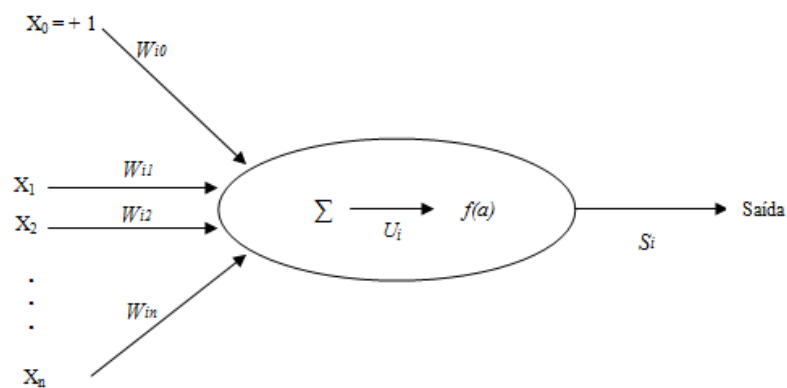


Figura 2-22 - Neurónio artificial (Gershenson, 2003).

Um conjunto de ligações (w_{ij}) representam as sinapses ou ligações entre neurónios. Cada ligação tem relacionada um peso representando a força do sinal enviado, sendo positivo para as sinapses de excitação e negativo se a sinapse é inibidora. Assim, o sinal de entrada (x) é multiplicado pelo peso correspondente w , onde i denota o nó do objecto de estudo e j o nó de onde partiu o sinal. Por norma, existem pseudo ligações relacionadas com neurónios virtuais cujo valor é constante e igual a 1. O círculo que representa o neurónio, reduz os n argumentos de entrada e os estímulos a um único valor de saída. As RNAs utilizam a função adição (Σ), pesando todas as entradas numa combinação linear. O nível de activação do neurónio é definido por uma função de activação $f(a)$, que limita a amplitude da saída do neurónio, introduzindo uma componente não linear.

Sendo que U consiste na soma pesada das entradas no neurónio (2-15).

$$U_i = \sum(W_0xW_{i0}, X_1xW_{i1}, X_2xW_{i2}, \dots, X_nxW_{in}) \quad (2-15)$$

Temos S como o resultado do cálculo do valor de ativação do neurónio através da utilização da função de ativação (2-16).

$$S_i = fa(U_i) \quad (2-16)$$

O funcionamento do neurónio é realizado através de uma operação simples: o neurónio recebe sinais das ligações de entrada e calcula um novo valor de ativação que é enviado através da ligação de saída. O valor é o resultado do cálculo do valor (S) de ativação do neurónio usando a função de ativação, que tem como argumento de entrada o valor da soma pesada dos valores de entrada (Chester, 1993).

As RNAs surgem como uma boa escolha na resolução de problemas de classificação, segmentação e previsão, quando os resultados do modelo são mais importantes do que compreender o modelo em si e a forma como funciona (Santos e Azevedo, 2005).

2.4.4.5 *K-Nearest Neighbours*

O algoritmo K-NN (vizinhos mais próximos), ou vulgarmente denominada de K-NN, baseia as suas previsões na localização das k observações que estão mais próximas do item a ser predito (Figura 2-23).

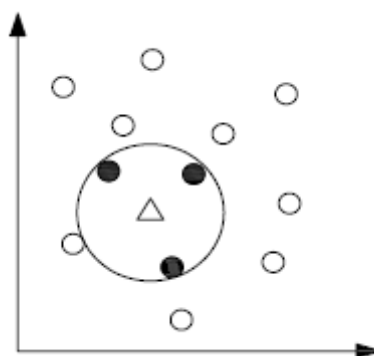


Figura 2-23 - Ilustração do princípio básico da técnica K-NN (Marques et al., 2009).

Considerando um dado conjunto D de tuplas de dados para treino, cada elemento de D consiste numa tupla $(x_1, x_2, \dots, x_n, c)$, onde c é a classe à qual pertence a tupla (x_1, \dots, x_n) .

Seja $Y = (y_1, \dots, y_n)$ uma nova tupla, ainda não classificada, para a sua classificação calcula-se as distâncias de Y a todas as tuplas de treino e considera-se as k tuplas de treino mais próximas a Y . De entre estas k tuplas, é verificada qual a classe que surge com mais frequência. A tupla Y será classificada dentro desta mesma classe.

A distância entre duas tuplas é calculada utilizando uma determinada medida de distância. As medidas de distância de uma maneira geral podem ser definidas como medidas de similaridade e de dissimilaridade. A primeira define o grau de semelhança entre as instâncias e realiza o agrupamento de acordo com a sua coesão e, a segunda, de acordo com as diferenças dos atributos das instâncias (Witten e Frank, 2005).

A noção de distância de similaridade mais comum consiste na distância euclidiana (2-17).

$$d(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2-17)$$

Normalmente é necessário normalizar os valores de cada atributo com o objetivo de que todos se situem num mesmo intervalo de variação, eliminando a elevada disparidade entre os valores dos diferentes atributos, o que poderia influir tendenciosamente no cálculo da distância.

O processo de normalização é simples. Considerando v um valor do atributo A que aparece na tupla Y , calcula-se o valor $v' \in [0; 1]$ da seguinte forma (2-18):

$$v' = \frac{v - \min_A}{\max_A - \min_A} \quad (2-18)$$

Onde \min_A e \max_A consistem nos valores mínimos e máximos que pode assumir o atributo A .

2.4.5 Abordagens para validação de resultados

Após a geração dos modelos, é necessário avaliar o desempenho dos mesmos. A principal dificuldade na previsão do desempenho esperado de um classificador num novo problema relaciona-se com quantidade limitada de dados disponíveis e o facto de a amostra poder não ser

suficientemente representativa. Portanto, efetuando uma simples divisão de dados em dados de treino e teste, gerando o modelo sobre o conjunto de dados de treino e avaliando do seu desempenho nos dados de teste pode tornar-se, na maioria das situações, insuficiente, uma vez que pode fornecer uma estimativa significativamente tendenciosa.

Na literatura surgem diferentes estratégias para validação de aprendizagem ou abordagens para avaliar a performance de um modelo, Estatística simples, a validação com a divisão da amostra (*hould-out validation*), a validação cruzada (*cross-validation*) e a *bootstrapping* (Létourneau et al., 1999). Todas envolvem a média do desempenho ao longo de várias iterações treino-teste. A ideia principal relativa à predição esperada do desempenho de um modelo consiste em efetuar “algo” na nova amostra de casos e que não foi "visto" pelo processo de aprendizagem efetuado durante a fase de treino. Isso ocorre porque as estimativas realizadas no conjunto de treino fornecem valores demasiado otimistas.

Na *hould-out validation*, a base de dados é dividida em duas novas bases de dados, uma de treino e outra de teste ou validação. Neste tipo de abordagem o conjunto de treino é utilizado para a construção de modelos e depois verificar a sua performance na classificação de ocorrências existentes no grupo de teste. Esta estimativa é baseada na performance esperada por parte modelo, gerado em novos casos. Este tipo de validação reduz de forma considerável o número de instâncias de forma a construir um modelo. Não sendo, assim, adequado para aplicações em pequenos conjuntos de dados. A *hould-out validation* assume que as ocorrências são idenpendentes umas das outras. Como pontos fortes temos a sua simplicidade e rapidez, embora produza uma redução efetiva dos dados disponíveis para treino.

A *cross-validation* utiliza os dados disponíveis de forma mais eficiente que a *hould-out validation*. Começa por dividir aleatoriamente os dados em k subconjuntos de tamanho aproximadamente igual. O modelo é treinado com $k-1$ subconjuntos e posteriormente é testado com o subconjunto remanescente. O processo é repetido até terem sido utilizados todos os subconjuntos na fase de teste. A performance esperada é calculada com base nos diferentes resultados obtidos em todas as repetições. A *cross-validation* requer mais tempo de computação do que a *hould-out validation*, no entanto fornece uma estimativa de desempenho mais robusta (Cortez, 2002).

A validação por *bootstrapping* é ainda mais exigente a nível de tempo de computação que a validação por *cross-validation*. Na sua forma mais simples, as análises (treino e teste) são

efetuadas efetuando aleatoriamente as trocas dos subconjuntos. São geralmente necessárias milhares de repetições.

Existem várias métricas ou medidas de avaliação de teste aos modelos que permitem saber o interesse de cada modelo e avaliar a sua performance na classificação dos dados (Kohavi e Provost, 1998).

A matriz de confusão consiste numa tabela onde são expressos os resultados obtidos sendo uma ferramenta comumente utilizada de forma a avaliar a performance de um modelo. Cada linha da matriz representa as instâncias previstas de uma classe enquanto cada coluna da matriz representa as instâncias reais de uma classe. Um dos benefícios da matriz de confusão está relacionado com a sua simplicidade de análise, principalmente se o sistema prevê duas classes. No caso de sistemas que preveem mais do que duas classes, estas podem ser reduzidas a duas, a classe alvo, designada de classe positiva, e as restantes agrupadas para formar uma só classe, designada de classe negativa. Esta estratégia permite comparar, mais facilmente, as previsões da classe positiva com as restantes classes, sendo possível construir uma tabela para cada classe. Assim sendo, a matriz de confusão consiste numa 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)) (Figura 2-24) (Zhang e Su, 2012).

		Valor Atual		Total
		<i>p</i>	<i>n</i>	
Resultado da predição	<i>p'</i>	<i>True Positive</i>	<i>False Positive</i>	<i>P'</i>
	<i>n'</i>	<i>False Negative</i>	<i>True Negative</i>	<i>N'</i>
Total		<i>P</i>	<i>N</i>	

TN – Instâncias negativas classificadas como negativas
 FP – Instâncias negativas classificadas como positivas
 FN – Instâncias positivas classificadas como negativas
 TP – Instâncias positivas classificadas como positivas

Figura 2-24 - Matriz de confusão. Adaptado de Han e Kamber (2011).

A matriz de confusão reflete a quantidade de classificações corretas e erradas sobre um conjunto de exemplos T . O número de acertos, para cada classe, é assinalado na diagonal principal da matriz $M(C_i, C_j)$, estes acertos podem ser verdadeiros positivos (correspondem ao número de exemplos positivos classificados como tal, também designados por TP – *true positive*) e verdadeiros negativos (correspondem ao número de exemplos negativos classificados como tal, também designados por TN – *true negative*). Os restantes elementos $M(C_i, C_j)$, para $i \neq j$, representam erros na classificação que podem ser falsos positivos (correspondem ao número de exemplos positivos classificados como negativos, também designados por FP – *false positive*) ou falsos negativos (correspondem ao número de exemplos negativos classificados como positivos, também designados por FN – *false negative*). A matriz de confusão ideal possui todos os elementos classificadores a nulo, representando a inexistência de erros (Pinto, 2005).

Tendo como base aos valores presentes na matriz de confusão é possível derivar um leque de métricas para avaliação de resultados, tais como:

- Exatidão

Segundo Zhang e Su (2012), a Exatidão (*accuracy*) consiste no acerto do sistema considerando a proporção de instâncias corretamente classificadas no total dos registos. Esta medida é calculada pela expressão (2-19):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2-19)$$

- Sensibilidade, Precisão e Medida F1

Segundo Zhang e Su (2012), Sensibilidade e Precisão (*Recall* e *Precision*) consistem em duas medidas de avaliação amplamente utilizadas. *Recall* pode ser considerada uma medida de completude enquanto *Precision* pode ser considerada como uma medida de precisão ou de fidelidade. *Recall* mede a capacidade de um sistema encontrar o que se quer e *Precision* mede a capacidade do mesmo sistema em rejeitar o que não se quer. Pode-se obter estas medidas a partir da matriz de confusão, equações (2-20) e (2-21).

$$Recall = \frac{TP}{FN + TP} \quad (2-20)$$

$$Precision = \frac{TP}{FP + TP} \quad (2-21)$$

F1-Measure é definida como a média harmónica entre os valores de *Recall* e do *Precision*. É expressa pela seguinte equação (2-22):

$$F1 - Measure = \frac{2 * Recall * Precision}{Recall + Precision} \quad (2-22)$$

A média harmónica é mais intuitiva do que a média aritmética quando envolve o cálculo da média de rácios. Esta medida é apropriada para situações onde é desejada a média das taxas ou percentagens.

- Falso Positivo e Verdadeiro Positivo

Segundo Delen et al. (2005), o Falso Positivo (*False Positive Rate* – FPR) é um indicador que permite saber qual a percentagem de instâncias que são erradamente classificadas como positivas, isto é, são erradamente classificadas como pertencentes à classe alvo. Pode ser obtido através da expressão (2-23):

$$FPR = \frac{FP}{FP + TN} \quad (2-23)$$

O indicador que permite saber qual a percentagem de instâncias negativas classificadas corretamente é o TNR – *True Negative Rate*. É definido através da expressão (2-24):

$$TNR = \frac{TN}{FP + TN} \quad (2-24)$$

Este indicador é conhecido como o indicador da “especificidade”. O indicador FPR pode também ser obtido através da expressão (2-25):

$$FPR = 1 - TNR \quad (2-25)$$

O Verdadeiro Positivo (*True Positive Rate*) é um indicador que permite saber qual a percentagem de instâncias classificadas corretamente como positivas. Este indicador é o mesmo que o *Recall*. Podemos obtê-lo através da expressão (2-26):

$$TPR = \frac{TP}{FN + TP} \quad (2-26)$$

Este é o indicador que representa a “sensibilidade” do sistema.

2.5 Algumas aplicações do *data mining*

Na literatura existem muitas abordagens que focam o uso de conceitos de DM na otimização de atividades de fabricação (Harding et al., 2006). Não é tão comum encontrar abordagens que usem estes conceitos de forma a melhorar a capacidade de prever comportamentos baseados em dados históricos. Na realidade, quanto à existência de possibilidade de uma colaboração distribuída de empresas independentes, com partilha de dados entre elas, mesmo que sejam concorrentes, os exemplos são ainda mais reduzidos. A utilização de bases de dados conjuntamente com abordagens estatísticas consiste numa ferramenta bem enraizada no âmbito da engenharia. A atividade de manutenção, bem como todas as questões inerentes, foram influenciadas e evoluíram baseadas na informação e nos desenvolvimentos de canais de comunicação (ICT - *Information and Communication Technology*).

Na literatura surgem alguns trabalhos que podem ser considerados como referências no estudo e desenvolvimento de soluções de aplicação do DM em diversas áreas da engenharia industrial, tais como a área de sistemas de produção, de sistemas de apoio à decisão e melhoria da qualidade. No entanto, a sua aplicação no campo da deteção de falhas na manutenção industrial iniciou-se na década de 90 (Harding et al., 2006).

Devido à quantidade de referências existentes na literatura, não é possível retratar muitos dos trabalhos elaborados, cujos resultados apresentam elevada relevância. No entanto, de seguida, são apresentados aqueles que apresentam similaridades com o presente estudo.

Batanov et al. (1993), apresentam um protótipo de um sistema denominado de EXPERT-MM. O sistema é constituído por uma base de dados contendo registos relativos a ocorrências de falhas em diferentes máquinas, bem como dados relativos ao seu comportamento na altura da

ocorrência. Este protótipo trabalha sobre dados históricos devolvendo como *output* sugestões para um planeamento apropriado de ações de manutenção preventiva.

O desenho de uma base de dados para aplicação de métodos otimizados de manutenção é apresentado por Hsu e Kuo (1995). Os autores sugerem que 100% das ações de inspeção devem ter início após a produção de n peças e quando a percentagem de peças defeituosas atinge um determinado valor limiar. Ações de manutenção preventiva deverão ser levadas a cabo por forma controlar de novo todo o processo. Quando o processo se encontra novamente controlado e são elaboradas novamente n peças, todo o procedimento poderá ser repetido.

Outro trabalho, este elaborado por Sylvain et al. (1999), utiliza diferentes algoritmos de DM, tais como AD, regressão, IR e RNA de forma a prever a falha de um determinado componente existente num avião, utilizando dados adquiridos através da utilização de sensores. Os resultados obtidos levaram à elaboração de políticas de manutenção preventiva capazes de atuar antes da falha de qualquer componente monitorizado.

Romanowski e Nagi (2001) aplicaram o DM ao domínio da manutenção de forma a identificar os subsistemas responsáveis pela baixa disponibilidade de determinados equipamentos. Os autores recomendam a utilização de planos de manutenção preventiva. A maioria dos dados relativos às falhas desses mesmos equipamentos foram obtidos recorrendo a sensores. Neste estudo foram utilizadas AD para a elaboração do modelo a aplicar.

Zhang e Kusiak (2012) apresentam um estudo relacionado com a manutenção em aerogeradores utilizando um sistema de supervisão e aquisição de dados denominado de SCADA (*supervisory control and data acquisition*). Foram desenvolvidos modelos de monitorização utilizando dados relativos a vibrações recolhidos em diferentes partes de um conjunto de aerogeradores. Utilizaram modelos de monitorização baseados em *clustering*. Estes modelos permitem identificar dados de vibração anormais. Neste estudo foi possível determinar valores limites (máximo e mínimo) de forma a detetar anomalias nas vibrações dos aerogeradores. Estes modelos são capazes de detetar anomalias, bem como determinar o início da vibração anormal.

Reffat et al. (2004) apresentam uma investigação relacionada com as potencialidades da aplicação de técnicas de DM sobre dados referentes a atividades de manutenção de edifícios, de forma a identificar os impedimentos para um melhor desempenho dos ativos usados na sua construção. Os benefícios para a indústria da construção assenta na obtenção de novo conhecimento existente em bases de dados de forma a melhorar a eficiência dos processos de

manutenção e assim projetar futuros planos de manutenção incorporando o novo conhecimento gerado.

Bansal et al. (2004), descrevem a criação de um sistema de manutenção preditiva baseado na aplicação de redes neuronais. Verificou-se a capacidade das redes neuronais aprenderem funções de mapeamento não-linear para previsão de parâmetros de sistema de uma máquina. Com esta aproximação é evitada a custosa necessidade de medição de parâmetros de sistema. Este sistema foi unicamente validado utilizando um sistema virtual, sendo necessário implementá-lo num sistema produtivo real.

Um outro trabalho destes mesmos autores, Bansal et al. (2006), apresenta uma alteração em relação ao trabalho anterior. Os autores verificaram que a performance relativa à utilização de redes neuronais para aprendizagem de funções de mapeamento não-linear para previsão de parâmetros de sistema de uma máquina, depende em larga escala da qualidade dos dados de treino do respetivo modelo. Desta forma, neste trabalho, os autores apresentam a utilização de um novo algoritmo denominado de BJEST (*Bansal-Jones Estimation*). Este trabalho demonstrou uma melhoria significativa da utilização deste algoritmo relativamente à consistência e precisão dos resultados obtidos.

Fu et al. (2004) apresentam um sistema denominado de ICMMS (*intelligent-control-maintenance-management system*) em unidades de geração hidroelétricas, utilizando políticas de manutenção preventiva condicionada ou preditiva. O sistema faz uso de toda a informação de controlo, manutenção e de aspetos técnicos de gestão de forma a efetuar ações de manutenção no tempo e local corretos. É gerado e aplicado um modelo de RNA de forma a implementar ações de manutenção preditiva garantindo uma performance ideal de todo o sistema.

Por fim, Cai et al. (2011) apresentam um estudo onde são utilizadas RB para identificação de falhas de um produto operando sobre distintas configurações e condições. Neste estudo são verificadas as relações existentes entre as diferentes variáveis analisadas e capazes de devolver a variável alvo de estudo. Através de um estudo de caso os autores confrontam a utilização das RB em relação aos modelos tradicionais para criação de AD, demonstrando com os resultados obtidos a melhor performance na utilização das RB.

Capítulo 3

3. O sistema de predição

Na literatura muitos são os trabalhos que se concentram na otimização de ações de manutenção de forma a melhorar o desempenho do planeamento dessas mesmas ações. Não nos esqueçamos porém que após todo o processo de implementação e otimização de ações de manutenção é necessário encetar um conjunto de novas ações de previsão de falha de forma a melhorar e otimizar as ações tomadas por parte das equipas de manutenção. É nesse campo de atuação que este trabalho pretende incidir, contribuindo para a melhor identificação de uma futura possível avaria de um equipamento, podendo desta forma atuar com medidas preventivas baseadas em dados para a tomada de decisão e assim contribuir para uma otimização do planeamento de ações de manutenção.

O objetivo final deste capítulo é a especificação completa do funcionamento do sistema de predição designado por SPAMUF, sistema de predição de avarias em máquinas de unidades fabris globalmente dispersas.

Numa primeira parte dão-se a conhecer quais os pressupostos assumidos no desenvolvimento do sistema, realizando-se também uma especificação da estrutura funcional do mesmo. São focadas superficialmente várias questões pertinentes, com o objetivo de alertar o leitor para os vários temas marginais ao alvo de estudo desta tese. Este procedimento baseia-se na necessidade de se ter consciência de problemas colaterais que poderão ser impeditivos de um desempenho adequado do sistema. Em simultâneo, esta ação pretende diminuir o número de fatores a controlar no decorrer da especificação do sistema proposto. Nesta fase de desenvolvimento, considera-se irreal a tentativa de controlar todas as ocorrências que poderão surgir no decorrer da ação de operação. Assim, com o procedimento referido pretende-se delimitar e focalizar a área de intervenção deste trabalho. A operacionalidade do modelo e as conclusões retiradas na parte final desta tese, apenas se manterão válidas, enquanto as condições estabelecidas pelos pressupostos se mantiverem inalteradas. O mesmo raciocínio é válido em relação às regras de funcionamento especificadas para cada uma das várias entidades intervenientes. Na secção de trabalhos futuros serão apontados caminhos para futuras investigações que poderão passar, precisamente, pelo levantamento de algumas das restrições impostas neste estudo.

Neste capítulo é explicitado o funcionamento do sistema proposto com base em diagramas IDEF0. São definidas as características do sistema, a função individual de cada atividade

constituente com a explicação da sua contribuição para o sistema global e o relacionamento que tem com as demais atividades do sistema. Será dado realce às atividades consideradas de maior importância no funcionamento do sistema, mediante a sua caracterização exaustiva, não deixando de explicar todas as restantes. Toda a explicação relativa ao funcionamento do sistema proposto baseia-se nos diagramas IDEF0 gerados. Inicia-se a explanação pelo processo de criação e definição de objetivos e estrutura funcional do sistema, seguida da seleção, incorporação e rejeição de parceiros, isto é, de empresas inseridas no sistema que com isto beneficiam das potencialidades do mesmo. Em seguida entra-se na parte nuclear do trabalho, que é a especificação da fase de operação. É neste ponto que se descrevem as ações realizadas, no sentido de planejar a execução do processo preditivo recorrendo à seleção da técnica de DM mais adequada.

Este capítulo termina com a apresentação das ferramentas e tecnologias utilizadas e necessárias na implementação de todo o sistema.

3.1 Visão global

Existem diferentes técnicas e métodos para auxiliar na implementação da manutenção preditiva, isto é, para prever futuras falhas de um equipamento ou parte dele ou para prever a degradação da sua *performance* partindo da sua condição atual e da sua história funcional (Fornaro et al., 2004). Surge como característica comum de todas estas técnicas a necessidade de existência de uma base de dados histórica contendo toda a informação de interesse relativa aos equipamentos existentes nas diferentes linhas de produção, bem como todos os seus dados de monitorização (Efthymiou et al., 2012). Assim sendo, a disponibilidade de bases de dados contendo informações de interesse fornecidas através de dispositivos existentes em diferentes linhas de produção e em diferentes ambientes produtivos, consiste num elemento de fundamental importância para o funcionamento eficaz desta ferramenta de diagnóstico. Em particular, esta informação é necessária para a execução eficiente de técnicas de manutenção preditiva.

A Figura 3-1 representa a estrutura lógica do sistema SPAMUF, dividida em três diferentes níveis.

Nível 1: Alarme.

A este nível, dados de monitorização contínua são periodicamente medidos e registados através de agentes automáticos de recolha de dados relativamente aos equipamentos existentes nas diferentes linhas de produção, podendo ser controlados diferentes componentes de um mesmo

equipamento. Quando um equipamento/componente falha, são recolhidos dados e um alarme é gerado e difundido à equipa ou responsável de manutenção. Um sinal de falha é transmitido em conjunto com um valor identificado/monitorizado e com a identificação do respetivo equipamento.

Nível 2: Ações de manutenção.

Um agente automático de recolha de dados instalado no equipamento recolhe dados relativos à intervenção efetuada no equipamento em falha, de forma a devolver informação útil relativa ao modo de falha e trabalho efetuado. De igual modo são registados dados relativos a intervenções preventivas efetuadas no mesmo equipamento registando toda a informação relevante acerca da ação realizada. Dados relativos à intervenção no equipamento são transmitidos via infraestrutura de dados existente e armazenados numa base de dados histórica para utilização no nível seguinte.

Nível 3: Manutenção preditiva.

Algoritmos e técnicas de DM são aplicados tendo como base os dados históricos de intervenções em conjunto com uma base de dados de monitorização.

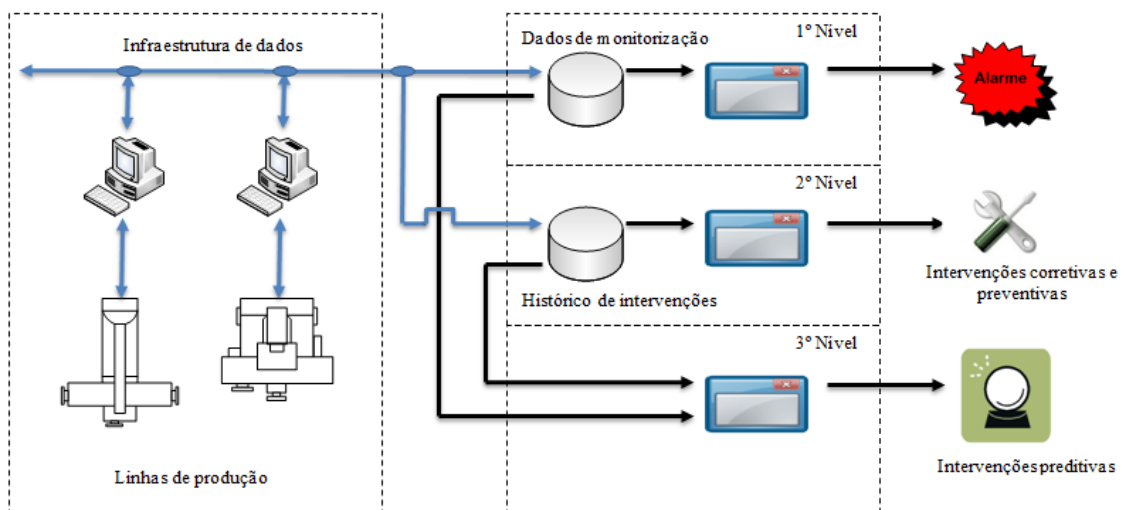


Figura 3-1 - Relações entre processos, dados e intervenções.

Com este trabalho não é pretendido devolver informação detalhada acerca de atividades de manutenção corretivas ou preventivas. O objetivo deste trabalho consiste em prever uma falha antes da sua ocorrência, alertando o responsável pela manutenção da necessidade de intervir sobre o equipamento antes de a falha ocorrer, podendo desta forma serem ajustados os planos de ações preventivas já existentes. Em suma, este trabalho pretende operacionalizar o 3º nível apresentado na figura anterior.

O sistema que se pretende desenvolver seguirá a estrutura proposta por Ruiz et al. (2014) no que diz respeito às fases do processo de DCBD (ver Figura 3-2) prevendo a forma como o sistema

gerado poderá afetar os processos de gestão de manutenção de um qualquer parceiro integrante deste mesmo sistema tendo em consideração o planeamento de ações de manutenção preventivas e todas as ações corretivas levadas a cabo independentemente de serem previstas ou não pelo sistema.

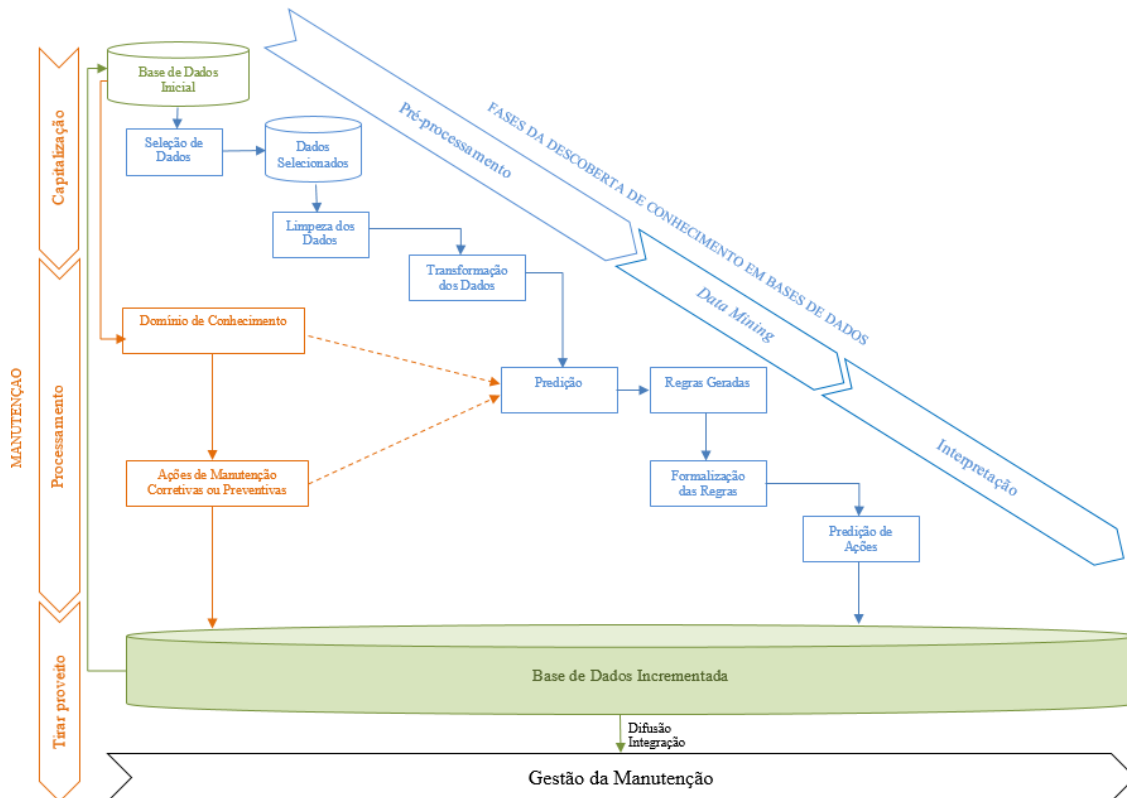


Figura 3-2 - Influência do sistema na gestão da manutenção. Adaptado de Ruiz et al. (2014).

A figura apresenta um sistema iterativo, onde a DCBD em consonância com as ações de manutenção programadas e não programadas, levadas a cabo por parte de um ou mais parceiros, podem afetar todo o processo de gestão da manutenção existente e executado por um determinado parceiro, podendo desta forma acarretar uma reorganização desse mesmo processo de gestão. A figura é demonstrativa de todo o procedimento que se inicia na criação de uma base de dados inicial e agregadora de dados de intervenções corretivas e preventivas, bem como de dados de monitorização. Todo o processo de DCBD será levado a cabo até serem gerados modelos de predição, através dos quais são efetuadas previsões de ações preventivas condicionadas a fim de evitar uma falha de um determinado equipamento que possivelmente será traduzida numa avaria. No processo preditivo deverá ser tomado em consideração todo o conhecimento adquirido por parte das equipas de manutenção e consequentes ações efetuadas e não previstas pelo sistema. Todos estes registos de ações são incrementados na base de dados que servirá de base

a todas as iterações do sistema. Tal como dito anteriormente, a difusão e integração do conhecimento adquirido com o sistema poderá afetar a gestão de processos de manutenção levados a cabo por um ou mais parceiros.

Numa visão mais generalista, o sistema agrega os dados iniciais num conjunto de bases de dados distribuídas (Figura 3-3) e pertencentes a cada organização. A angariação dos dados é efetuada a nível local e posteriormente são transferidos para um patamar superior e agregador de dados de cada uma das estruturas locais.

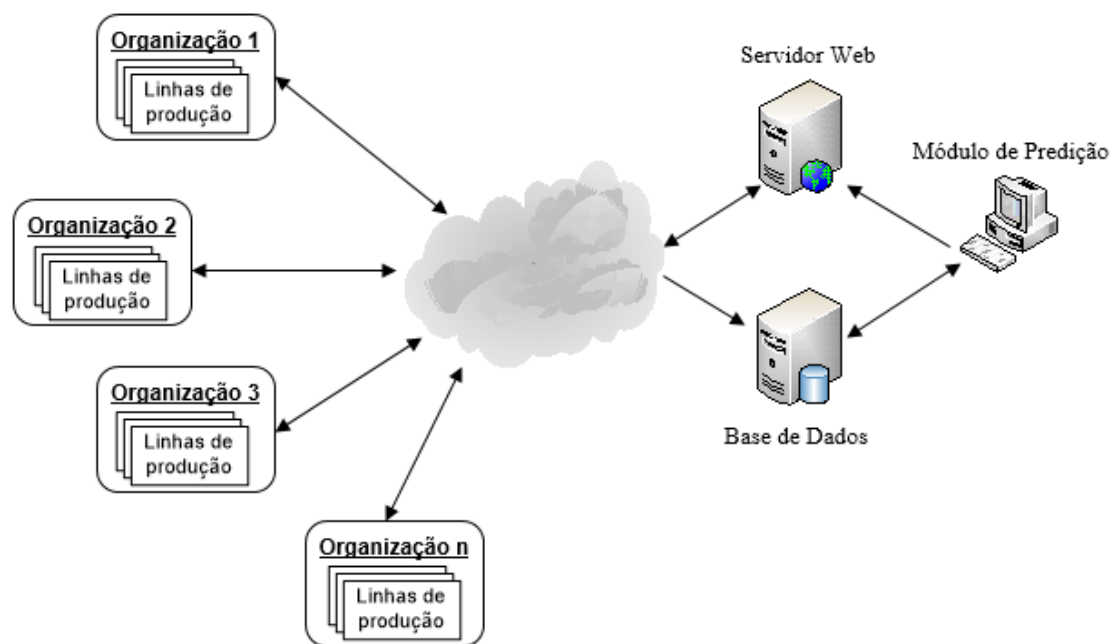


Figura 3-3 - Esquematização geral do projeto.

Sobre esse patamar de dados incide o módulo de predição elaborado com o intuito de descobrir conhecimento implícito e oculto de forma a gerar previsões de comportamento e possíveis futuras ocorrências de falhas. A eventualidade da ocorrência de eventos é disponibilizada pelo sistema de predição através de um portal *front-end* (Servidor *Web*), à totalidade das organizações, sob a forma de sistema de suporte à decisão.

3.2 Especificação funcional

De seguida e recorrendo à esquematização através de gráficos IDEF0, é realizada a descrição funcional do sistema proposto. Esta metodologia tem como objetivo a modelação de processos para um desenvolvimento assente e sustentado, descrevendo todo o ciclo de vida de

desenvolvimento de um sistema de forma gráfica. A metodologia IDEF0 pertence a uma família de linguagens de modelação na área da engenharia de sistemas e *software*, a família IDEF. Cobre uma ampla gama de aplicações, desde modelos funcionais de sistemas a descrição de dados, simulação, análise/desenho orientada a objetos e aquisição de conhecimento.

A metodologia mais reconhecida e utilizada da família IDEF é a metodologia IDEF0, uma linguagem de modelação funcional baseada em análise estruturada e técnicas de esquematização (SADT¹⁰) idealizada de forma a modelar as decisões, ações e atividades de uma organização ou sistema (Marca e McGowan, 1987).

3.2.1 A Metodologia IDEF0

Uma vez que o esquema concetual do sistema apresentado nesta tese foi realizado usando a metodologia IDEF0, pretende-se nesta breve subsecção explicar as motivações que levaram à utilização desta metodologia. Para além disso, são fornecidos elementos sobre as regras que regulam o uso desta ferramenta e metodologia de leitura de um esquema IDEF0.

A *Integration Definition for Function Modeling* (IDEF) é uma família integrada de métodos de modelação com aplicabilidade em vários domínios, indo desde o setor industrial até à área comercial (Browne, 1996; Waltman e Presley, 1993). Em Dezembro de 1993 o *Computer Systems Laboratory do National Institute of Standards and Technology* (NIST) normalizou a metodologia IDEF0 através da publicação n.º 183 do *Federal Information Processing Standards Publications* (FIPS PUBS). A metodologia IDEF0 emergiu da técnica SADT e tem uma grande utilização nos dias que correm (Pires, 2004; Serifi et al., 2009; Oliveira e Rosa, 2010; Bevilacqua et al., 2012; Shang et al., 2011; Shehab et al., 2013; Maull et al., 2013; Grabenstetter e Usher, 2013).

A metodologia IDEF0 é utilizada de forma a garantir uma completa visão dos conteúdos de informação requeridos (Cho e Lee, 1999). É uma ferramenta de análise de processos de qualquer natureza, mediante a sequência de diagramas inter-relacionados logicamente, iniciados pela função macro do sistema apresentado. Este diagrama é desdobrado em outros diagramas mais detalhados a fim de completar o objetivo do mapeamento.

Os objectivos primários da norma acima referida (publicação n.º183) são (FIPS-PUBS, 1993):

¹⁰ *Structured Analysis and Design Technique*. Consiste numa metodologia na área da engenharia de sistemas e engenharia de software de forma a descrever sistemas através de uma hierarquia de funções. Consiste numa notação esquemática projetada especificamente para o auxílio na descrição e compreensão de sistemas.

1. Documentar e clarificar a técnica de modelação IDEF0 e especificar a sua correta utilização;
2. Fornecer meios para uma modelação completa e consistente das funções de um sistema ou área de estudo, bem como definir os dados e objetos que relacionam essas funções;
3. Fornecer uma linguagem de modelação que seja independente dos métodos ou ferramentas a utilizar;
4. Fornecer uma linguagem de modelação com as seguintes características:
 - a) Genérica (permite analisar sistemas e áreas temáticas de grande variedade, objetivo e complexidade);
 - b) Rigorosa e precisa (permite produzir modelos correctos e utilizáveis);
 - c) Conceptual (Para representação de requisitos funcionais independentemente da sua implementação física ou organizacional);
 - d) Flexível (uma vez que permite suportar diversas fases do ciclo de vida de um projeto).

3.2.1.1 Justificação da utilização da metodologia IDEF0

A utilização desta metodologia de normalização é fortemente recomendada em projetos que requeiram uma técnica de modelação para análise, desenvolvimento, reengenharia ou integração de sistemas de informação (Liu e Fang, 2006). Uma vez que nesta tese (subsecção 3.2.2) se realiza a especificação de um sistema de predição de futuras falhas e considerando também os objetivos da metodologia IDEF0, crê-se estarem reunidas as condições indicadas para a utilização desta ferramenta. Acrescenta-se ainda como fator de incentivo ao seu uso as características da linguagem IDEF0 definidas na norma (FIPS-PUBS, 1993):

- É compreensiva e expressiva, capaz de representar graficamente uma grande variedade de negócios, produção e outros tipos de operações de empresas em qualquer nível de detalhe.
- É coerente e simples, fornecendo expressões rigorosas e precisas e promovendo consistência no uso e interpretação.
- Melhora a comunicação entre analistas de sistemas, programadores e utilizadores.
- Está suficientemente provada e testada através de uma utilização de vários anos da Força Área dos EUA e pela indústria privada.
- Pode ser gerada por uma grande variedade de ferramentas de software existentes no mercado.

Mais pormenores referentes a esta técnica poderão ser encontrados no livro *The Practical Guide to Business Process Reengineering Using Idef0* (Feldmann e Tieso, 1998) e no texto da *Federal Information Processing Standards Publications* (FIPS-PUBS, 1993).

3.2.1.2 Sintaxe e semântica

Existem cinco elementos dentro do esquema funcional IDEF0 (Figura 3-4). São eles (Chidambaram et al., 1999):

- Atividade, designada também por função, ação, processo ou operação, representada por caixas;
- Entradas, que são representadas por setas que fluem para o lado esquerdo da caixa que representa a atividade;
- Saídas, que são representadas por setas que saem pelo lado direito da atividade;
- Controlos representados pelas setas que entram na parte de cima da atividade e cuja função é a de controlar ou limitar o funcionamento da ação;
- Mecanismos que consistem nas setas que fluem para a parte inferior da actividade e podem ser vistos como elementos que permitem que a atividade se desenvolva.

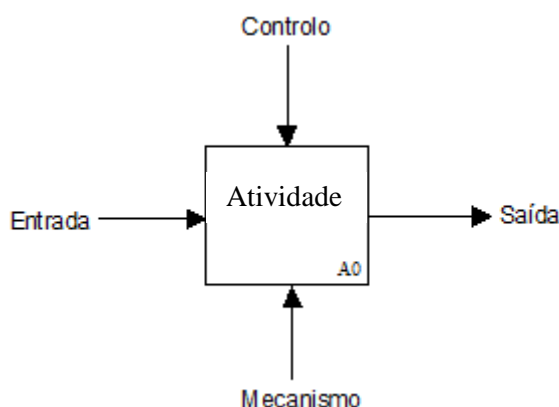


Figura 3-4 - Nomenclatura da técnica IDEF0.

Esta técnica baseia-se numa decomposição hierárquica, partindo do geral até ao específico (Pires, 2004). Uma qualquer atividade pode ser decomposta em subatividades e essas subatividades resultantes podem também ser subdivididas. Este procedimento mantém-se até ao nível de detalhe pretendido ou razoável. O diagrama de nível hierárquico mais elevado é denominado diagrama de contexto ou diagrama A-0 e representa-se por uma única caixa onde se resumem todas as funções. O diagrama A0 representa a primeira decomposição do sistema. Para

além do diagrama A-0, todos os restantes devem ter entre 3 a 6 atividades numeradas. Esta limitação deve-se ao seguinte: se uma atividade é dividida em mais de seis subatividades, então está excessivamente pormenorizada neste nível tornando a sua visualização incompreensível. Por este motivo, deve-se considerar outro nível hierarquicamente inferior onde se reunirão algumas das actividades presentes. Se pelo contrário existirem menos de três sub-actividades, então com certeza que se estão a representar poucos detalhes, motivo pelo qual se deve reunir essas actividades apenas numa, hierarquicamente superior.

A localização das caixas num determinado diagrama bem como as setas que as interligam, não implicam uma sequência temporal. Iterações, retorno de informação e processos contínuos podem ser representados por setas. A saída de uma determinada atividade pode ser utilizada numa atividade que a precede no esquema, com o papel de a ativar.

Na Figura 3-5 apresenta-se um modelo hipotético onde se podem ver os diagramas A-0, A2 e A23. Por exemplo se em A0 existirem quatro atividades, estas serão designadas por A1, A2, A3 e A4. Todas as subatividades de A2 devem seguir a designação A2x (com x a variar entre 1 e 6). No modelo apresentado pela figura será A21, A22 e A23.

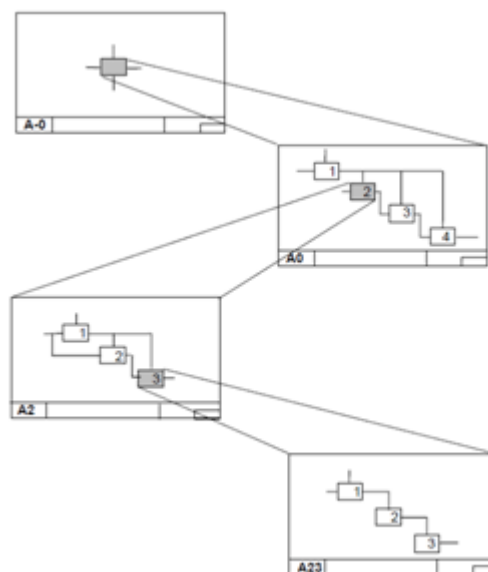


Figura 3-5 - Decomposição pelo método IDEF0. Adaptado de Waltman e Presley (1993).

O exemplo da Figura 3-5, que se pretendeu o mais simples possível por propósitos introdutórios, inclui já alguns detalhes que se considera importante explicar. Para se descrever de uma forma adequada um diagrama realizado em IDEF0, é necessária uma maior atenção às setas

que realizam a interligação entre as atividades. As setas podem por exemplo fornecer dados a mais do que uma atividade (circunferências cinza na Figura 3-6).

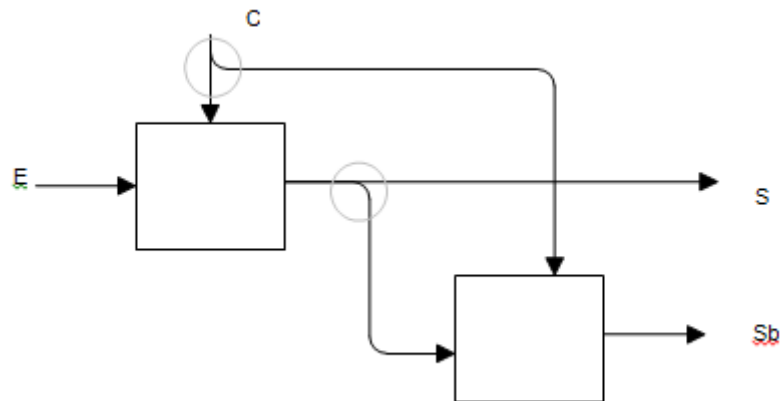


Figura 3-6 - Ramificação de setas.

Por outro lado, diferentes setas também se podem unir numa só, significando neste caso que diferentes atividades podem produzir saídas do mesmo tipo (circunferência cinza na Figura 3-7).

Há que fazer menção à possibilidade de entradas, saídas, controlos e mecanismos múltiplos.

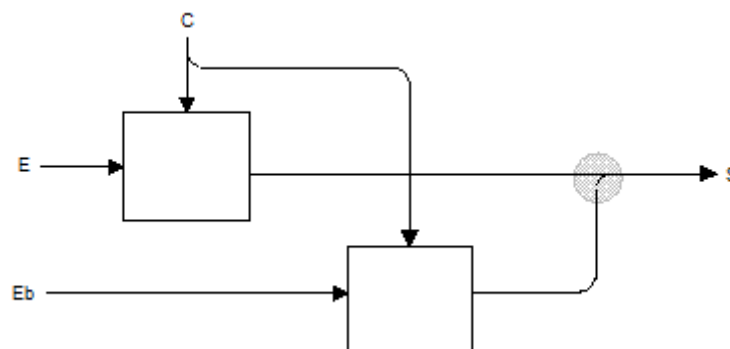


Figura 3-7 - Junção de setas.

Na Figura 3-8 é possível ver-se a repercussão de entradas, saídas, controlos e mecanismos múltiplos de um diagrama A-0 no diagrama seguinte hierarquicamente inferior. Na imagem apresenta-se um exemplo hipotético onde é possível visualizar a utilização de junções e ramificações de setas, bem como setas que, saindo de uma atividade, são utilizadas como controlo noutra atividade (linha cinza clara designada por AA). É possível também ver-se uma seta que, saindo de uma atividade, é utilizada como entrada noutra atividade (linha cinza clara designada por BB). O facto de a linha AA fazer o percurso por cima e a linha BB por baixo é fruto do seguimento da norma.

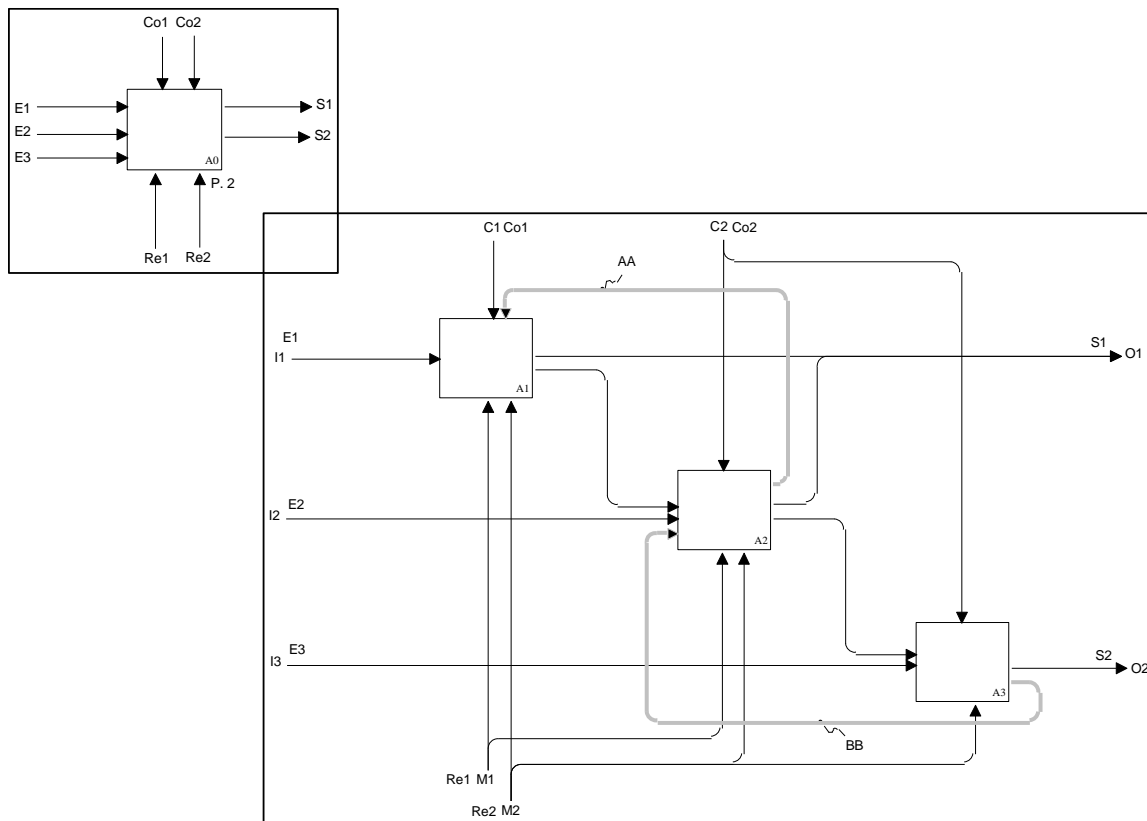


Figura 3-8 - Entradas, saídas, controles e mecanismos múltiplos.

Na especificação completa de um determinado sistema, a metodologia IDEF recorre a três elementos que lhe conferem consistência, sistematização e coerência. Trata-se concretamente dos esquemas ou diagramas IDEF0, do texto explicativo e de um glossário. Os diagramas são realizados conforme indicado nos parágrafos anteriores. O texto associado aos diagramas pretende apresentar uma explicação resumida do diagrama, dando relevo a determinadas características fundamentais deste. O glossário é utilizado para definir determinados rótulos associados a setas e permitir desta forma uma interpretação correta do diagrama.

3.2.2 Descrição funcional

Nesta subsecção é apresentada a esquematização funcional do sistema através da utilização da metodologia anteriormente apresentada. Aplicando esta metodologia é possível efetuar uma abordagem *Top-Down* de todo o sistema sendo apresentadas e destrinchadas todas as atividades e subatividades que o constituem.

Na sua representação mais geral, o funcionamento global do Sistema de Predição de Avarias em Máquinas de Unidades Fabris Globalmente Dispersas é representado pelo diagrama IDEF0 da Figura 3-9.

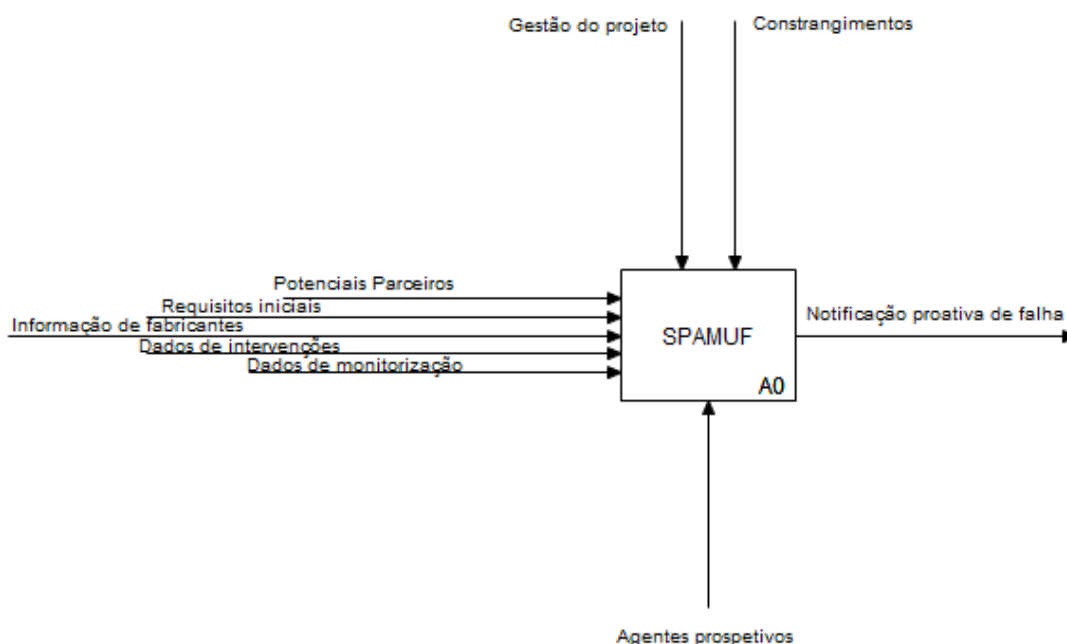


Figura 3-9 - Funcionalidade global do sistema.

A figura mostra que as entradas principais para o funcionamento do sistema podem ser classificadas conforme a sua natureza, isto é, i) ao nível de gestão de parceiros é fundamental agregar diferentes organizações com interesse em integrar o sistema e desta forma obter resultados capazes de se traduzirem em melhorias no planeamento das suas ações de manutenção, utilizando para tal um conjunto de dados referentes a todos os parceiros integrantes do sistema. ii) ao nível funcional e de forma a garantir o funcionamento do sistema é fundamental estabelecer todos os requisitos iniciais relacionados com a esquematização funcional onde são estabelecidas todas as necessidades técnicas a serem utilizadas e assim garantir o seu normal funcionamento. iii) por fim é fundamental “alimentá-lo” com dados referentes a ações de manutenção e de monitorização dos diferentes equipamentos de cada parceiro, base de todo o processo preditivo e objetivo primordial do sistema. Dados relativos a informação de diferentes fabricantes poderão ser tidos em conta, visto normalmente conterem dados relevantes acerca de ações de manutenção preventiva dos diferentes equipamentos integrantes do sistema.

Após a descrição dos diferentes fluxos de entrada é fundamental neste momento focar o principal fluxo de saída do sistema, a notificação proativa de falha. Esta saída consiste no envio ao parceiro, ou proporcionar a visualização por parte do mesmo, de alertas relativas a possíveis falhas futuras nos equipamentos, podendo desta forma, as equipas de manutenção efetuar ações de manutenção preventivas condicionadas e assim evitar a ocorrência dessa mesma falha evitando assim a ocorrência de uma avaria num equipamento.

O significado dos fluxos de entrada e saída representados na Figura 3-9 encontram-se sistematizados na Tabela 3-1.

Tabela 3-1 - Fluxos na representação global do sistema.

Atividade	Entradas	Saídas
A0	<ul style="list-style-type: none"> • <u>Potenciais parceiros</u>: Informação referente aos potenciais parceiros a integrar no sistema. Informação relativa à área de atividade, caracterização de recursos, especificações funcionais relativas à área de manutenção e possíveis constrangimentos. • <u>Requisitos iniciais</u>: Estabelecimento de critérios relativos aos requisitos para criação do sistema e desenho funcional do sistema, estabelecendo as necessidades técnicas e funcionais do sistema de forma a garantir o seu normal funcionamento. • <u>Dados de Manutenção</u>: Dados relativos a ações de manutenção corretiva e preventiva a fornecer ao sistema relativos aos parceiros. • <u>Dados de Monitorização</u>: Dados relativos a ações monitorização a fornecer ao sistema relativos aos parceiros. • <u>Informação de fabricantes</u>: Informação relativa a ações corretivas e preventivas de fabricantes dos equipamentos inseridos no sistema e pertencentes aos diferentes parceiros. 	<ul style="list-style-type: none"> • <u>Notificação proativa de falha</u>: Objetivo principal do sistema que consiste no envio de alertas relativos a possíveis falhas, alertando da necessidade de elaboração de ações de manutenção preventivas condicionadas sobre os equipamentos aos parceiros integrantes do sistema.

Como se poderá verificar mais à frente nesta subsecção, na descrição detalhada de cada subatividade surgirão mais fluxos de entrada e de saída, bem como mecanismos e controlos, que se encontram omissos a este nível de representação. Assim sendo, quando forem focadas e explanadas as diferentes subatividades que compõem o sistema os novos fluxos serão focados e descritos de forma a tornar mais fácil a compreensão da esquematização apresentada.

A Tabela 3-2 inclui a lista dos mecanismos e informação de controlo representados na Figura 3-9.

Tabela 3-2 - Mecanismos e informação de controlo do sistema.

Controlos	<ul style="list-style-type: none"> • <u>Gestão do projeto</u>: Procedimentos de gestão definidos inicialmente de forma a criar e gerir o sistema e atingir os objetivos pretendidos. • <u>Constrangimentos</u>: Identificação de constrangimentos funcionais de forma a garantir o normal funcionamento do sistema, o tipo de competências que podem/devem ser implementadas para melhorar o desempenho global do sistema.
Mecanismos	<ul style="list-style-type: none"> • <u>Agentes prospetivos</u>: Agentes automáticos para recolha de dados pelo sistema.

Na atividade principal representada na Figura 3-10 existem três grandes atividades (A1,A2 e A3) que coincidem com as três fases principais identificadas, bem como todo um conjunto de fluxos de informação (setas) que inter-relacionam cada uma dessas fases e que permitem a sua operacionalidade.

Dada a complexidade que o sistema encerra é necessário recorrer-se à decomposição hierárquica de cada atividade. Segue-se a explicação de cada uma das subatividades repetindo-se este procedimento para cada uma das subatividades resultantes até se atingir o nível de explicação desejado.

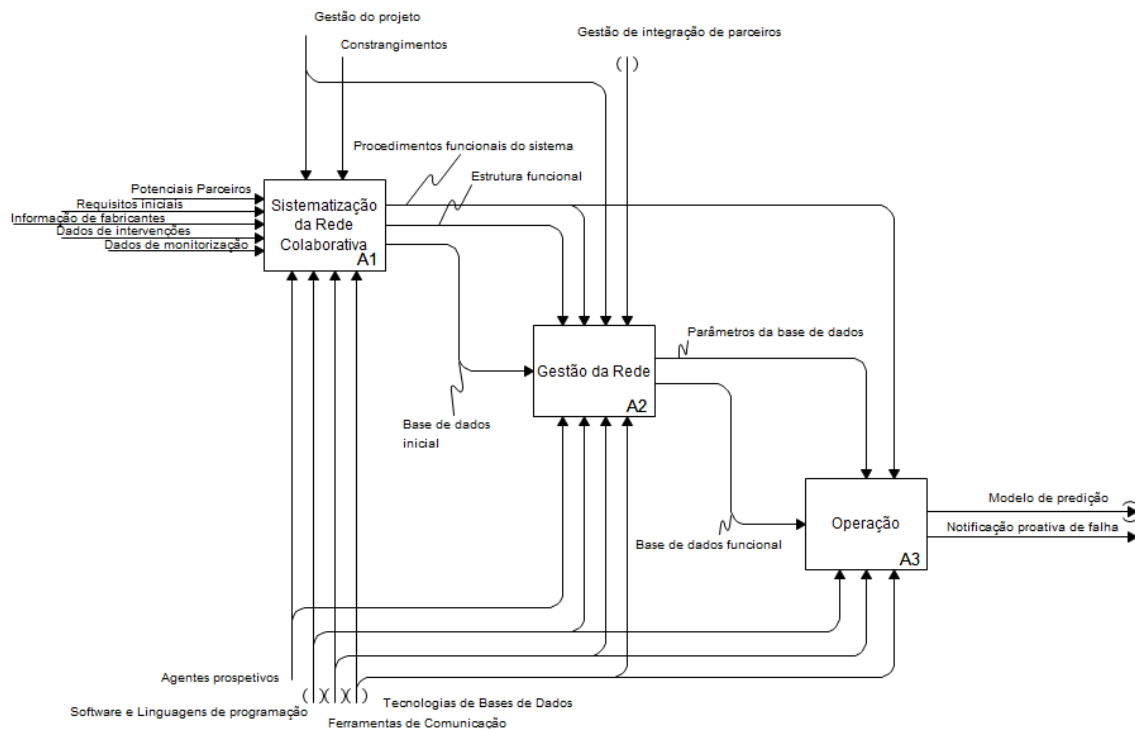


Figura 3-10 - Atividade A0 - SPAMUF.

Como se poderá verificar através da visualização da Figura 3-10, a atividade A1 corresponde à atividade responsável por inicializar o sistema recebendo como fluxos de entrada os descritos anteriormente e *inputs* da atividade A0 – SPAMUF. A atividade A1 gera como saídas os procedimentos funcionais do sistema e o estabelecimento de toda a estrutura funcional do mesmo, servindo estas saídas de controlo das atividades seguintes, A2 e A3 respetivamente. A atividade A1 gera igualmente a base de dados inicial do sistema que servirá de entrada a atividade A2, gestão da rede, atividade responsável por efetuar toda a gestão da base de dados inicial do sistema e efetuar todos os procedimentos de pré-processamento e limpeza de dados capazes de gerar uma base de dados funcional e alimentando a atividade A3, atividade esta onde é gerado todo o processo de predição do sistema. A atividade A2, gestão da rede, gera também como saída os parâmetros relativos à criação da base de dados funcional, controlando a atividade A3, atividade responsável por gerar as principais saídas do sistema, o modelo de predição gerado e a notificação proactiva de falha de um componente ou equipamento inserido no sistema.

Seguidamente é efetuada uma descrição mais detalhada das atividades apresentadas a este nível de representação.

Atividade A1 – Sistematização da Rede Colaborativa

Esta atividade corresponde à inicialização do sistema e gestão de parcerias para o sistema proposto, é feita a incorporação e seleção dos possíveis parceiros do sistema tendo em consideração os constrangimentos estabelecidos aquando da definição inicial do sistema relativamente a diversos aspetos tecnológicos tais como, a obrigatoriedade do uso de bases de dados normalizadas, a utilização de agentes automáticos para recolha de dados, o desenvolvimento e aplicação de *software* sem custos de licenciamento e a utilização de ferramentas de comunicação. Em suma, nesta atividade são definidos e refinados aspetos relacionados com a seleção e gestão dos parceiros integrantes do sistema e é constituída por dois subníveis a explanar mais à frente nesta subsecção.

Atividade A2 – Gestão da Rede

Esta atividade é constituída por três subactividades posteriormente apresentadas, responsáveis pela análise dos dados incorporados no sistema e pela criação agrupamentos de dados de forma a tornar a base de dados mais adequada ao processo de DM e operado na atividade seguinte.

Atividade A3 – Operação

Esta atividade está composta por 4 subníveis e relaciona-se com a gestão de dados a ser elaborada pelo sistema de forma a alimentar a função de DM e posterior disseminação de resultados. A atividade de DM está sujeita a solicitações efetuadas por parte de um ou mais parceiros, sendo gerados alertas especificamente direcionados a quem os solicita. Paralelamente o sistema é capaz de gerar alertas não solicitados e disseminados a todos os parceiros integrados no sistema. Tal como as atividades anteriores, no decorrer do capítulo serão explanadas e explicitadas todas as subatividades que a constituem.

O significado do novo fluxo de saída da atividade A0 representados na Figura 3-10 encontra-se sistematizado na Tabela 3-3. Salienta-se que os fluxos de saídas correspondem aos fluxos de saída da subatividade A3. Todos os fluxos de saída das subatividades A1 e A2 são fluxos internos da atividade A0 e serão descritos mais à frente nesta tese.

Tabela 3-3 - Fluxo de saída atividade A0.

Atividade	Saídas
A3	<ul style="list-style-type: none"> • <u>Modelo de predição</u>: Modelo de predição de conhecimento gerado através da aplicação do DM e que servem de base para a geração das notificações proativas de falha.

A Tabela 3-4 inclui a lista dos novos mecanismos e informação de controlo representados na Figura 3-10.

Tabela 3-4 - Mecanismos e informação de controlo da atividade A0.

Controlos	<ul style="list-style-type: none"> • <u>Gestão de integração de parceiros</u>: Procedimentos de gestão de parceiros de forma a auxiliar o processo de avaliação e validação dos cumprimentos dos requisitos funcionais dos parceiros, garantindo o normal funcionamento do sistema (Atividade A2).
Mecanismos	<ul style="list-style-type: none"> • <u>Software e linguagens de programação</u>: Necessidade de desenvolvimento e utilização de diferentes programas e linguagens de forma a implementar possíveis agentes automáticos de recolha de dados, o sistema de predição de conhecimento e a plataforma de disseminação de resultados. • <u>Ferramentas de comunicação</u>: Utilização de ferramentas comunicação para fluxo de dados entre os parceiros e o módulo de DM gerado, bem como para difusão dos resultados obtidos. • <u>Tecnologias de bases de dados</u>: Aplicação de conhecimentos a nível de implementação e manuseamento de ferramentas de base de dados de forma a gerar, manusear e manter a base de dados do sistema.

De seguida é apresentada e explicitada a atividade A1 – Sistematização da Rede Colaborativa focando todas as suas subactividades. A atividade é composta por três subatividades e dois subníveis: a subatividade A11 – Definição e Inicialização do Sistema, a subatividade A12 – Integração de Parceiros e a subatividade A13 – Configuração da Rede (Figura 3-11).

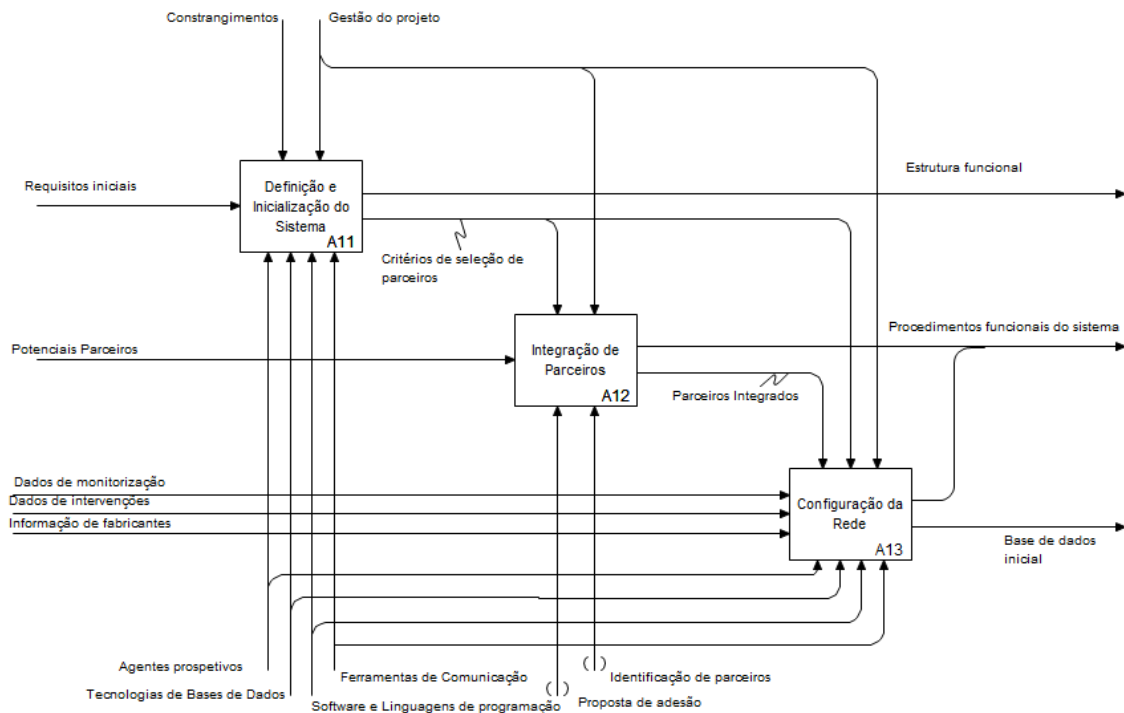


Figura 3-11 - Subatividades A1 – Sistematização da Rede Colaborativa.

Subatividade A11 – Definição e Inicialização do Sistema

É nesta subatividade que é estabelecida toda a estrutura funcional do sistema tendo em consideração todas as limitações técnicas e funcionais do mesmo. Limitações relativas às infraestruturas de comunicação a adotar por parte dos parceiros, bem como na capacidade de integração de tecnologias de agentes prospetivos de dados de forma a alimentar automaticamente o sistema com dados relativos a ações de manutenção e monitorização dos diferentes equipamentos sobre os quais serão geradas as saídas principais do sistema. Como descrito anteriormente, esta saída controla a atividade A2 e A3. De igual forma, esta subatividade gera como saída os critérios para a seleção de parceiros a integrar no sistema e que controla as subatividades A12 e A13. Estes critérios baseiam-se na capacidade de integração de um parceiro no sistema relativamente às capacidades técnicas e informáticas de forma a não causar possíveis entropias no sistema e afetando assim o desempenho global do mesmo.

Subatividade A12 – Integração de Parceiros

Nesta subatividade é elaborada a avaliação dos potenciais parceiros e posterior integração, gerando como saída os procedimentos funcionais do sistema visto este ter necessidade de se reorganizar tendo em consideração o número de parceiros nele integrados. A segunda saída

gerada por esta subatividade consiste na informação relativa aos parceiros integrados no sistema e que controla a subatividade A13.

Subatividade A13 – Configuração da Rede

Responsável pela criação da base de dados do sistema reorganizando a estrutura funcional do mesmo e conseqüentemente a sua gestão funcional. Devido à integração e ao possível abandono de parceiros do sistema, a subatividade A13 é constantemente afetada visto existir necessidade de reorganização da base de dados com a incorporação ou eliminação de dados desses mesmos parceiros.

A Tabela 3-5 inclui a lista dos novos fluxos de saída representados na Figura 3-11.

Tabela 3-5 - Fluxos de saída da subatividade A1.

Subatividade	Saídas
A11	<ul style="list-style-type: none"> • <u>Estrutura funcional</u>: Definição da estrutura funcional do sistema considerando os objetivos propostos e os requisitos funcionais do mesmo.
A12	<ul style="list-style-type: none"> • <u>Procedimentos funcionais do sistema</u>: Fluxo que estabelece todos os procedimentos funcionais a efetuar devido às necessárias adequações a efetuar no sistema relativamente aos parceiros integrados no mesmo.
A13	<ul style="list-style-type: none"> • <u>Base de dados inicial</u>: Fluxo referente à base de dados inicial do sistema integrando os dados referentes aos parceiros integrantes do sistema.

A Tabela 3-6 inclui a lista dos novos mecanismos representados na Figura 3-11.

Tabela 3-6 - Novos mecanismos apresentados na subatividade A1.

Mecanismos	<ul style="list-style-type: none"> • <u>Proposta de adesão</u>: Documento elaborado de forma a oficializar a integração de um potencial parceiro no sistema. • <u>Identificação de parceiros</u>: Mecanismo de identificação de potenciais parceiros capazes de serem integrados no sistema. Estabelecimento de diretrizes funcionais de forma a garantir a incorporação de parceiros.
------------	--

Seguidamente ir-se-á focar a subatividade A11 – Definição e Inicialização do Sistema, que é constituída por três subatividades e apresentadas na Figura 3-12.

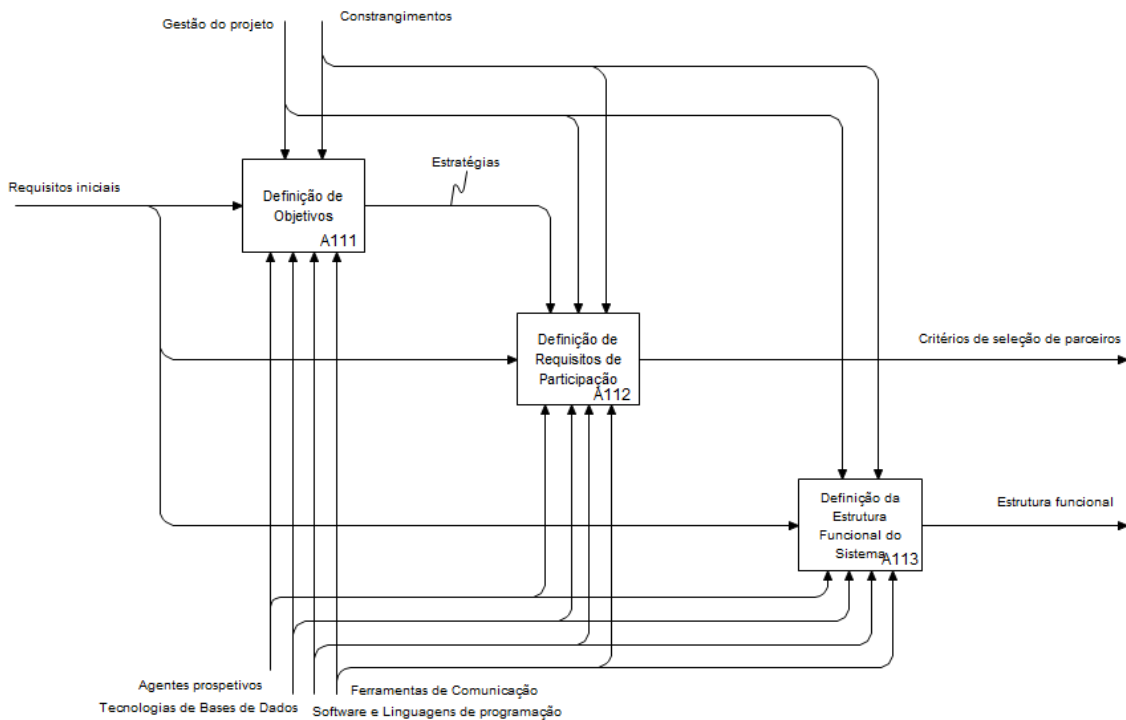


Figura 3-12 - Subatividades A11 – Definição e Inicialização do Sistema.

Subatividade A111 – Definição de Objetivos

Responsável por definir os objetivos e desenvolver as estratégias de forma a obter os resultados propostos e está integrada no processo decisório que controla a subatividade A112.

Subatividade A112 - Definição de Requisitos de Participação

Responsável por estabelecer os critérios para a seleção de parceiros. A capacidade deste integrar o sistema incorporando todas as tecnologias informáticas de suporte bem como a existência de uma infraestrutura de dados capaz de ser utilizada para os fins propostos.

Subatividade A113 – Definição da Estrutura Funcional do Sistema

Estabelece a arquitetura funcional do sistema tendo em conta as especificações iniciais e os constrangimentos referentes à seleção de parceiros que poderão afetar essa mesma estrutura funcional. Nesta subatividade são efetuadas as definições do sistema tendo em conta os requisitos do mesmo, bem como dos parceiros a integrá-lo.

Com a descrição da subatividade A113 terminou-se a explanação e descrição da subatividade A11 – Definição e Inicialização do Sistema, sendo seguidamente focada e descrita a subatividade seguinte e existente ao mesmo nível de representação, a subatividade A12 – Integração de Parceiros (presente na Figura 3-11).

A subatividade A12 (Figura 3-13) é constituída por três novas subatividades, a subatividade A121 – Identificação de Potenciais Parceiros, a subatividade A122 – Avaliação de Potenciais Parcerias e a subatividade A123 – Decisão de Parceria.

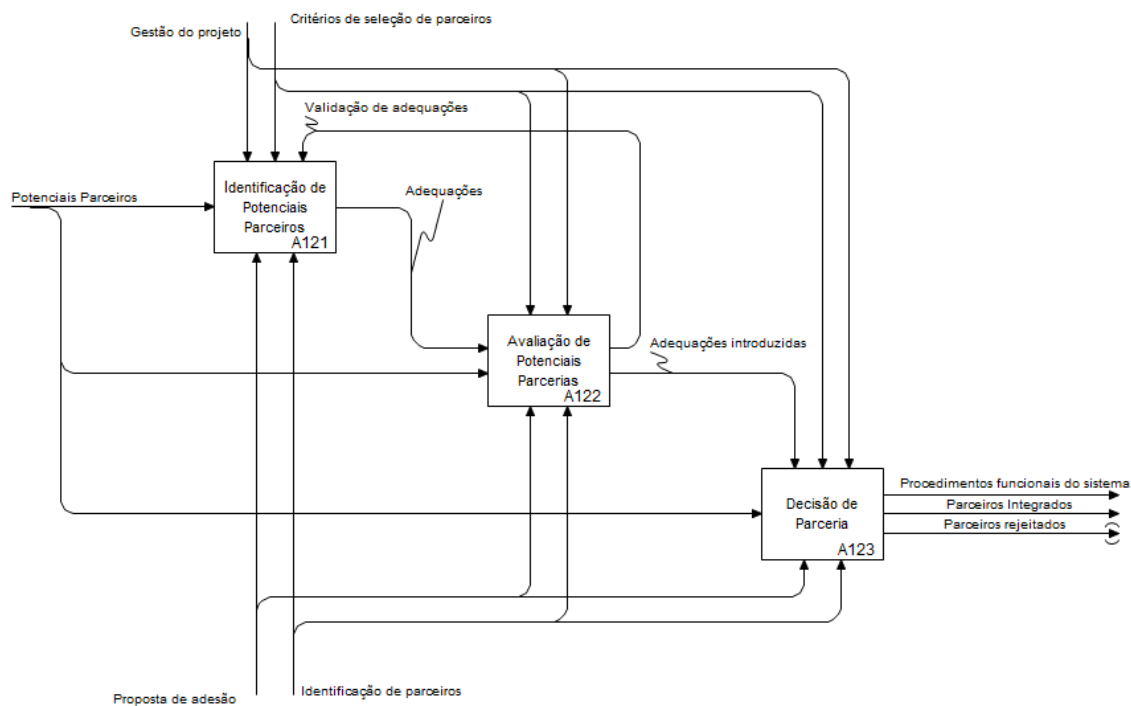


Figura 3-13 - Subatividades A12 – Integração de Parceiros.

Como entrada de todas as subatividades aqui representadas surgem os potenciais parceiros a integrar no sistema, sendo necessário identificar os mesmos e verificar a sua mais valia para o sistema (subatividade A121).

Seguidamente é efetuada uma descrição das subatividades que constituem a subatividade A12.

Subatividade A121 – Identificação de Potenciais Parceiros

Subatividade responsável pela verificação da adequabilidade de cada potencial parceiro perante os parâmetros de seleção e gestão do projeto, devolvendo como saída as adequações necessárias a elaborar por parte dos parceiros de forma a serem integrados no sistema e que serve de entrada à subatividade A122. Estas adequações devem verificar-se a nível de infraestrutura de dados, capacidade de integração de agentes automáticos para recolha de dados e comunicações.

Subatividade A122 – Avaliação de Potenciais Parcerias.

Esta subatividade é responsável por avaliar as possíveis adequações efetuadas pelos potenciais parceiros, validando essas mesmas adequações de forma a elaborar essas mesmas parcerias, A subatividade A122 devolve como saída as adequações introduzidas que controlam a subatividade A123.

Subatividade A123 – Decisão de Parceria

Subatividade responsável por estabelecer ou rejeitar essas mesmas parcerias após a análise das adequações necessárias a elaborar por parte de um potencial parceiro. A integração de um novo parceiro no sistema afeta a gestão funcional do mesmo que englobará todas as alterações funcionais inerente às novas parcerias e às já existentes que afetam toda a infraestrutura de dados que alimenta o processo preditivo.

Com a descrição da subatividade A123 termina a explanação da subatividade A12 – Integração de Parceiros e como verificado na Figura 3-11, surge como subatividade final a subatividade A13- Configuração da Rede.

Tal como descrito anteriormente, esta subatividade tem como função a agregação de dados de monitorização, manutenção e informação de fabricantes relativos aos equipamentos sob os quais os dados dizem respeito, utilizando para tal a definição estrutural relativa à arquitetura de dados a agregar dos parceiros já integrados no sistema, tendo como única saída a base de dados inicial do sistema e que serve de entrada à atividade A2 – Gestão da Rede.

Com a descrição da subatividade A13 conclui-se a exposição da atividade A1.

Seguidamente é abordada a descrição da atividade A2 – Gestão da rede, responsável pela gestão da base de dados inicial do sistema, efetuando alterações na estrutura organizacional dos dados de forma a melhorar a performance da atividade seguinte. A atividade é constituída por três novas subatividades e apresentadas na Figura 3-14.

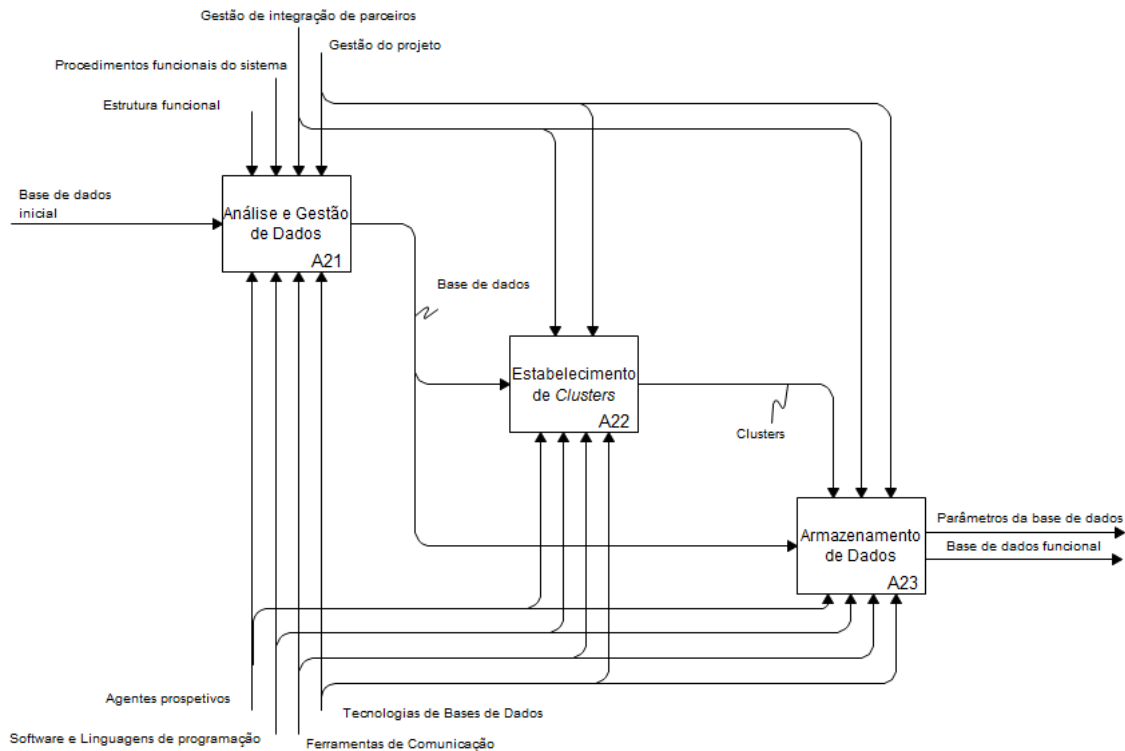


Figura 3-14 - Subatividades A2 - Gestão da Rede.

Subatividade A21 – Análise e Gestão de Dados

Recebe como entrada a base de dados inicial do sistema e consoante os parceiros integrados é elaborada uma seleção dos dados a serem integrados de forma definitiva no sistema dando origem à base de dados do sistema. Esta seleção de dados passa pela filtragem de campos que de alguma forma não sejam considerados relevantes e desta forma não sobrecarregar o sistema com o armazenamento de dados irrelevantes ao processo preditivo. A subatividade gera como saída a base de dados que serve de entrada às subatividades seguintes.

Subatividade A22 – Estabelecimento de *Clusters*

Efetua a análise da base de dados gerada na subatividade anterior e estabelece os agrupamentos (*clusters*) de dados relativos a cada tipo de equipamento proveniente do mesmo ou de diferentes parceiros e controlam a subatividade A23.

Subatividade A23 - Armazenamento de Dados

Esta subatividade recebe como entrada a base de dados resultante da subatividade A21 e conforme os agrupamentos criados na subatividade A22 efetua uma reestruturação do armazenamento de dados com a nova nomenclatura, isto é, fornece como saída a base de dados

funcional bem como os parâmetros estabelecidos e utilizados aquando da sua elaboração. Esta funcionalidade está representada na Figura 3-15.

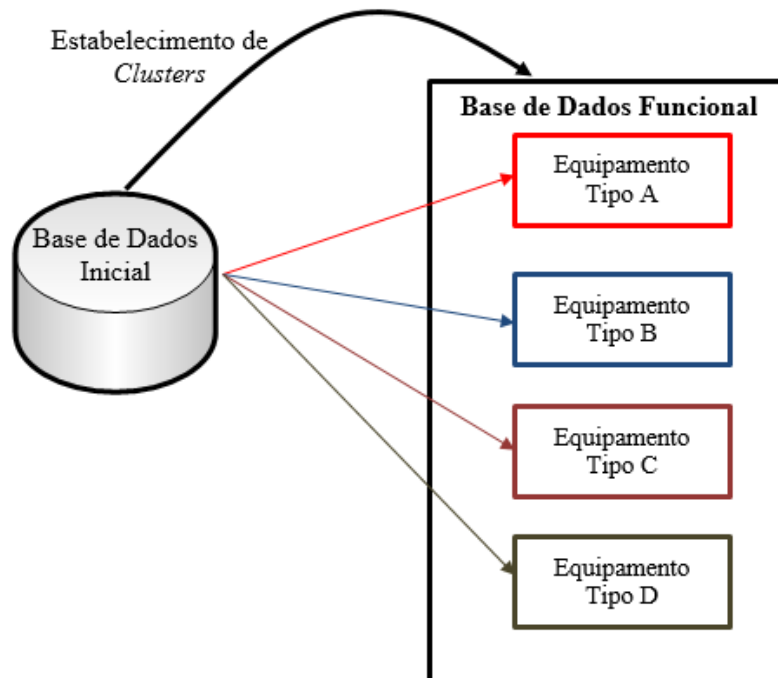


Figura 3-15 - Criação da base de dados funcional do sistema.

Neste ponto concluiu-se a apresentação e descrição da subatividade A2, tal como apresentado na Figura 3-10, sendo de seguida apresentada e analisada a atividade A3 – Operação. A decomposição desta atividade está representada na Figura 3-16. Conforme apresentado, a subatividade é constituída por três novas subatividades, a subatividade A31 – Gestão de Dados e Comunicação, a subatividade A32 – Predição de Conhecimento e a subatividade A33 – Organização dos Resultados, as quais serão a seguir descritas mais pormenorizadamente.

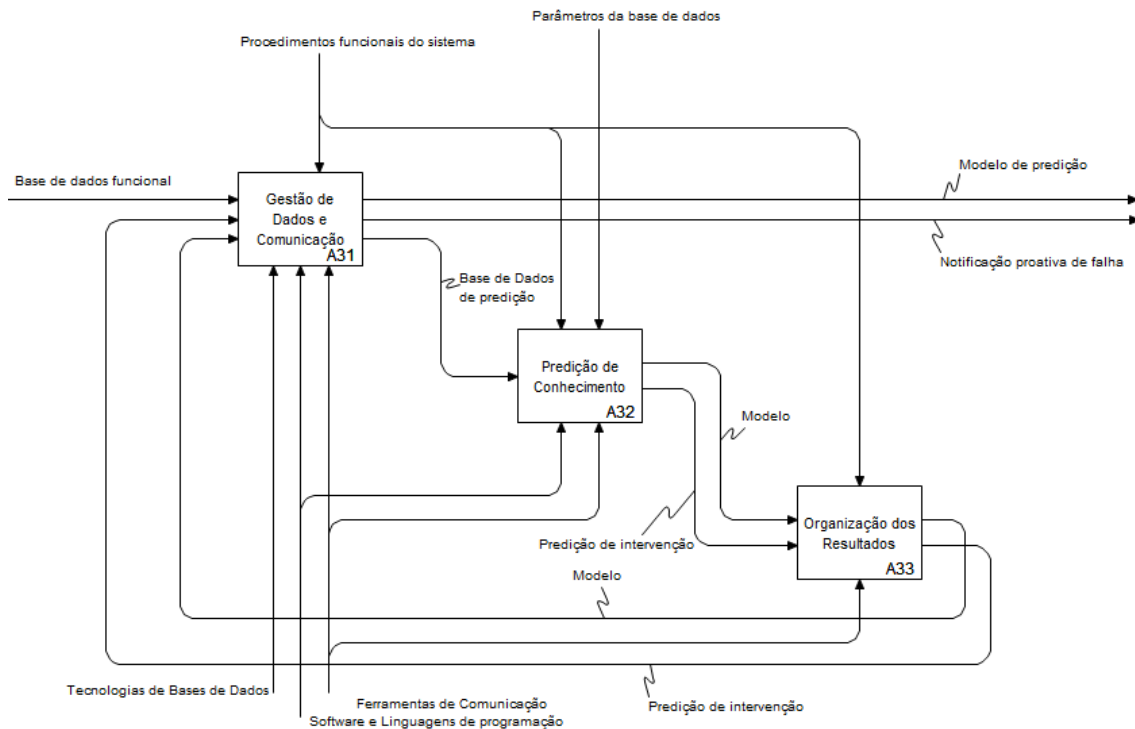


Figura 3-16 - Subatividades A3 - Operação.

Subatividade A31 – Gestão de Dados e Comunicação

Esta subatividade é responsável pela disponibilização da base de dados de predição sob a qual serão aplicadas as técnicas de DM, bem como pela difusão dos resultados através da plataforma *web* que serve de *interface* do sistema e que será apresentada mais à frente nesta tese.

Subatividade A32 – Predição de conhecimento

Subatividade central do processo preditivo. Nesta subatividade são gerados os modelos de predição e as predições de intervenções corretivas ou preventivas a serem disponibilizadas pelo sistema. Esta subatividade tem também como função o armazenamento do modelo predição gerado de forma a ser utilizado em futuras ações de predição.

Subatividade A33 – Organização dos Resultados

Após a geração das saídas pela subatividade A32, esta subatividade tem como função trabalhar e converter os modelos e as predições de intervenções em formas de visualização facilmente compreensíveis aos parceiros de sistema e posteriormente difundidas através da subatividade A31.

A figura seguinte, Figura 3-17, apresenta a decomposição da subatividade A31 – Gestão de Dados e Comunicação, em três novas subatividades.

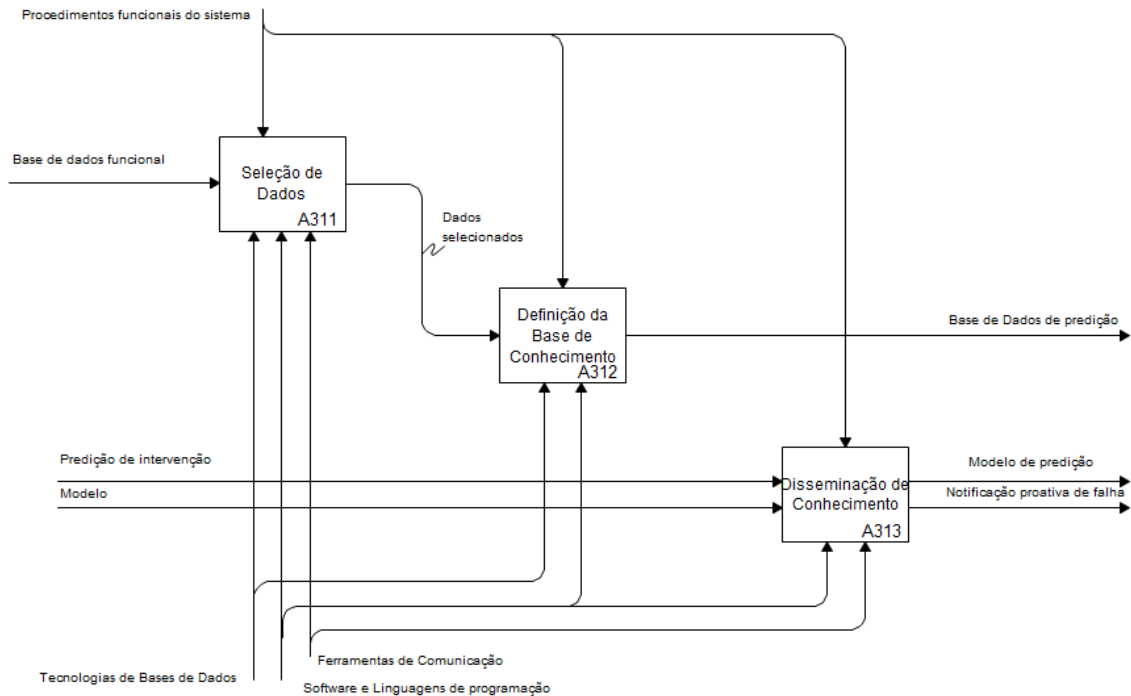


Figura 3-17 - Subatividades A31 - Gestão de Dados e Comunicação.

A subatividade A311 – Seleção de Dados

Responsável pela seleção de dados relevantes para o processo de predição de conhecimento. Elabora a seleção de dados para o estudo, sendo gerada a base de dados predição através da subatividade A312. A frequência de realização desta atividade é ditada pelos procedimentos funcionais do sistema.

A312 – Definição da Base de Conhecimento.

Após os dados selecionados a base de dados que serve de base ao processo preditivo é gerada através desta subatividade e que servirá de entrada à subatividade A32 – Predição de Conhecimento.

Subatividade A313 – Disseminação de Conhecimento

Recebe como entrada as saídas Modelo e Predição de Intervenção da subatividade A33 – Organização dos Resultados, sendo desta forma responsável pela apresentação do modelo de predição de forma compreensível aos parceiros e das notificações proativas de falha, principais saídas do sistema.

Com a descrição da subatividade A313 está concluída a descrição da subatividade A31. Assim vamos seguidamente focar a subatividade seguinte, a subatividade A32 - Predição de Conhecimento e representada na Figura 3-18.

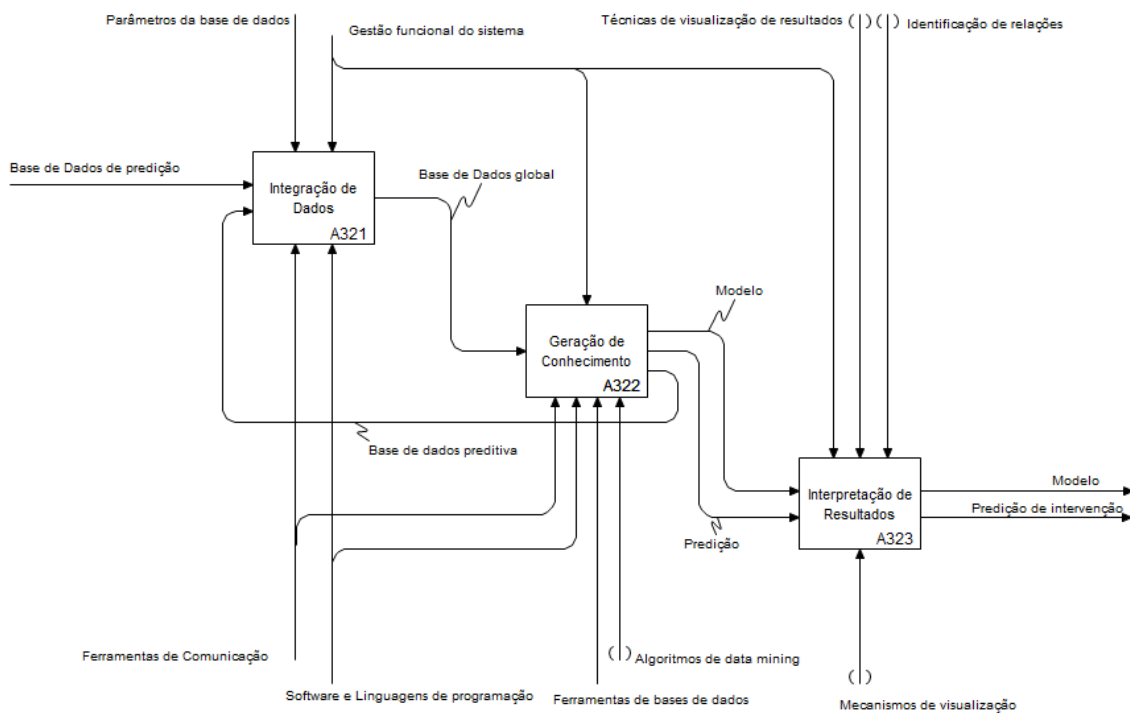


Figura 3-18 - Subatividades A32 – Predição de Conhecimento.

Esta subatividade é também constituída por três novas subatividades e explanadas seguidamente.

Subatividade A321 – Integração de Dados

Nesta subatividade é efetuado o processo de junção dos novos dados inseridos no sistema com aqueles já utilizados em anteriores atividades de predição já elaboradas de forma a que a subatividade seguinte, a subatividade A322, trabalhe com todos os dados existentes e assim atingir os resultados pretendidos com maior grau de precisão.

Subatividade A322 – Geração de Conhecimento

Recebe como entrada a base de dados global criada e gera como saída o modelo de DM produzido e a predição de avaria atingida com esse mesmo modelo. Esta subatividade é decomposta em mais dois subníveis e descritos mais à frente.

Subatividade A323 – Interpretação de Resultados

Responsável pela interpretação dos resultados da subatividade A322 recorrendo a mecanismos de visualização de forma a tornar as saídas mais compreensíveis aos utilizadores finais do sistema, isto é, às equipas ou responsável de manutenção de cada um dos parceiros.

A Tabela 3-7 inclui a lista dos novos mecanismos e informação de controlo apresentados na Figura 3-18.

Tabela 3-7 - Novos mecanismos e informação de controlo da subatividade A32.

Controlos	<ul style="list-style-type: none"> • <u>Técnicas de visualização de resultados</u>: Procedimentos e técnicas de visualização de forma auxiliar a todo o processo de visualização dos resultados obtidos. • <u>Identificação de relações</u>: Procedimentos e técnicas relativas à identificação de relações entre os dados de forma a melhor interpretação dos resultados obtidos.
Mecanismos	<ul style="list-style-type: none"> • <u>Algoritmos de <i>data mining</i></u>: Mecanismos relacionados com a aplicação de diferentes algoritmos de DM de forma a serem utilizados no sistema. • <u>Mecanismos de visualização</u>: Utilização de ferramentas capazes de gerar formas de visualização perceptíveis às saídas do sistema.

Seguidamente vamos focar o estudo na subatividade A322 – Geração de Conhecimento e todas as subatividades que a constituem através da representação dos seus subníveis. A subatividade é constituída por três novas subactividades que estão representadas na Figura 3-19.

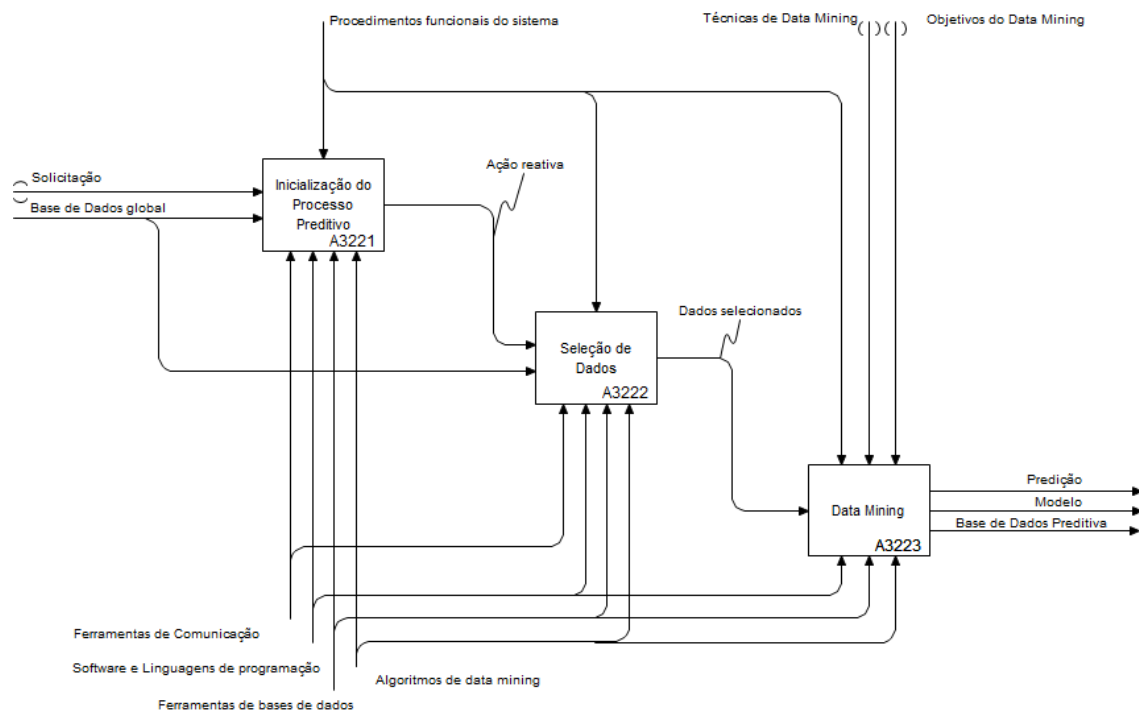


Figura 3-19 - Subatividade A322 - Geração de conhecimento.

Analisando a Figura 3-19 iremos passar à descrição mais pormenorizada das subatividades presentes e que a constituem.

Subatividade A3221 – Inicialização do Processo Preditivo

Nesta subatividade verifica-se a existência ou não de uma solicitação por parte de um parceiro (novo fluxo de entrada) de forma a ser efetuada uma ação reativa por parte do sistema, caso não exista nenhuma solicitação o sistema elaborará ações proativas de forma a gerar regras e padrões de conhecimento com todos os dados do sistema ou somente os necessários para o tipo de ação requerida.

A Figura 3-19 apresenta um novo fluxo de entrada e descrito na Tabela 3-8.

Tabela 3-8 - Novo fluxo de entrada da subatividade A3221.

Subatividade	Entradas
A3221	<ul style="list-style-type: none"> <u>Solicitação</u>: Sempre que um parceiro achar necessário pode solicitar ao sistema uma ação não planificada e fora dos parâmetros funcionais previamente estabelecidos.

A3222 – Seleção de dados

A Figura 3-20 apresenta um esquema que representa a seleção de dados, conforme a solicitação efetuada ao sistema realizada através desta subatividade.

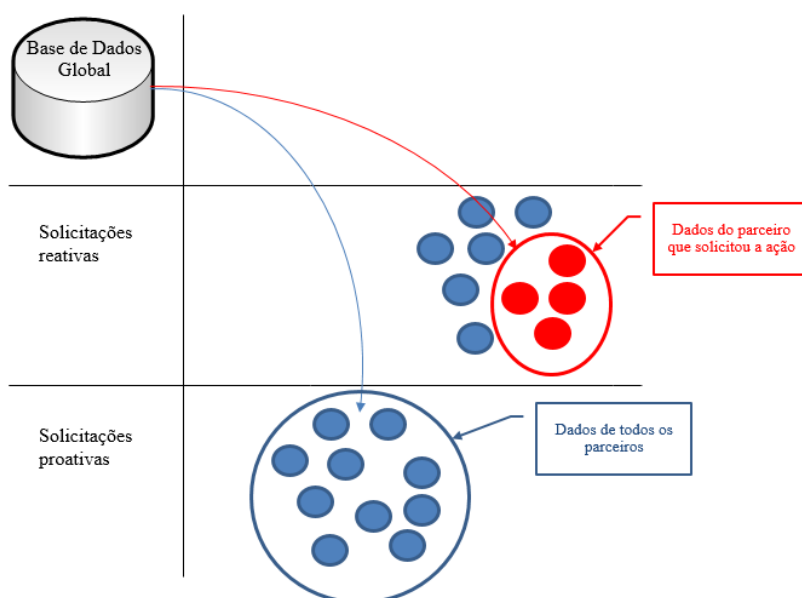


Figura 3-20 - Seleção de dados conforme solicitação ao sistema.

A figura mostra que, caso não seja efetuada nenhuma solicitação ao sistema, este é capaz de gerar automaticamente solicitações proativas a todos os parceiros utilizando para o efeito toda a base de dados existente. Caso um parceiro efetue uma solicitação reativa ao sistema, isto é, um pedido ao sistema fora dos parâmetros de atuação previamente estipulados, este selecionará os dados referentes ao parceiro em questão e irá gerar as saídas utilizando somente esse conjunto de dados.

Subatividade A3223 – Data Mining

Após o processo de seleção de dados, esta subatividade é responsável, através dos dados previamente selecionados e os modelos de predição novos ou existentes, por gerar as saídas principais do sistema.

As fases de desenvolvimento e metodologia da subatividade A3223 – Data Mining estão resumidas no seguinte fluxograma e apresentado na Figura 3-21.

Nesta subatividade surge dois novos fluxos de controlo que estão descritos na Tabela 3-9.

Tabela 3-9 - Novos fluxos de controlo da subatividade A322.

Controlos	<ul style="list-style-type: none">• <u>Técnicas de Data Mining</u>: Procedimentos de aplicação e programação das diferentes técnicas de DM presentes na ferramenta utilizada.• <u>Objetivos do Data Mining</u>: Procedimentos de seleção da técnica de DM adequada ao objetivo pretendido.
-----------	---

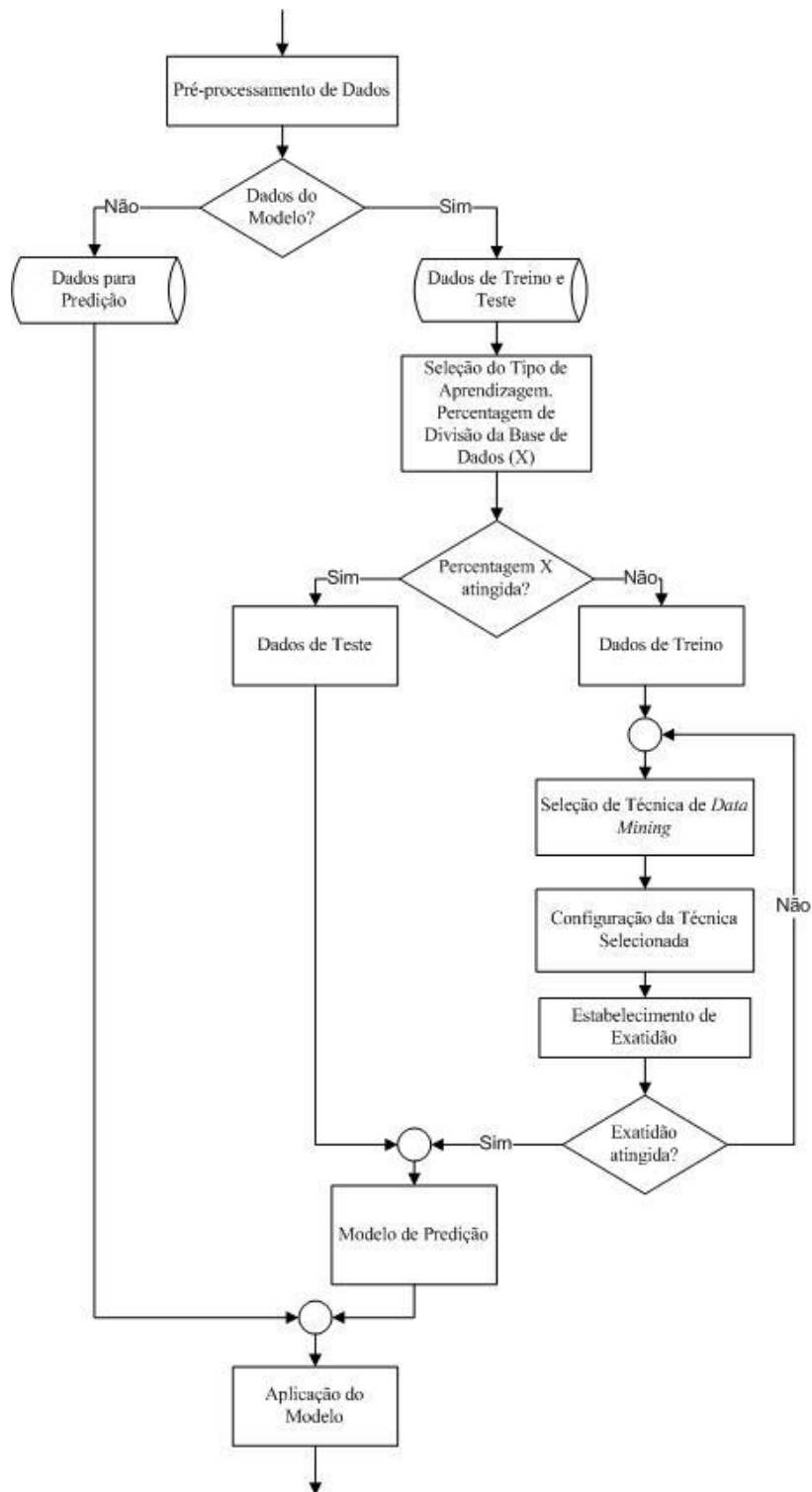


Figura 3-21 - Fluxograma da metodologia desenvolvida para implementação da Subatividade A3223 - *Data Mining*.

Para melhor compreender o funcionamento da subatividade A3223 – *Data Mining*, seguidamente é esquematizada a sua constituição, ultimo subnível da atividade A3 – Operação e de todo o sistema (Figura 3-22).

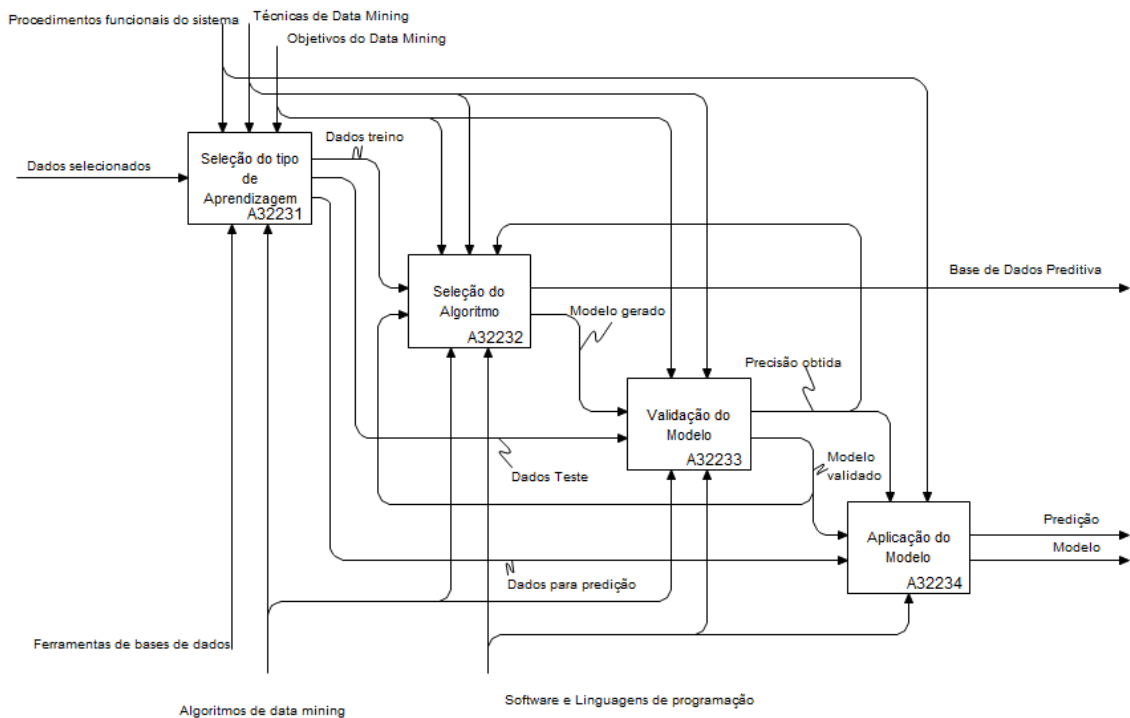


Figura 3-22 - Subatividade A3223 - *Data Mining*.

A subatividade A3223 é composta por quatro subatividades, tendo como entrada os dados selecionados na subatividade A3222 e como principais saídas o modelo de predição gerado e a predição de avaria gerada.

Subatividade A32231 – Seleção do Tipo de Aprendizagem

Nesta subatividade é efetuada a divisão da base de dados selecionada em dados de treino e teste e posteriormente aplicadas na geração do modelo de predição, e os dados a aplicar ao modelo de predição gerado de forma a devolver a predição a divulgar a parceiros. Tal como visto em capítulos anteriores desta tese, o DM preditivo carece da divisão da base de dados de suporte em dados de treino e validação, motivo pelo qual esta subatividade é essencial.

Subatividade A32232 – Seleção do Algoritmo

Recebe os dados de treino provenientes da subatividade A32231 e aplica o algoritmo de DM mais adequado aos dados em análise. Esta seleção é sempre controlada pela exatidão obtida na geração do modelo preditivo e avaliada na subatividade seguinte, a subatividade A32233.

A32233 – Validação do Modelo

O modelo gerado pela subatividade A32232 é validado através da aplicação dos dados de teste pelo sistema e após avaliada a exatidão obtida, o modelo gerado e validado é fornecido como entrada da subatividade A32234. O mesmo modelo é fornecido como entrada à subatividade

anterior podendo este ser reutilizado em futuras predições sem necessidade de geração de um novo modelo.

A32234 – Aplicação do Modelo

Esta subatividade é a responsável pela aplicação do modelo de predição validado aos dados de entrada para geração das predições, gerando as saídas principais do sistema, o modelo gerado, validado e aplicado e as predições elaboradas.

Com a descrição da subatividade A3223 – *Data Mining* conclui-se a apresentação e explanação do sistema de predição.

3.3 Recolha de dados

Nesta secção será focada na temática e a abordagem proposta relativamente à recolha de dados por parte do sistema, sendo apresentadas as tecnologias de suporte ao referido processo.

3.3.1 Tecnologias utilizadas

A recolha de dados de suporte ao sistema e seu incremento apoiam-se numa tecnologia denominada de Rede Privada Virtual, vulgarmente denominada de VPN (*Virtual Private Network*) e é efetuada a dois diferentes níveis:

- a) Nível 1 - Recolha de dados operacionais das diferentes linhas de produção das diferentes organizações inseridas no sistema. Entradas da atividade A1 – Sistematização da Rede Colaborativa, e descrita anteriormente;
- b) Nível 2 - Recolha de dados das bases de dados operacionais das diferentes organizações para a base de dados funcional do sistema de predição, isto é, a base de dados do sistema que é a entrada da atividade A3 – Operação.

Relativamente à recolha de dados respeitantes ao Nível 1 (Figura 3-23), a recolha automática de dados torna-se mais rápida e confiável devido à agregação dos seguintes fatores:

1. A evolução das redes *Ethernet*¹¹ e a utilização destas mesmas redes com padrão para a criação de redes locais;
2. A adoção por parte das organizações da linguagem *SQL* como meio de adquirir dados e guardá-los em vários repositórios;

¹¹ Arquitetura de interconexão de redes locais, sendo atualmente a mais popular. A informação é transmitida pelos diversos terminais baseando-se no envio de pacotes.

3. A facilidade na comunicação com os diferentes equipamentos através da comunicação direta com os controladores existentes ou através da criação de *software* dedicado a cada tipo de máquina. Atualmente e devido à evolução evidenciada na elaboração desses mesmos controladores, a primeira abordagem surge como a mais fácil de implementar.

Relativamente ao Nível 2, a transição de dados do sistema de armazenamento de dados das fábricas para a base de dados do sistema de previsão criado far-se-á recorrendo à criação de consultas em *SQL* e ao *XML* no sentido de normalizar os dados recolhidos. Também se recorrerá à utilização de VPN's permitindo encriptação de Alto-Nível, ou seja proteção de dados.

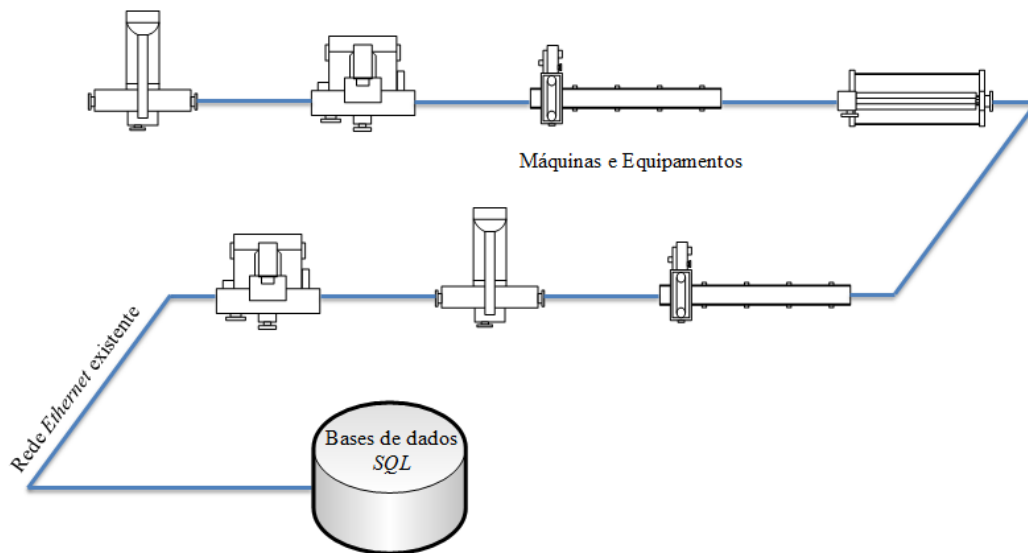


Figura 3-23 - Sistema de recolha automática de dados Nível 1.

Os dados serão oriundos de mais do que uma fonte, quer a nível de cada uma das organizações incorporadas no sistema, quer a nível de todas as organizações que poderão ser incorporadas. O extrator de dados a desenvolver deverá ser genérico podendo ser aplicável a qualquer esquema de base de dados, o que exigirá o desenvolvimento de uma linguagem de especificação apropriada que permita descrever a estrutura das fontes, a estrutura da base de dados comum e o esquema de transferência.

O extrator de dados é representado através de uma consulta SQL. A consulta poderá abranger a complexidade a desejar, e é ela que irá definir quais os registos e quais as colunas a serem exportadas. Seguidamente é apresentado o código fonte SQL genérico.

```
SELECT <lista_de_campos>  
FROM <lista_de_tabelas>  
WHERE <condições_da_seleção>  
ORDER BY <ordenação_dos_dados>
```

Se os dados estiverem estruturados com esquema relacional conhecido, o *SQL* fornece as ferramentas e a funcionalidade necessárias para uma boa manipulação desses mesmos dados. Por outro lado, se a estrutura não estiver estruturada ou desconhecida, será necessário considerar a modelagem dos dados. O *XML* surge como uma boa opção no desenho de um modelo independente da plataforma utilizada, de forma a garantir a portabilidade dos dados através da sua marcação estrutural e semântica.

3.3.1.1 XML (eXtensible Markup Language)

Segundo Heitlinger (2001), o *XML* é o formato universal para partilha de dados entre aplicações. Abrange conceitos para descrever, armazenar, intercambiar e manipular dados estruturados.

O *XML*, descreve uma classe de dados denominados de documentos *XML* e descreve parcialmente o comportamento que diferentes programas informáticos tomam quando os processam (Bray et al., 2000).

Num documento *XML* os dados são descritos em texto (*Unicode*), definindo-se “marcadores” que permitem não só descrever o conteúdo, sem dúvidas ou ambiguidades, como também fixar a estrutura lógica do mesmo, servindo para qualquer tipo de dados. Esta tecnologia tem a particularidade de ser extensível, uma vez que é o utilizador que define os “marcadores” e os seus atributos da forma pretendida (Carvalho, 2008).

A estruturação da informação através do *XML* é algo de muito importante, sendo ele capaz de definir uma estrutura de base e a gramática presente no documento, permitindo também definir o conteúdo textual dos elementos e a nomenclatura dos mesmos.

Os documentos *XML* não se destinam a ser visualizados directamente pelo utilizador. A aplicação mais importante do *XML* incide sobre a comunicação de computador a computador, ou seja, a partilha de dados e não o diálogo entre o Homem e o computador (Figura 3-24).

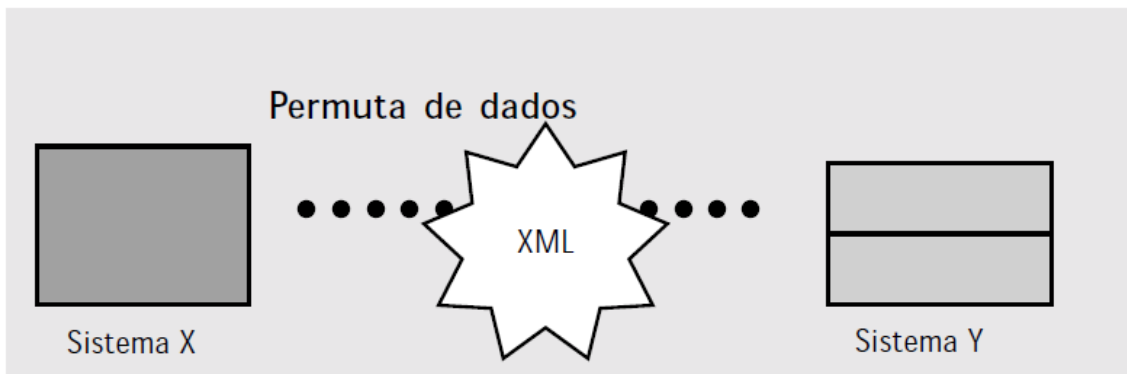


Figura 3-24 - Aplicação do XML. Adaptado de Heitlinger (2001).

O *XML* tem por objetivo conferir flexibilidade e eficiência às aplicações *Web*. De entre os benefícios, quer para o programador quer para o utilizador, pode-se salientar (Heitlinger, 2001):

- A possibilidade de buscas mais eficientes. Os dados em *XML* podem ser "etiquetados", o que permite que, por exemplo, se perquise os tipos de máquinas de um dado fabricante.
- O desenvolvimento de aplicações *Web* mais flexíveis. Inclui a integração de dados de fontes completamente diferentes, de múltiplas aplicações, a computação e manipulação local dos dados, e múltiplas formas de visualização e atualização granulares do conteúdo.
- A distribuição dos dados pela rede de forma mais comprimida e escalável. Actualmente é praticamente impossível a procura em múltiplos bases de dados incompatíveis. O *XML* permite que tais dados possam ser facilmente combinados.
- A escalabilidade. Devido ao facto dos documentos em *XML* separarem completamente os dados da forma como são visualizados, os criadores de aplicações podem torná-las muito poderosas e interativas, permitindo ao utilizador visualizar os dados da forma que mais lhe agrade.
- Múltiplas formas de visualização de dados. Os dados podem ser visualizados de diferentes formas uma vez que o *XML* define somente os dados e não o aspeto visual. A interpretação visual pode ser facultada de várias formas, de acordo com as aplicações.
- A compreensão. A compreensão de documentos em *XML* é fácil devido à natureza repetitiva das *tags*¹² usadas para definir a estrutura dos dados.

¹² Estruturas de linguagem de marcação que consistem em breves instruções, tendo uma marca de início e outra de fim.

3.3.1.2 Rede Privada Virtual

Ferguson e Huston (1998), caracterizam a *Virtual Private Network* (Rede Privada Virtual) – *VPN* de forma simples e informal:

“Uma VPN consiste numa rede privada construída sobre uma infraestrutura de rede pública, tipo a Internet”.

A designação desta rede é composta por três palavras que a definem:

- i) A palavra rede encerra toda a infraestrutura entre os dois pontos que se pretendem ligar.
- ii) A palavra privada significa que a informação transmitida é mantida confidencial e só pode ser lida por utilizadores autorizados.
- iii) A palavra virtual confere uma vertente dinâmica à ligação. A ligação pode ser alterada e adaptada a diferentes circunstâncias. Quando existe necessidade da ligação esta é efetuada, mantendo-se até ao momento em que não seja mais necessária. Esta postura, para além de incorporar uma componente dinâmica essencial na resposta às alterações que ocorrem nas necessidades de comunicação de hoje, permite também eliminar redundância na infraestrutura da rede e diminuir os custos.

De forma mais abrangente, uma *VPN* consiste numa rede de comunicações privada normalmente utilizada por uma qualquer organização ou um conjunto de organizações e/ou instituições, construída em cima de uma rede de comunicações pública (como por exemplo, a *Internet*) (Scott et al., 1999).

É possível interligar redes privadas distintas ou até mesmos computadores únicos de forma a criar uma rede particular através de “túneis virtuais” por onde flui a informação. As redes *VPN* têm a vantagem de terem custos reduzidos em relação a uma rede dedicada, *WAN*, principalmente quando as distâncias físicas entre os pontos da rede são grandes, além disso, devido à utilização de uma rede pública na sua infraestrutura, os dados podem ser visualizados por qualquer computador da rede pública. De forma a garantir a privacidade e segurança de uma *VPN*, utiliza-se a criptografia. Através do uso de criptografia os dados podem ser transmitidos com segurança através da rede. Além da vantagem financeira, as *VPN's* podem ser facilmente expandidas, sem custos adicionais para com a infraestrutura, podem suportar todo o tipo de computação móvel.

É então necessário atuar de forma a “blindar” a informação enquanto ela circula no domínio da rede pública, tornando-a dessa forma, informação privada, íntegra e autêntica. Esta atuação

vai no sentido de criar uma rede privada, em cima de uma rede pública, contornando todos os potenciais problemas que esta comporta.

3.3.1.2.1 Segurança

A principal característica das *VPN's* é a segurança. Os principais aspetos de segurança consistem na (Netgear, 2005):

- i) Autenticidade: Verifica se quem recebe ou envia a informação tem autorização para o fazer;
- ii) Confidencialidade: Protege as informações através do uso de criptografia;
- iii) Integridade: Assegura que os dados não se alteram durante o tráfego;
- iv) Controlo de acesso: Restringe o acesso a pessoas autorizadas;
- v) Disponibilidade: Mantém os dados sempre disponíveis;
- vi) Não-repúdio: Impede que algum utilizador envolvido na transação negue a sua participação no evento.

Além dos ataques de *hackers*¹³, uma rede está sujeita a vandalismos realizados por membros desta, sendo assim muito importante dispor de uma boa metodologia de segurança. Entre os métodos de segurança possíveis estão a criptografia e o uso de *Firewall*¹⁴.

3.3.1.2.2 Criptografia

O único método disponível que oferece proteção tanto no armazenamento, como no envio de informações através de uma rede pública ou através da *Internet*, é a criptografia.

O tráfego de dados encaminhado pela rede pública e utilizando protocolos padrão, não é necessariamente seguro. No entanto, uma *VPN* segura utiliza protocolos de criptografia através de um *Túnel* de forma a garantir confidencialidade, autenticação e integridade necessárias para garantir a privacidade das comunicações requeridas. Quando adequadamente implementados, estes protocolos podem assegurar comunicações seguras através de redes inseguras.

O processo de criar *Túneis* é designado na literatura inglesa por "*tunneling*". É uma tecnologia que permite que uma rede envie os seus dados através das ligações de outra rede. O *Túnel* esconde a estrutura original e a infraestrutura de encaminhamento de pacotes da *Internet* através

¹³ Indivíduo que se dedica, com intensidade incomum, a conhecer e modificar os aspetos internos de dispositivos, programas e redes de computadores.

¹⁴ Dispositivo pertencente a uma rede de computadores que tem por objetivo aplicar políticas de segurança a um determinado ponto da rede.

de um processo de transformação do pacote original. De uma forma simples, o seu papel é o de permitir o encapsulamento dos pacotes, mediante a adição de um novo cabeçalho IP^{15} ao pacote original. O IP original fica assim “diluído” nos dados (Figura 3-25). No entanto, este procedimento por si só não garante que os dados não sejam interceptados e adulterados. Para atingir esse objetivo é necessário o recurso a técnicas de criptografia (Pires, 2004).

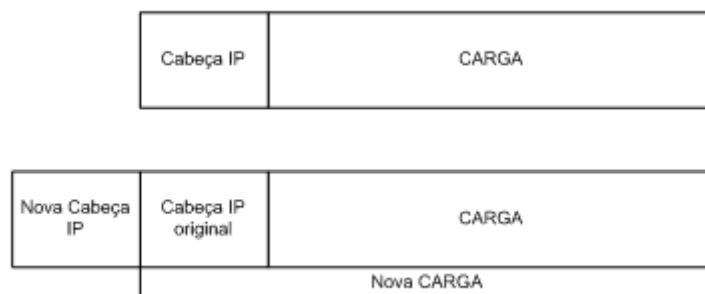


Figura 3-25 - Encapsulamento de um pacote.

Quando se fala em criptografia, fala-se também no conceito de chaves, pois elas são quem fecham e abrem dados sujeitos a criptografia. Existem dois métodos para trabalhar com chaves criptográficas:

- a) Criptografia de chaves simétricas. Este método, conhecido também como criptografia tradicional, funciona sob a premissa que o remetente e o destinatário se preparam antecipadamente para o uso da chave. Para o funcionamento deste método, todas as pessoas envolvidas devem conhecer a chave, pois quando uma mensagem criptografada chega ao destino, ela só pode ser aberta por quem possui essa mesma chave. Este método não é muito eficiente em ligações inseguras, no entanto, quando é utilizada sobre ligações seguras este tipo de criptografia é eficiente.
- b) Criptografia de chaves assimétricas ou de chave pública. Este método funciona com uma chave para criptografar, e outra para descriptografar a mesma mensagem. No sistema de chave pública, cada pessoa tem que ter duas chaves, uma que fica publicamente disponível, e outra, que deve ser mantida em conhecimento próprio.

A utilização de VPN 's por parte de diferentes organizações é uma realidade e a sua implementação pode assegurar comunicações seguras através de redes inseguras. As interligações e o acesso às bases de dados operacionais através da utilização de VPN 's sobre a

¹⁵Protocolo de comunicação usado entre duas ou mais máquinas ligadas a uma rede para encaminhamento dos dados.

Internet, geram custos insignificantes na implementação e manutenção. A utilização de *VPN's* oferece também elevada confiabilidade, integridade e disponibilidade dos dados transmitidos na rede. Os sistemas de comunicação através de *VPN's* estão a ser amplamente utilizados em diversos setores, pois oferecem absoluta segurança e elevada fiabilidade (Pires, 2004).

3.4 Desenvolvimento do interface para disponibilização de resultados

Na presente secção serão explanadas todas as ferramentas utilizadas na implementação do *website* através do qual o sistema de predição fornece os resultados obtidos, apresentando de forma amigável e segura os resultados aos utilizadores finais do sistema, i.e. as equipas ou responsáveis de manutenção.

Através deste interface *web*, os resultados obtidos pelo protótipo de predição são disponibilizados aos diferentes parceiros. A Informação é disponibilizada através de acessos controlados por *login* e palavra-passe, assegurando desta forma confidencialidade no acesso aos resultados obtidos.

Os *outputs* do sistema de predição são fornecidos em formato *html*¹⁶ pela ferramenta de DM e exportados diretamente para o servidor *web*, sendo fornecido pelo *interface web* os acessos controlados à informação disponibilizada.

O interface *web* do sistema está disponível no endereço “<http://www.esa.ipb.pt/~bastos/spamuf>” (Figura 3-26).

¹⁶ Abreviatura de *HyperText Markup Language*, que significa Linguagem de Marcação de Hipertexto. É uma linguagem de marcação utilizada para produzir páginas na *Web*.

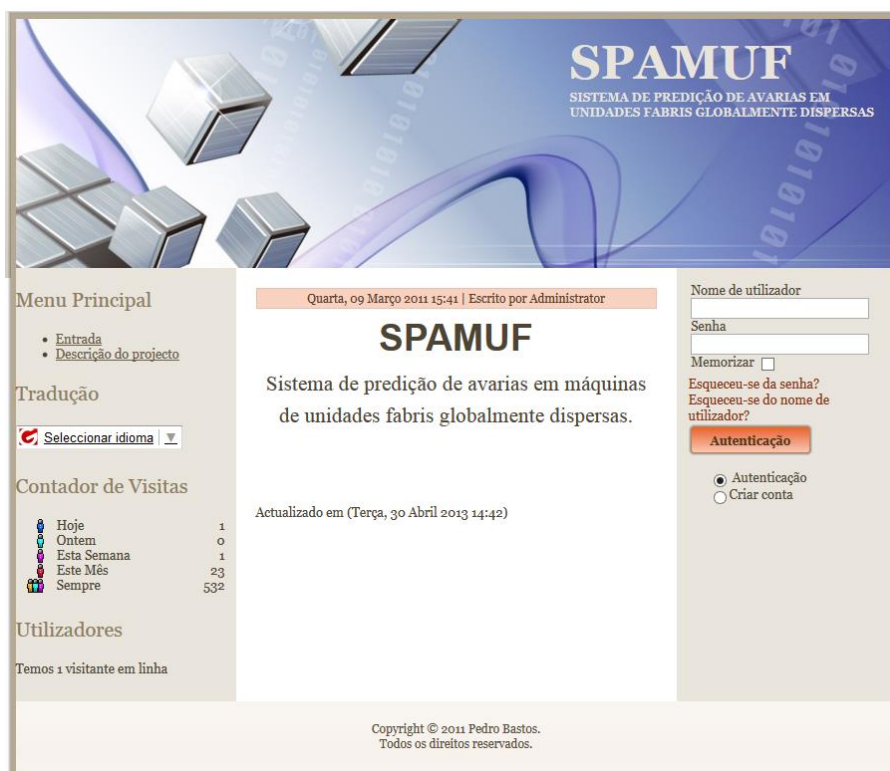


Figura 3-26 - Portal do protótipo do sistema de predição.

3.4.1 Ferramenta de desenvolvimento

O portal foi desenvolvido em *Joomla* versão 1.5.22, atualmente na versão 3.2, que consiste num Sistema de Gestão de Conteúdo ou CMS (*Content Management System*) desenvolvido a partir do Mambo¹⁷. É escrito em *PHP*¹⁸ e corre num servidor web *Apache*¹⁹ conectado a uma base de dados *MySQL*²⁰. A ferramenta proporciona dois *interfaces*, o *website* onde acedem os utilizadores e denominado de *Front-end* e a secção de administração do portal denominada de *Back-end*.

Com este tipo de plataforma o programador somente necessita de prever características básicas para o *website*. Outro tipo de características como login de utilizadores, criação, edição e publicação de conteúdos, etc., não necessitam de implementação. Estes recursos estão pré-programados e disponíveis na plataforma para serem utilizados. O *Joomla* conta com extensões que não estão na sua instalação básica, mas que podem ser adquiridos através de *download* de módulos e componentes que aumentam as suas funcionalidades. Resumidamente, o *Joomla* é

¹⁷ Consiste num sistema completo de gestão de conteúdos para criação de simples páginas *Web* a complexas aplicações empresariais.

¹⁸ Acrónimo recursivo para "*PHP: Hypertext Preprocessor*", originalmente *Personal Home Page*) é uma linguagem interpretada livre e utilizada para gerar conteúdo dinâmico na *World Wide Web*.

¹⁹ Versão do sistema operativo Linux. O Apache é um servidor web gratuito, criado sem fins lucrativos pela Apache Software Foundation, com o intuito de despoletar o crescimento da *World Wide Web*.

²⁰ Sistema gestor de bases de dados (SGBD), que utiliza a linguagem SQL (*Structured Query Language*) como interface.

um “esqueleto” de um *website* pré-programado e com recursos básicos, com fácil manutenção e administração via *web* e tem as seguintes características principais:

- Código aberto;
- Arquivamento para conteúdos não utilizados (Lixeira);
- Sistema de publicação para o conteúdo;
- Sumário de conteúdo no formato RSS ²¹;
- Busca otimizada (qualquer palavra registada);
- *Interface* traduzido em várias línguas;
- Fácil instalação para novos *templates*, módulos e componentes;
- Hierarquia para grupos de utilizadores da plataforma implementada;
- Estatísticas básicas de visitantes;
- Extensões livres em diversos *websites* (Módulos, Componentes, *Templates*, Traduções, etc.).

Todo o interface de *Back-end* do portal é disponibilizado através do endereço “<http://www.esa.ipb.pt/~bastos/spamuf/administrator>”, através de acessos controlados e está representado no Anexo A.

3.4.2 Linguagens aplicadas

Para o desenvolvimento e funcionamento do *interface* deste sistema foram aplicadas, tal como referido anteriormente, várias linguagens de programação, designadamente *PHP*, *HTML* e *MySQL*.

De modo a melhor compreensão dos motivos das escolhas efetuadas, bem como das suas potencialidades, são apresentados nas subsecções seguintes conceitos fundamentais de funcionamento de cada uma destas linguagens.

3.4.2.1 A linguagem *PHP*

O *PHP*, atualmente na sua versão 5, foi desenvolvido em 1994 por Rasmus Lerdorf, com o intuito de criar as páginas *web* da sua página pessoal. Inicialmente foi desenvolvido em *PERL*, sendo posteriormente reescrito em linguagem *C*. É uma linguagem de *Script* cujo código é inserido

²¹ RSS consiste num subconjunto de "dialetos" XML que servem para agregar conteúdo, podendo ser acedido mediante programas ou *sites* agregadores.

em páginas *HTML*, sendo o seu conteúdo interpretado no servidor *http*²². O resultado é posteriormente enviado para o posto cliente sob a forma de página *HTML* (Serrão e Marques, 2000).

Esta linguagem tem como principal característica o facto de ser capaz de processar os dados existentes numa base de dados e, ao mesmo tempo, estar embebida no código *HTML*, possibilitando desta forma a disponibilização da informação ao utilizador. Pelo facto de o *PHP* permitir facilmente a interface com o *MySQL* ou outras bases de dados, possibilita-lhe, mediante a interação com uma dessas base de dados, disponibilizar informação dinamicamente formatada. Em suma, o sucesso do *PHP* sustenta-se no facto de embeber o código *PHP* no *HTML*, processar dados provenientes de formulários, comunicar e interatuar com bases de dados de forma persistente e efetuar cálculos complexos (Serrão e Marques, 2000).

3.4.2.2 O *HTML*

Outro elemento fundamental deste interface é sem dúvida a componente elaborada em *HTML* e responsável pela apresentação dos *outputs* do sistema. Não se trata de uma linguagem de programação, mas de uma linguagem de marcação.

O *HTML* surgiu em 1990, é uma linguagem estática, interpretada pelos *browsers* e vastamente utilizada na criação de páginas *web*.

Os documentos são arquivos em forma de texto simples, com *tags* especiais ou códigos que um *browser* utiliza para interpretar e exibir informações de forma perceptível ao utilizador final.

A estrutura de um documento *HTML* está normalmente dividida em duas partes:

- Cabeçalho (*Header*) – Contém informações básicas acerca do documento, como por exemplo o título, tipo de codificação da página, formatação, funções provenientes de ficheiros externos, descrição da página, etc.
- Corpo (*Body*) – Contém tudo o que é apresentado na página *web*, como por exemplo texto, tabelas, imagens, campos de seleção múltipla, hiperligações, etc.

De forma a tornar uma página *HTML* dinâmica é possível incorporar blocos elaborados em *PHP*, *CSS*²³ ou *javascrip*.

²² Protocolo de Transferência de Hipertexto - é um protocolo de comunicação utilizado para sistemas de informação de hipermédia distribuídos e colaborativos. A sua utilização para a obtenção de recursos interligados levou ao estabelecimento da *World Wide Web*.

²³ *CSS* é uma linguagem para estilos que define o *layout* de documentos *HTML*. Por exemplo, *CSS* controla fontes, cores, margens, linhas, alturas, larguras, imagens de fundo, posicionamentos, etc.

3.4.2.3 O MySQL

O *MySQL* é um sistema de gestão de bases de dados relacional utilizando a linguagem *SQL* ²⁴ como base do seu funcionamento. Atualmente é o tipo de base de dados mais utilizado, sobretudo em aplicações *web*. A ligação entre este e o *PHP* é tão forte que normalmente um produto é associado ao outro.

O *software MySQL* funciona tipicamente numa postura cliente servidor, que se baseia num servidor *SQL* multitarefa suportando diferentes acessos (multiutilizador), diversos programas cliente e bibliotecas, ferramentas de administração várias, bem como diferentes e várias interfaces de programação, para lá de estar catalogado como sendo extremamente rápido (Pires, 2004).

3.5 Ferramenta para elaboração de modelos preditivos

Seguidamente é apresentada a ferramenta utilizada na elaboração do protótipo de predição, apresentando uma breve descrição da mesma, as suas principais características e razões da sua utilização.

3.5.1 O RapidMiner

O *RapidMiner* ²⁵ consiste num *software* capaz de desenvolver eficazmente modelos preditivos especializados e de associação entre vários atributos em vários ramos de negócio e implementá-los em operações de forma a melhorar o processo de tomada de decisão. A versão utilizada consiste na versão 5.3.015 *free Community Edition* licenciada sobre *AGPL* ²⁶ versão 3. Consiste numa ferramenta *opensource* amplamente utilizada em variados domínios de aplicação. É caracterizada pela sua flexibilidade e o seu ambiente gráfico capaz de simplificar a elaboração das tarefas mais complexas.

Este *software* é implementado em *Java* o que possibilita uma utilização versátil, em qualquer sistema operativo ou ambiente de trabalho.

Esta ferramenta fornece aos seus utilizadores:

²⁴ Linguagem de pesquisa declarativa padrão para bases de dados relacionais. Muitas das características originais do SQL foram inspiradas na álgebra relacional.

²⁵ *RapidMiner V5.2*. Disponível em <http://rapid-i.com>. Acesso em 01/03/2012.

²⁶ GNU *Affero General Public License*, é a designação da licença para *software* livre idealizada por Richard Matthew Stallman em 1989, no âmbito do projeto GNU (*GNU is Not Unix*) da *Free Software Foundation* (FSF). Copyright © 2007 Free Software Foundation, Inc. <http://fsf.org/>

- i) Uma solução global para o desenho e implementação de um processo de DM. Permite a definição de todas as fases incluídas na metodologia *CRISP-DM*.
- ii) Um interface gráfico muito intuitiva e flexível para o desenho de um processo de DM.
- iii) Mais de 500 operadores que implementam as mais diversas técnicas e algoritmos. Ligação directa com a biblioteca de classes de aprendizagem automática *Weka*²⁷, uma das mais utilizadas na comunidade científica especializada.
- iv) Acesso as mais diversas fontes de dados: *Excel, Access, Oracle, Microsoft SQL Server, SPSS*, etc.
- v) Mais de 20 métodos para visualização de dados e modelos.
- vi) Repositórios de processos, dados e meta-dados.

Todos os processos de DM podem ser executados através do seu interface gráfico (modo *GUI* – *Graphic User Interface*), utilizando a linha de comandos de *DOS*²⁸ ou acedendo via uma aplicação *Java*. Esta ferramenta suporta a metodologia *PMML*, onde cada processo de DM construído é armazenado num ficheiro *XML*.

3.5.2 Interface gráfico do RapidMiner

O interface gráfico do utilizador disponibiliza, de forma agregada, diversas funcionalidades. É intuitivo construir diferentes fases e tarefas autonomamente, após a definição pelo utilizador.

Sem a necessidade de requisitos desenvolvidos em linguagem de programação para a implementação de um determinado processo, a ferramenta fornece um conjunto alargado de algoritmos integrados em operadores. Estes algoritmos provêm de recentes trabalhos de investigação que abordam problemáticas específicas a certas fases do processo. Destaca-se, como exemplo, a disponibilidade de novos algoritmos de modelação como é o caso das RN ou a otimização computacional de algoritmos tradicionais, permitindo maior eficiência na utilização de volumes de dados de grande dimensão (Akthar e Hahne, 2012).

Ao utilizar esta ferramenta é possível definir todo um processo de tratamento de dados, inserindo operadores responsáveis por:

- i) Operações de entrada e saída;
- ii) Algoritmos de aprendizagem (supervisionados ou não);

²⁷ O pacote de *software Weka (Waikato Environment for Knowledge Analysis)* começou a ser escrito em 1993, usando *Java*, na Universidade de Waikato, Nova Zelândia sendo adquirido posteriormente por uma empresa no final de 2006. O *Weka* encontra-se licenciado ao abrigo da *General Public License* sendo portanto possível estudar e alterar o respetivo código fonte.

²⁸ *Disk Operating System*, é um acrónimo para vários sistemas operativos que dominaram o mercado para compatíveis IBM PC entre 1981 e 1995.

- iii) Funções de selecção e pré-processamento de dados;
- iv) Validação;
- v) Visualização.

Cada operador pode ser individualmente configurado existindo diferentes ambientes para configuração ou definição do processo e execução ou visualização do mesmo.

O interface divide-se segundo três perspectivas, sendo fácil e rápido alternar entre os ecrãs correspondentes:

- *welcome* consiste no ecrã de boas-vindas;
- *design* consiste no ecrã principal onde se constrói todo o processo de DM;
- *results* possibilita a análise dos resultados, reunindo todos os *outputs* do processo de DM, possibilitando a utilização de algumas ferramentas de análise gráfica.

Cada ecrã recorre a abas onde são disponibilizadas várias vistas, disponíveis através do comando “VIEW → SHOW VIEW”.

O seu interface pode ser visualizado na Figura 3-27.

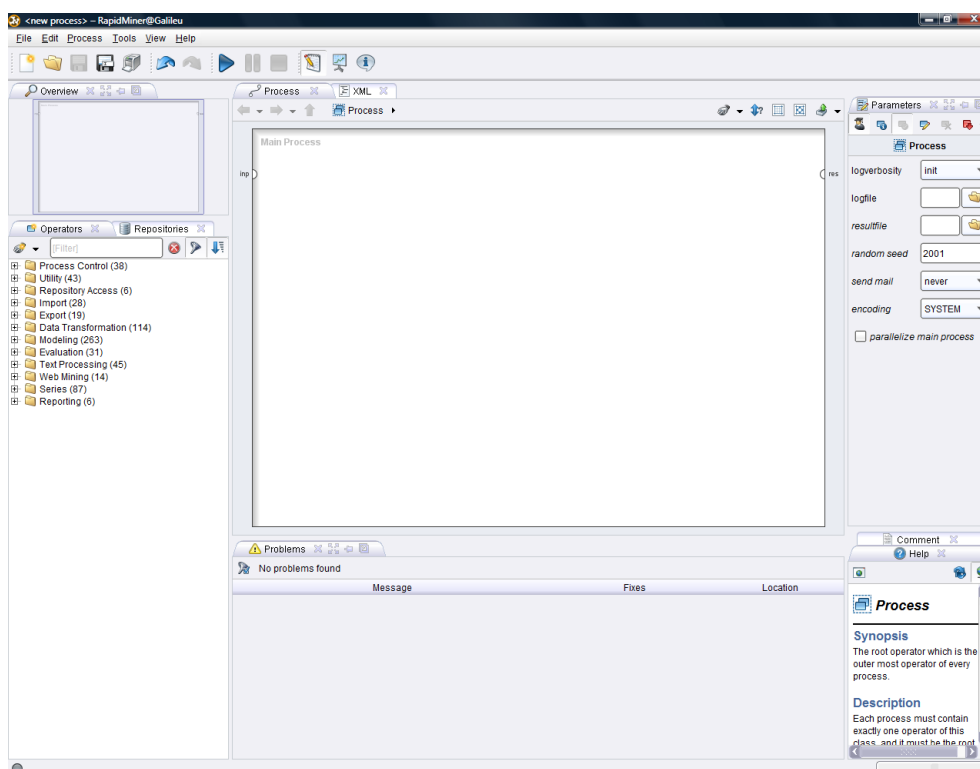


Figura 3-27 - Interface do *RapidMiner* V5.3.015.

O interface de *design* engloba os operadores. Esta designação referir-se-á aos elementos que englobam as ferramentas úteis à implementação do processo de DM. São constituídos por uma ou mais instruções directas (escrita, concatenação, etc.) e implementam um ou mais algoritmos

para pré-processamento, modelação e avaliação. Apresentam-se como um ícone gráfico, com um *design* que inclui a designação da ferramenta e o símbolo alusivo à família de operadores a que pertence, bem como os conectores para os *inputs/ outputs* de e para outros operadores.

Na Figura 3-27 pode-se observar as abas que podem ser seleccionadas durante a implementação de um processo de DM e que estão descritas no Anexo B deste documento.

Resumidamente, o *RapidMiner* opera através da elaboração de processos (desenho de projetos de DM). Isto significa que, para realizar uma classificação de dados ou predição de ocorrências nesta ferramenta, é necessário incluir operadores para cada processo que antecede a tarefa final desejada. Alguns operadores possuem parâmetros configuráveis. A lista de operadores, a sua sequência de execução e os parâmetros são armazenados num arquivo *XML* de fácil visualização e edição.

A elaboração do sistema de predição em *RapidMiner* é produzido através da combinação de um conjunto de operadores. A ferramenta permite a representação do processo através de um sistema em árvore de operadores ou por um ambiente gráfico de fluxo de processo (*work flow*). Em ambos os casos, a estrutura do processo é ainda descrita internamente em *XML*, permitindo adicionalmente o desenvolvimento do processo nesta linguagem. É de salientar de igual forma a possibilidade de definir pontos de interrupção do processo que permitem inspeccionar praticamente todos os resultados intermédios e a capacidade de combinar e agrupar operadores em blocos autónomos, disponíveis para processos posteriores.

Capítulo 4

4. Desenvolvimento e implementação do módulo de predição

No capítulo anterior concretizou-se a especificação funcional do sistema de predição, processo cujo desenvolvimento possibilitou a consciencialização da existência de desafios adicionais, com grande pertinência para o foco principal deste trabalho, que basicamente se centra na capacidade de predição da ocorrência de uma falha num equipamento pertencente a uma linha de montagem, posicionada algures, numa unidade integrada nesta proposta de rede, ou como hoje se define, nesta nuvem funcional. Esta frase de grande simplicidade sintática, encerra em si, uma crescente dificuldade funcional e de implementação proporcionalmente desafiante com o incremento da concretização do conceito que lhe assiste.

Neste capítulo é apresentada a implementação da principal atividade do sistema de predição, responsável por todo o processo de geração de alertas de ações de manutenção preventiva condicionada, sendo capaz de responder às solicitações de um ou mais parceiros. O módulo de predição é testado e validado através da elaboração de um caso prático e posterior discussão dos resultados obtidos.

4.1 A Metodologia de conceção

O processo de desenvolvimento e implementação de conceptualizações passa, necessariamente, pela construção de um elemento demonstrador, suficientemente fiel ao sistema conceptual e suficientemente realista que possibilite amadurecer as ideias necessariamente especulativas que devem acompanhar a construção de conhecimento, de sistemas ou de funcionalidades. Como dificuldades adicionais ao desafio génese que se pretende resolver nesta abordagem, surgem por exemplo, a dispersão e heterogeneidade de dados provenientes da tendencialmente crescente panóplia de parceiros putativamente integrantes do projeto e a capacidade de armazenar ou, no limite, replicar de forma mais ou menos definitiva, os dados das suas organizações numa base de dados que assegure, por um lado a robustez e integridade necessárias a um sistema desta complexidade e por outro, tornar a sua utilização e manuseamento em processos simples e rápidos, ou seja utilizáveis em tempo útil.

Outro desafio que é necessário acomodar de forma concorrente ao problema estruturante do trabalho prende-se com a crescente quantidade de dados que perfazem a base de dados de

suporte à função de predição. A previsível e necessária alimentação do sistema com dados oriundos de novas ações de manutenção e monitorização implica um acréscimo de complexidade no processo de compreensão da informação e consequentemente na forma de inferir conhecimento.

É expectável que os atuais CMMS se deparem com dificuldades acrescidas quando submetidos a processos que impliquem lidar com as exigências conceptualmente previstas no sistema, nomeadamente ao nível de predição de ocorrência de falhas, seja relativamente às quantidades de dados armazenados e à capacidade de transformação desses dados em informação relevante, seja relativamente à incorporação de ferramentas e estratégias automatizadas que auxiliem as atividades que constituem o processo de inferir novo conhecimento válido a partir desses mesmos dados.

A análise efetuada evidencia que a integração de dados de diferentes proveniências, gerados através de políticas de manutenção que obedecem a entendimentos, culturas, hábitos, patamares de desenvolvimento distintos, heterogéneos e não normalizados, em cada interveniente, com o objetivo de inferir e conseguir um crescimento de conhecimento, resulta num problema de grande complexidade. Este desafio assume-se como uma etapa de índole estruturante à boa articulação de todo o sistema proposto, perspectiva pela qual a apresentação de soluções nesse sentido se assume como um contributo parcelar no sentido de uma proposta de articulação que resolva funcionalidades e potencie a colaboração e integração funcional.

O módulo de predição apresentado foca todo o processo de DCBD, partindo da seleção dos dados a utilizar na geração do modelo de predição aplicando diferentes cenários de aprendizagem, recorrendo à aplicação de distintos algoritmos de predição, até à interpretação dos resultados obtidos. O módulo fornece soluções independentemente da proveniência dos dados utilizados, disponibilizando desta forma resultados que visam tirar proveito de diferentes ambientes produtivos. Adicionalmente, a solução apresentada permite que dados gerados por diferentes parceiros coexistam na mesma base de dados, sem questionar a integridade dos mesmos.

Para provar a validade do módulo de predição e consequentemente do sistema proposto procedeu-se a um conjunto de testes representativos do funcionamento deste módulo em distintos cenários de aprendizagem, considerando a aplicação e parametrização de diferentes algoritmos de predição, confrontando os resultados obtidos com dados de ocorrências reais conhecidas. Conscientes da existência de distintas abordagens possíveis para autenticar os procedimentos de validação, com base na literatura e face às possibilidades de aplicação de um modelo a mais

casos, acadêmicos ou industriais, de forma a confirmar a sua validade, consideraram-se as abordagens de Chrobok et al. (2004), Artuso et al. (2011), Milne et al. (1998), Çiflikli e Kahya-Özyirmidokuz (2010), Yee et al. (2013) e Du et al. (2011) como as que melhor se adequam à dinâmica, às variáveis, à escala potencial de intervenientes no sistema global, face aos meios práticos disponibilizados para esse trabalho de validação. Chrobok et al. (2004) apresentam um trabalho onde são elaborados diferentes métodos de previsão de tráfego automóvel utilizando dados reais. A abordagem fundamental do trabalho apresentado pelos autores consiste no desenvolvimento duma heurística com base na análise estatística de dados históricos relativos ao tráfego diário no centro da cidade de Duisburgo, Alemanha, durante um período de dois anos. Para validação dos resultados obtidos os autores confrontaram os resultados atingidos com dados reais conhecidos possibilitando a posterior utilização do modelo, com confiança, em cenários especulativos. Milne et al. (1998) descrevem um trabalho aplicado à indústria de produção de papel com a realização de um estudo de previsão do surgimento de defeitos no papel produzido numa empresa, tendo como base dados relativos ao seu processo produtivo. O trabalho baseia-se na deteção de correlações entre os diferentes dados produtivos e dados históricos de ocorrências de defeitos, sendo estes inseridos num sistema de predição que analisa dados reais e deteta padrões que indiciam o surgimento de defeitos no papel produzido. O trabalho assentou na utilização de dados do histórico, sendo os resultados obtidos, confrontados e analisados relativamente aos acontecimentos reais respeitantes ao período considerado. Para validação dos resultados, os autores efetuam a comparação dos modelos gerados com as ocorrências reais registadas e relativas ao período em análise.

A comparação da especificação apresentada neste trabalho com outras abordagens enumeradas e descritas acima, poderá abrir novos horizontes de investigação e proporcionar a evolução do tema, visto o módulo ser suficientemente genérico para se poder aplicar em diversos cenários englobando diversos parceiros, diversos equipamentos/itens e com diferentes parametrizações. A sua aplicação fornece uma robusta ferramenta de incorporação de organizações num sistema preditivo, com uma relação estreita entre a configuração do sistema e os *outputs* esperados e com a especificação de níveis de agregação do mesmo. Esta modelação por níveis de agregação permite reduzir a quantidade de informação e a complexidade de processamento na elaboração dos resultados projetados, bem como criar um sistema distribuído de cooperação com criação e incorporação de um módulo inteligente de predição, para de uma forma dinâmica, responder a cada solicitação. Neste trabalho desenvolveram-se diferentes

modelos de predição de falhas cujos *outputs* se confrontaram com ocorrências reais, criando-se assim aproximações dinâmicas acerca da ocorrência de uma falha a cada solicitação efetuada.

4.2 O módulo preditivo

Esta secção descreve o trabalho realizado no sentido de avaliar a consistência de um sistema tipicamente experimental, cuja operacionalidade se baseia em algumas das funcionalidades definidas ao longo do capítulo 3. Convém lembrar que o objetivo primeiro com que se partiu para este trabalho consiste na especificação de um sistema que permite a deteção da possibilidade de ocorrência de falhas de um determinado equipamento e assim encetar a elaboração de intervenções de manutenção corretivas ou preventivas a fim de debelar essa falha, evitando desta forma o surgimento de uma avaria. A adequabilidade do sistema proposto traduzir-se-á numa maior valia relativamente aos atuais sistemas de manutenção quando utilizados em ambientes colaborativos.

Como foi referido, um dos baluartes desta abordagem é a DCBD, que para a geração do modelo de predição, foi desenvolvido segundo a metodologia CRISP-DM apresentada na subsecção 2.4.2.1. A ferramenta de DM selecionada e utilizada para o desenvolvimento do caso de estudo está naturalmente alinhada com a metodologia aplicada, apesar de esta ser independente da plataforma tecnológica em que os processos de DCBD são desenvolvidos.

O modelo de predição assenta na geração de dois processos: i) Um responsável por desenvolver o modelo capaz de prever a necessidade de elaborar uma intervenção numa determinada máquina considerando os parâmetros monitorizados e ii) O segundo responsável pela utilização do modelo gerado no primeiro, aplicando novas entradas de dados, gerando alertas relativos à necessidade de efetuar uma intervenção futura facilmente perceptíveis ao responsável de manutenção. A Figura 4-1 esquematiza e especifica funcionalmente a intervenção e relacionamento de ambos os processos no domínio do modelo de predição.

Conscientes que o procedimento de DCBD é iterativo por natureza, com sucessivas repetições que visam melhorar o resultado final, é imprescindível a reformulação de ambos os processos, sempre que necessário, através da aplicação e aperfeiçoamento dos diferentes cenários de aprendizagem no intuito de se obterem resultados cada vez mais precisos.

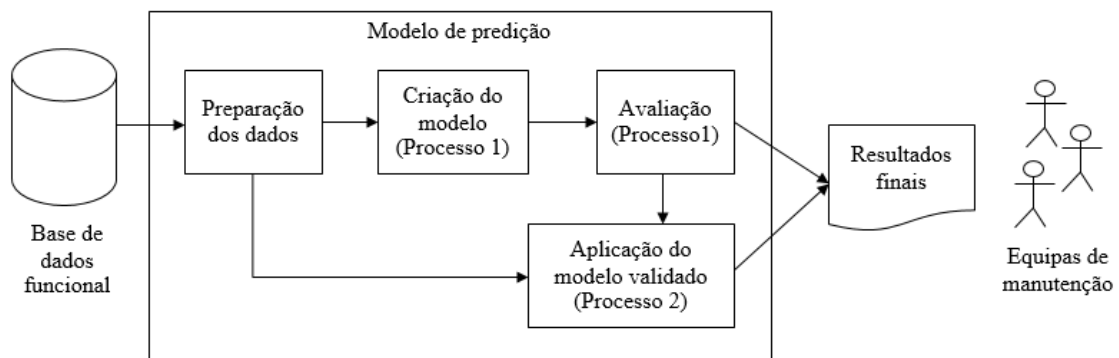


Figura 4-1 - O modelo de predição e os processos associados.

Concluída esta reflexão aos pressupostos para a elaboração do módulo preditivo do sistema, detalha-se a construção dos processos que sustentam os conceitos definidos como pilares da metodologia de predição.

4.2.1 Processo 1 – Criação do modelo preditivo

O processo 1 consiste na criação e validação de um modelo de predição que tem como variável de estudo uma variável “alvo” relacionada com os modos de falha do equipamento ou item. Basicamente, este processo é elaborado para cada agrupamento de máquinas idênticas, inseridas no sistema em função do tipo, resultando dessa agregação um número de processos equivalente ao número de agrupamentos existentes. A principal função de cada processo é a otimização da ação de predição, pretendendo-se inferir um padrão de comportamento futuro em função dos casos ocorridos no passado conjugados com dados de monitorização de diferentes parâmetros dessas mesmas máquinas.

A verificação da valorização para o sistema dos campos presentes na base de dados funcional de suporte ao processo de criação do modelo preditivo, bem como a verificação do desempenho em diferentes cenários de aprendizagem através da aplicação de quatro diferentes operadores de DM preditivo em cada um deles (i.e. AD, IR, redes *Naive Bayes* e RNA), sustenta-se na geração de vários ciclos completos de DCBD (Figura 4-2).

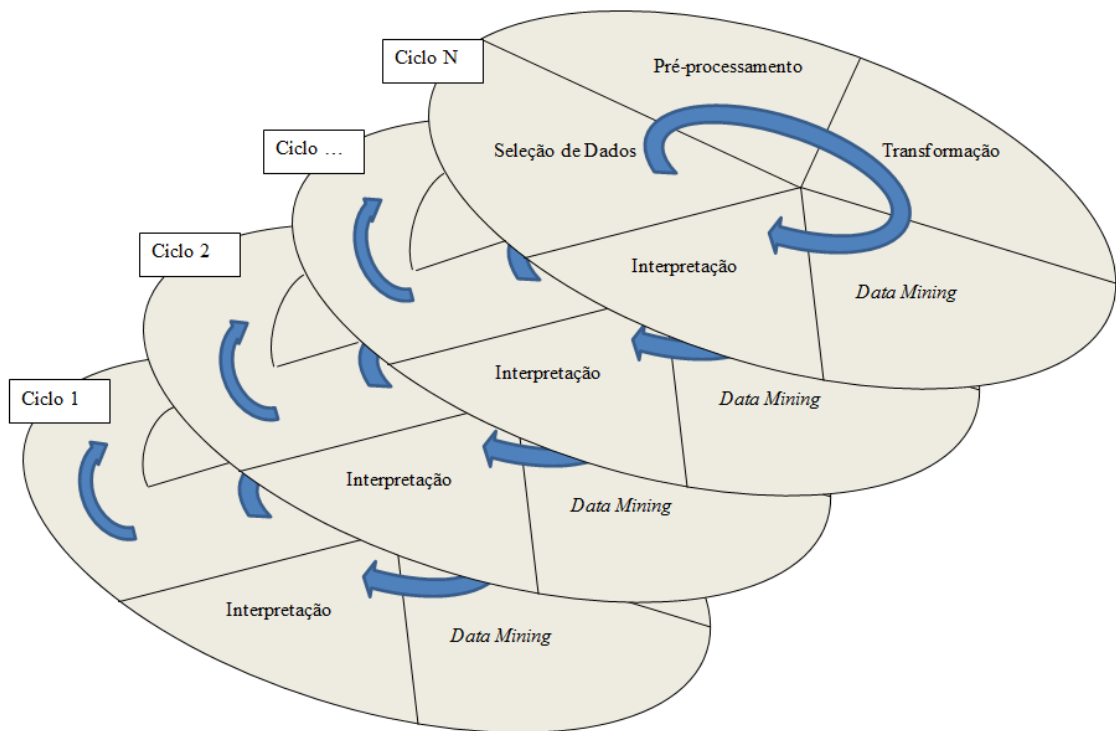


Figura 4-2- Ciclos de DCBD em modelo espiral evolutivo (de baixo para cima Ciclos 1 a N).

Na Figura 4-2 observa-se a evolução de cada ciclo de DCBD executado. Cada ciclo envolve a realização de avaliações dos campos selecionados para o estudo, aperfeiçoando a seleção de campos para o próximo ciclo, bem como a revisão da estratégia de treino através da aplicação de diferentes cenários de aprendizagem, incorporando diferentes algoritmos/operadores para DM preditivo e posterior validação dos resultados através da precisão obtida.

A Figura 4-3 apresenta o processo 1 implementado em *RapidMiner*, cujo código *XML* é apresentado no anexo C deste documento. Esta representação é sustentada unicamente na aplicação e ligação de todos os operadores necessários à programação do processo, estabelecendo os caminhos necessários ao fluxo de dados, partindo da sua origem até aos resultados finais obtidos com o seu processamento. É transversal a todos os ciclos de DCBM necessários, evidenciando desta forma a independência dos campos e dos cenários de aprendizagem a selecionar.

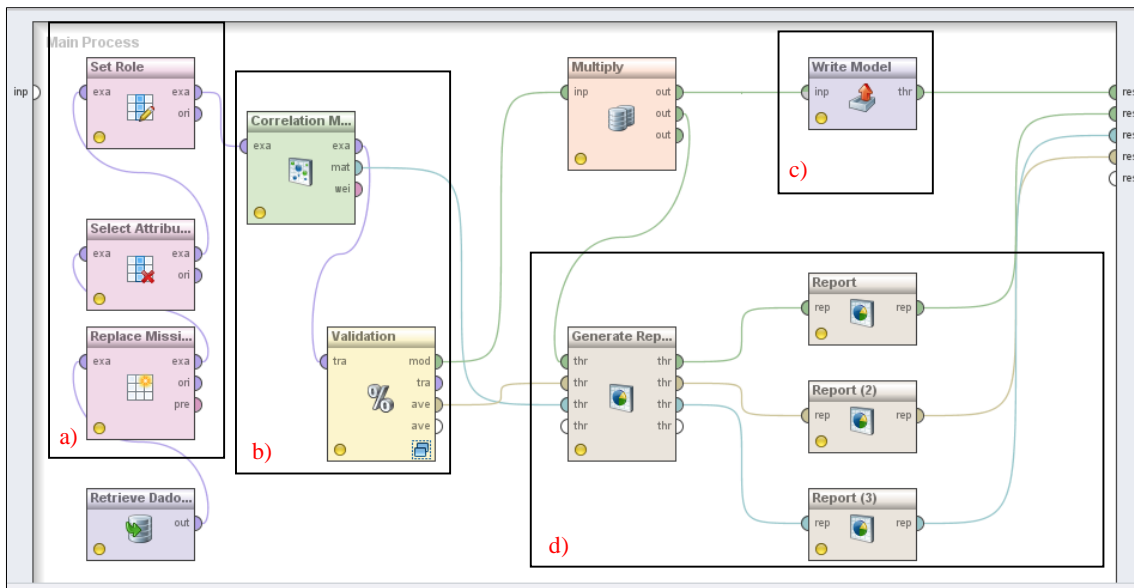


Figura 4-3 - Processo de geração do modelo de previsão.

Basicamente o processo é dividido em quatro blocos principais, designados por: a) Seleção, pré-processamento e transformação dos dados, b) DM, c) Armazenamento do modelo gerado e d) *Output* e interpretação dos resultados obtidos. É iniciado com a entrada da base de dados funcional, englobado no processo de recolha de dados, os quais normalmente podem ser provenientes de sistemas diversos, em diferentes locais, ou mesmo em diferentes suportes. Esta fase de recolha é de grande exigência, sendo por essa razão demorada. Os desafios da sua execução podem ser enumerados da seguinte forma: i) Diversidade no formato dos dados onde a evolução dos sistemas tem proporcionado ao longo dos anos diferentes formas e formatos para armazenamento e codificação dos dados. Esta heterogeneidade de formatos condiciona o processo de recolha, na medida que obriga a processos específicos para cada formato; ii) Acesso aos dados de forma a potenciar a sua utilização, sendo necessário que os mesmos estejam acessíveis ao sistema informático onde se desenvolverá o trabalho de análise. Este facto obriga à colocação dos dados num sistema ao qual o sistema informático de suporte ao modelo preditivo possa aceder *on-line*; e iii) Granularidade dos dados, que representa o nível de detalhe em que os dados são armazenados. Sistemas de informação tradicionais baseados em bases de dados, para que possam assegurar as suas tarefas, armazenam os dados com o maior detalhe possível. A granularidade influencia diretamente a velocidade do acesso à informação e o volume de dados armazenados.

A Figura 4-4 define a estrutura da base de dados funcional, e respetivos relacionamentos, assumindo-se como responsável operacional pela consistência e coerência na agregação de todos os dados oriundos dos vários parceiros, procedimento conceptualizado em 3.2, mais propriamente na descrição da subatividade A21 – Análise e Gestão de Dados.

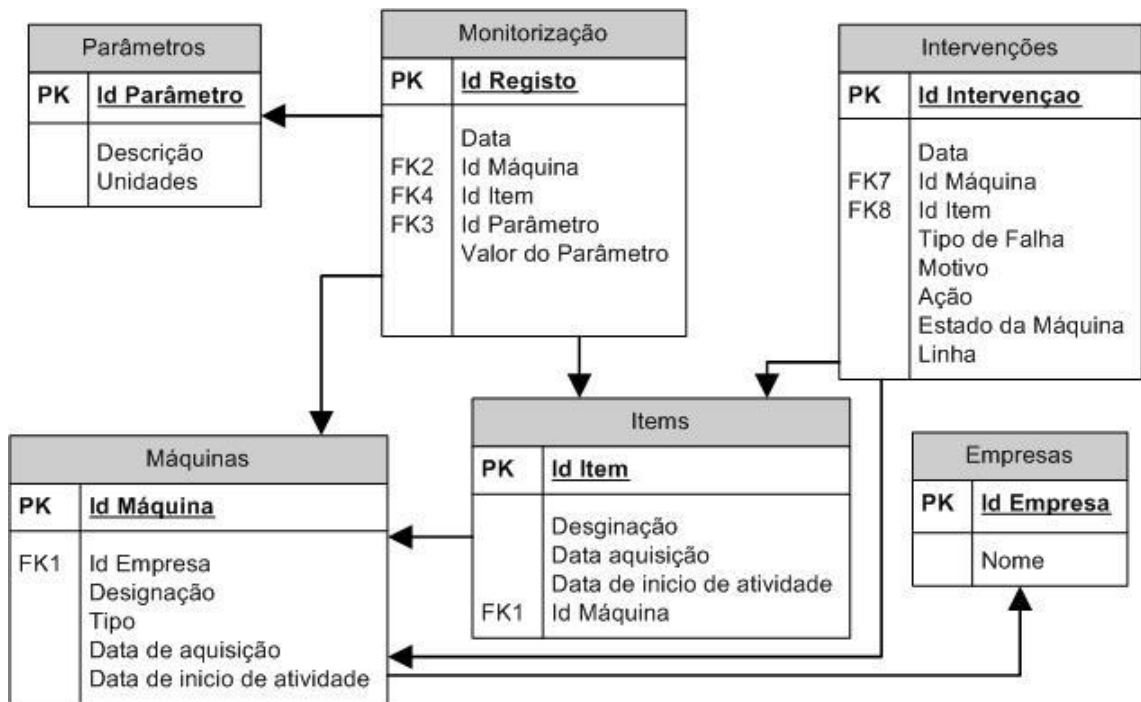


Figura 4-4 - Esquema relacional da base de dados funcional do sistema.

Uma das questões mais pertinentes na operacionalização do protótipo, e que se estenderá a uma futura utilização industrial, prende-se com a complexidade associada ao armazenamento dos dados monitorizados e à seleção dos intervalos de dados que acrescentem, de facto, relevância à base de produção de conhecimento.

A sensibilidade na complexidade do armazenamento de dados é diretamente afetada pelo aumento de parceiros inseridos no sistema, e conseqüente incremento de dados relativos a máquinas/items, pelo que a validação da relevância dos registos e conseqüente seleção se assume como preponderante.

Seguindo as orientações genéricas descritas para cada um dos processos ao longo da parte inicial desta secção, face às amostragens passíveis de serem conseguidas, e numa perspetiva de operacionalização, assume-se um período de tempo correspondente a uma janela relevante de

dados, isto é, um conjunto de dados significativos para o processo de aprendizagem do modelo preditivo e assim serem capazes de proporcionar a obtenção de modelos com elevada precisão.

Numa primeira fase, com a participação de poucos intervenientes serão utilizados dados relativos a um período de tempo relevante relativo a dados de monitorização e ações de manutenção, sendo periodicamente descartados os registos mais antigos e incorporados novos registos. Com a incorporação de mais parceiros e conseqüentemente registos relativos a mais máquinas/itens, a frequência de limpeza de dados irá ser encurtada. Sempre que é efetuada uma reformulação da base de dados de suporte, o modelo de predição será novamente gerado através da execução do processo aqui apresentado. O esquema relacional apresentado na Figura 4-4 poderá em qualquer momento sofrer alterações através da inclusão ou remoção de atributos e desta forma traduzir uma evolução no respeitante à criação dos modelos preditivos. Atributos como o tempo acumulado de funcionamento do equipamento/item ou o tempo decorrido desde a última avaria, fornecidos como variáveis de entrada à elaboração dos modelos de predição, poderão traduzir melhorias no processo preditivo do comportamento de equipamentos/itens com idades próximas.

A fase seguinte, análise, pré-processamento e transformação dos dados (Figura 4-3 a)), assume-se como crucial visto permitir a compreensão dos dados através da sua análise e descrição, bem como a seleção dos dados sobre os quais se irá desenvolver todo o trabalho de DM. Com esse intuito aplica-se o operador *<Replace Missing Values>* para atribuir o valor médio relativamente a todos os parâmetros utilizados no estudo e assim garantir que não existem campos com valores omissos, que afetem o desempenho do processo aquando da geração do modelo de predição. O passo seguinte, com recurso ao operador *<Select Attributes>* filtra os atributos sem relevância para a geração do modelo, sendo unicamente selecionados os considerados relevantes. Segundo Lee et al. (1999), a seleção de atributos consiste em obter um subconjunto dos mesmos sobre o qual o algoritmo utilizado na modelação irá trabalhar. As razões que justificam o uso de métodos para a seleção de atributos são genericamente resultantes do facto de muitos dos algoritmos não funcionarem bem com grandes quantidades de atributos, pelo que a seleção de um conjunto de atributos pode melhorar o desempenho do mesmo, e também pelo conhecimento induzido por alguns algoritmos ser frequentemente mais compreensível quando o número de atributos é mais reduzido.

Para além dos métodos enunciados existe ainda a sensibilidade do analista de DM. A experiência e o conhecimento adquirido ao longo do processo confere ao analista de DM uma

capacidade de perspetivar o tipo e a quantidade de dados que melhor se adapta à situação (Lee et al., 1999). De um modo geral, o pré-processamento dos dados é um processo semiautomático, dependendo da capacidade do analista em identificar os problemas presentes nos dados e de seleccionar os métodos mais apropriados para solucionar cada um dos casos. Resumidamente, a transformação de dados consiste em trabalhar a representação dos dados a fim de superar quaisquer limitações existentes nos algoritmos empregues na modelação.

Na segunda fase, através do operador *<Set Role>* define-se qual o atributo “alvo” de estudo, o atributo “Tipo de Falha”. Isto é, estabelece-se o *label* pretendido para o modelo, o atributo responsável pelo armazenamento de valores relativos à descrição da não ocorrência ou ocorrência de um tipo de falha.

Na fase seguinte, com a utilização do operador *<Correlation Matrix>* é aplicada uma técnica estatística capaz de mostrar quão fortemente estão relacionados pares de atributos, possibilitando assim determinar a correlação entre os atributos disponíveis e produzir um vetor de pesos e consequentemente otimizar a geração do modelo através da sua aplicação, i.e. face a um determinado conjunto de parâmetros medidos são definidos automaticamente um conjunto de pesos que indicam a dependência existente entre os atributos disponíveis na ocorrência de falhas de um item (Figura 4-3 b)). Os pesos determinados pela matriz de correlação são passados para o operador *<Validation>* facilitando desta forma a geração do modelo de aprendizagem nos diferentes cenários (fase de modelação) que assentará nos pesos calculados de cada atributo. A fase de modelação consiste na seleção de técnicas que permitam alcançar os objetivos propostos para a DCBD e que devem encontrar-se enquadrados com os objetivos traçados. A utilização do operador *<Validation>* assegura a utilização de uma aprendizagem supervisionada na criação do modelo preditivo. Este operador efetua de forma automática a divisão da base de dados em 70% para treino (aprendizagem) do modelo gerado e os restantes 30% para teste (validação) do mesmo avaliando-se o grau de robustez dos modelos gerados.

Como a fase de modelação acarreta por vezes horas de processamento, o modelo é gravado para um ficheiro binário, para posteriormente ser aplicado a futuras predições, sem necessidade de efetuar novamente o processo de geração do mesmo. Para este fim recorre-se ao operador *<Write Model>* (Figura 4-3 c)).

Na fase final do processo, interpretação e apresentação de resultados, os *outputs* (Figura 4-3 d)) são gerados em formato *htm* e enviados automaticamente para a plataforma *web* do sistema, onde é visualizada uma representação gráfica do modelo gerado, a matriz de confusão, a matriz

de correlação, bem como o grau de confiança obtido. A sumarização dos resultados obtidos e disponibilizada através do bloco *d)* do processo responsável pela criação do modelo preditivo, não consiste no principal *output* de sistema, no entanto, através da visualização destes resultados, o responsável de manutenção poderá obter informação acerca da performance do sistema na geração do respetivo modelo.

Concluída a explanação do processo relativo à criação do modelo preditivo é seguidamente focada a elaboração do segundo processo do módulo de predição. É explicitado o processo relativo à aplicação do modelo preditivo a novos dados de entrada por parte de um parceiro e posterior geração de predições.

4.2.2 Processo 2 – Aplicação do modelo preditivo

A Figura 4-5 apresenta o processo 2 implementado em *RapidMiner*, sendo o seu código *XML* apresentado no anexo D deste documento. Recorrendo à mesma abordagem que foi utilizada no processo anterior, esta representação é sustentada unicamente na aplicação e ligação de todos os operadores necessários à programação do processo de forma a estabelecer os caminhos necessários ao fluxo de dados.

O processo é igualmente dividido em quatro blocos principais: a) Seleção, pré-processamento e transformação dos dados, b) *Input* do modelo gerado pelo processo anterior, c) Aplicação do modelo às novas entradas de dados (DM) e d) *Output* e interpretação dos resultados obtidos.

Na fase de pré-processamento e transformação de dados (Figura 4-5 a)) é fornecido como *input* do processo um conjunto de novas entradas, isto é, um conjunto de dados (dados relativos a vários parâmetros monitorizados) respeitantes a um determinado instante ou intervalo de tempo. Da mesma forma que no processo anteriormente descrito, o operador *<Replace Missing Values>* é aplicado, e através do operador *<Select Attributes>*, são selecionados os campos previamente sinalizados como importantes para o processo preditivo. Como entrada deste processo é também fornecido o modelo gerado no processo anterior através da incorporação do operador *<Read Model>* (Figura 4-5 b)). Posteriormente é aplicado o modelo através do operador *<Apply Model>* (Figura 4-5 c)) aos novos registos de monitorização, transpondo para o sistema a capacidade de devolução de uma previsão da necessidade ou não de efetuar uma intervenção a um determinado equipamento/item. Como neste processo não é gerado o modelo de predição, apenas é utilizado de forma a funcionar como “professor” no processo preditivo, os *outputs* são obtidos de forma imediata, podendo desta forma as equipas de manutenção obter os alertas de falha em tempo

real sempre que dados de monitorização são inseridos no sistema para com isto desencadear ações de manutenção.

Através do bloco de *output* e configuração dos resultados (Figura 4-5 d)) o alerta do sistema é fornecido através de tabelas de fácil compreensão aos diferentes responsáveis de manutenção dos diferentes parceiros inseridos no sistema, sendo capaz de propor a necessidade ou não de uma intervenção sobre o equipamento, bem como o grau de confiança dessa mesma previsão.

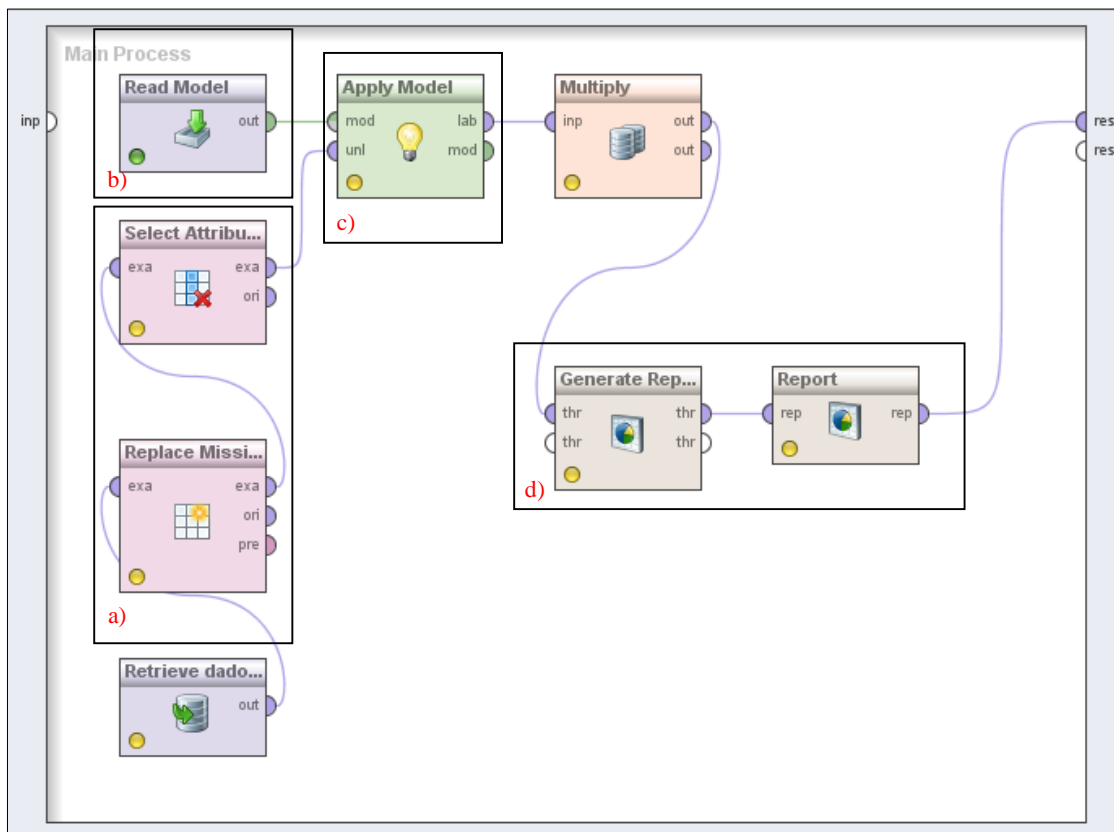


Figura 4-5 - Processo de previsão.

Concluída a descrição relativa ao processo aqui focado, finaliza-se a explanação dos processos que fazem parte da implementação do módulo de previsão, peça fulcral na engrenagem do sistema proposto e apresentado. De forma a testar o módulo de previsão exposto, é de seguida apresentado um caso prático que serviu de suporte ao trabalho até aqui realizado. O caso prático é tratado como um ensaio piloto para uma aplicação mais ampla, utilizado na preparação do modelo de previsão, auxiliando a sua construção prática.

4.3 Caso prático

Em consonância com Hartley (2004), um caso prático, ou um caso de estudo, é basicamente uma investigação detalhada de um todo de uma situação circunscrita mas não isolada, ou seja com interações e influências, desencadeada com vista a prover uma análise do contexto e dos processos envolvidos na realidade em estudo.

O desafio que se nos depara nesta etapa do trabalho é ambicioso e fulcral e consiste na concretização e validação de módulo de predição com base em testes funcionais. As orientações da implementação basearam-se nas especificações produzidas na secção 3.2 e a contextualização do sistema de predição com o modelo preditivo apresentado na secção anterior. Os dados utilizados são originários da empresa Bosch Car Multimédia Portugal S.A. sediada em Braga, disponibilizados pelo departamento de manutenção mediante protocolo de cooperação. Este protocolo possibilitou o acesso a dados reais da atividade da empresa relativos a operações de manutenção efetuadas em todas as máquinas/equipamentos existentes nas linhas de produção da empresa. Possibilitou ainda o acesso a dados provenientes do sistema de monitorização continua, já implementado na empresa e em pleno funcionamento desde 2010, que recorre à utilização de agentes automáticos para aquisição de dados, agregados já no âmbito deste trabalho, numa base de dados para tratamento posterior. Numa fase embrionária, uma das primeiras dificuldades encontradas teve origem no processo de aproximação entre a abordagem produzida conceptualmente e a realidade funcional e identidade corporativa encontrada na empresa. Na necessária ação de convergência, com ênfase nas ações de manutenção e no seu registo, face a alguma inconsistência e ausência de normalização de dados, evoluiu-se, em colaboração com a empresa, para um novo sistema normalizado de registo de todas as ações de manutenção abrindo a possibilidade de integração no presente trabalho.

Na concretização desta realidade, o processo de planeamento, criação e implementação de uma aplicação de gestão de dados relativos ao registo de ações de manutenção foi realizado paralelamente ao caso de estudo no âmbito de uma Tese de Mestrado Integrado em Engenharia Eletrónica Industrial e Computadores do departamento de Electrónica Industrial da Universidade do Minho. O resultado desse trabalho - “Sistema de gestão da manutenção dos equipamentos das linhas de montagem da Bosch Car Multimédia” (Costa, 2012) encontra-se em pleno funcionamento desde o início de 2012, momento a partir do qual foi possível a recolha e utilização de dados formatados conforme os requisitos da arquitetura proposta.

De forma a não afetar o normal funcionamento da empresa, nesta fase de teste e validação do módulo preditivo, efetuou-se a elaboração de modelos preditivos, e consequentes predições de ações de manutenção, recorrendo a uma base de dados estática, contendo os registos do ano de 2012, totalizando 91.136.166 registos de monitorização e 28.048 registos de intervenções corretivas e preventivas. Os dados de monitorização são recolhidos através de agentes automáticos, a cada segundo, sendo a data e hora registadas de forma automática em cada registo. Relativamente às intervenções efetuadas, os dados são recolhidos igualmente por agentes automáticos, ficando de igual forma registadas a data e hora da intervenção.

Face a esta dinâmica de dados, a elevada dimensão da base de dados resultante constitui uma das maiores limitações à sua utilização, basicamente devido à (in)capacidade para extrair informação relevante, para além daquilo que os processos tradicionais permitem. No entanto, e tipicamente, a concretização de um estudo utilizando DM não obriga a que todo o conjunto de dados existentes seja considerado para a construção de um modelo. A questão que de imediato pode ser colocada diz respeito à dimensão mínima necessária desse mesmo conjunto de dados. Ainda que não possa ser dada realisticamente uma dimensão mínima, uma aproximação possível será começar por uma pequena dimensão e sucessivamente ir efetuando o seu aumento, até abranger a totalidade do conjunto de dados disponíveis. Se o modelo representar com precisão os dados, o aumento do conjunto de dados não deverá alterar os resultados obtidos. Este processo de selecionar um subconjunto de dados a partir do conjunto inicial, pode ser feito através de técnicas de amostragem aleatória, normalmente incorporados nos sistemas de DM, ou através da seleção de uma amostra representativa e previamente conhecida. Após a análise dos dados existentes na base de dados da empresa, numa perspectiva de compreensão do comportamento do sistema perante uma configuração orientada à filosofia conceptual do modelo e facilmente controlável, limitou-se o número de registos para análise através da seleção de uma só máquina de uma linha de produção, obtendo-se a necessária simplicidade para a compreensão da resposta e robustez do sistema, posicionando-se como referência para posterior evolução e assim complexificar o caso prático aqui apresentado. Observando uma vez mais as orientações de Chrobok et al. (2004), Artuso et al. (2011), Milne et al. (1998), Çiflikli e Kahya-Özyirmidokuz (2010), Yee et al. (2013) e Du et al. (2011), garantiu-se, relativamente à máquina selecionada, a existência de um número representativo de dados. A máquina selecionada para teste foi submetida a cenários que reproduzem hipóteses com complexidades distintas, cujos resultados possibilitam um percurso de aprendizagem para a obtenção de confiabilidade na resposta e previsão do seu

comportamento. Esta dinâmica incute maior celeridade à construção do modelo uma vez que o tempo gasto na obtenção de resultados é reduzido. No cenário mais complexo, correspondente à fase mais avançada de teste, e com uma compreensão completa do comportamento da resposta, utilizaram-se todos os indicadores monitorizados na máquina e respetivos valores ao longo do tempo, independentemente de corresponderem ou não a uma intervenção. Como existem na base de dados utilizada, dados referentes a máquinas iguais em linhas de produção diferentes, os resultados poderão ser extrapolados, numa fase seguinte, para todas as máquinas similares.

Com esta premissa, os dados referentes ao sistema de monitorização foram reduzidos a 942.733 registos e os dados de intervenções a 574 registos que se traduziu num conjunto de dados extenso, sendo assim capaz de induzir um padrão genérico de predição.

Salienta-se que devido ao protocolo efetuado com a empresa, garantindo a confidencialidade dos dados utilizados neste caso de estudo, os valores dos campos utilizados foram mantidos com a sua nomenclatura base mas não se identificou o equipamento através da sua designação. Não foi operada qualquer transformação nas designações dos campos operados que se traduzisse numa dificuldade de compreensão dos resultados pela empresa.

4.3.1 Estrutura de dados

Tal como é apresentado na secção 3.2, as especificações da base de dados são produzidas e geridas na subatividade A13 (representado na Figura 3-11) garantindo uma base de dados uniformizada e agregadora de todos os dados das ações de manutenção corretiva e preventiva, bem como dos dados de monitorização dos diversos equipamentos dos parceiros inseridos no sistema. No entanto, devido à enorme dificuldade em agregar diferentes parceiros num trabalho académico, numa fase em que se está a desenvolver um sistema num cenário inovador na área da manutenção industrial e a ser utilizado em ações de manutenção, optou-se por adequar a subatividade A13 ao parceiro existente em detrimento de o obrigar a esforços de adequação e normalização dos seus dados ao sistema conceptual. Assim, tendo em conta a necessidade premente de obtenção de dados de forma a testar o sistema efetuaram-se esforços no sentido contrário, isto é, adequou-se o sistema ao parceiro. A base de dados de suporte ao processo preditivo foi implementada de acordo com a base de dados já existente e utilizada pelo parceiro na sua atividade diária.

A Figura 4-6 apresenta a base de dados inicial do caso prático com ênfase nas chaves e relacionamentos e não nos registos, tornando mais evidente os relacionamentos entre as diversas tabelas.

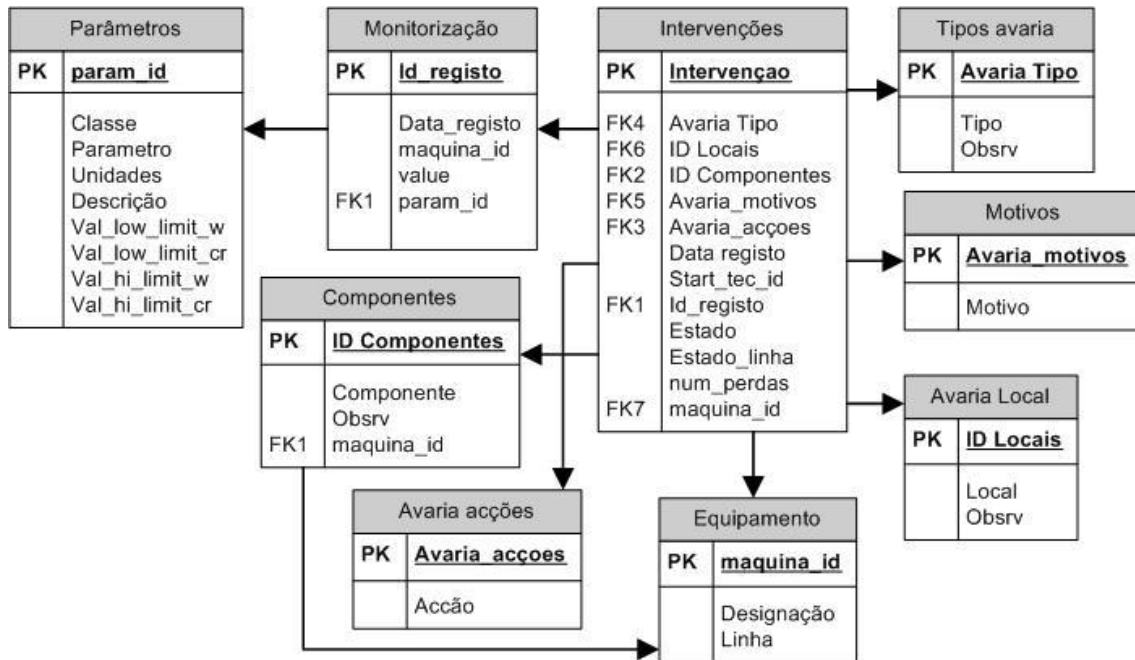


Figura 4-6 - Esquema relacional da base de dados do caso prático.

Durante o procedimento de gestão de dados e conseqüente migração funcional para operacionalização da arquitetura proposta, capaz de traduzir uma melhoria do processo preditivo, foi criado um *layer* intermedio, procurando uma melhor operacionalização do modelo. Descartaram-se na sua estrutura alguns atributos face à estrutura inicial proposta, resultando uma base de dados de predição assente em duas tabelas contendo dados de monitorização e de intervenções.

De forma a explicitar a estrutura referente às duas tabelas focadas anteriormente, que servem de base ao trabalho efetuado e resultantes do processo de seleção e pré-processamento dos dados, são apresentadas seguidamente nas Tabelas 4-1 e 4-2.

Tabela 4-1 - Estrutura da tabela de monitorização.

Campo	Descrição
Id_registro	Identificação de registo
Data_registro	Data de registo dos valores monitorizados
maq_id	Identificação da máquina
param_id	Identificação do parâmetro monitorizado
value	Valor monitorizado
Val_low_limit_w	Valor limite baixo de alerta
val_low_limit_cr	Valor crítico baixo de alerta
val_hi_limit_w	Valor limite alto de alerta
val_hi_limit_cr	Valor crítico alto de alerta

A Tabela 4-1 apresenta nove campos e a respetiva descrição. Os quatro últimos campos apresentados dizem respeito a valores limites e críticos de alerta, inferiores e superiores, definidos pela empresa. Estes valores são parametrizados pelo responsável de manutenção da empresa, otimizando o momento em que serão efetuadas as intervenções preventivas condicionadas sobre o equipamento/item monitorizado visto este apresentar algum indicio de falha. Basicamente estes intervalos servem para otimização do *timing* de intervenção sobre determinado equipamento, sendo estes valores de alerta refinados através do conhecimento empírico adquirido durante a tomada de decisão por parte do responsável de manutenção da empresa, encurtando os valores limite e críticos, e assim intervir com maior acuidade sobre determinado equipamento/item e com isto não afetar o normal funcionamento deste e da linha de produção onde está inserido. Basicamente a utilização destes limites têm um papel semelhante ao pretendido pelo modelo de predição gerado neste trabalho, i.e., indicar quando é que a falha está iminente, pelo que não representa qualquer interesse a sua integração no estudo.

No respeitante ao domínio de valores referentes ao campo “param_id”, estes estão relacionados com uma tabela denominada “livemonparams”, onde são registados diferentes parâmetros monitorizados e relativos a todas as máquinas existentes.

Ao nível do registo de intervenções, Tabela 4-2, salienta-se o campo “Avaria Tipo” que agrega registos de todas as intervenções corretivas efetuadas no equipamento/item, na sequência da ocorrência de avarias ou no âmbito da manutenção condicionada. Em termos conceptuais, o tipo de dados registados neste campo irá variar em função do detalhe da informação registada pelo parceiro. No caso concreto, o domínio de valores registados neste campo são: i) Elétrico, valor relacionado com intervenções levadas a cabo relativamente a anomalias de carácter elétrico; ii) Mecânico, identificação de intervenções de carácter mecânico; iii) *Setup*, relativas a afinações e parametrizações de funcionamento de um determinado equipamento/item e iv) Software, valor

relacionado com problemas de aplicações informáticas controladoras do equipamento/item. Relativamente a este ponto, a impossibilidade de uma identificação mais específica da falha é deve-se, mais uma vez, à falta de normalização de dados e a algumas incoerências no sistema de manutenção existente à data. Ultrapassada esta debilidade, poder-se-á evoluir para um patamar de funcionalidade mais conciso e uma conseqüente intervenção mais direcionada.

Tabela 4-2 - Estrutura da tabela de intervenções.

Campo	Descrição
Intervenção	Identificação de registo de intervenção
Avaria Tipo	Identificação do tipo de intervenção registada (Elétrico, Mecânico, <i>Setup</i> e <i>Software</i>)
ID Locais	Identificação do local da máquina onde ocorre a intervenção
ID componentes	Identificação do item sujeito a intervenção
Avaria_motivos	Registo do motivo da ocorrência da intervenção (Após intervenção externa, Após manutenção, Danificado, Desajustado, Desgaste, Erro, Erro de operação, Falha, Melhoria/Novos produtos, Partido e Sujidade)
Avaria_accoes	Ação efetuada pela equipa de manutenção (Ajuste, Alteração, Calibração, Configuração, Limpeza, <i>Restart</i> , Sem ação e Substituição)
Data_registro	Data de intervenção
Start_tec_id	Registo do id do técnico que efetuou a intervenção
maquina_id	Identificação da máquina
Linha	Identificação da linha de produção
Estado	Registo do estado da intervenção (Fechada e Incompleta)
Estado_linha	Identificação do estado da linha de produção (A 50%, A funcionar e Parada)
num_perdas	Registo do número de perdas com a intervenção

A base de dados de predição foi constituída através da agregação das duas tabelas apresentadas anteriormente e utilizando como campo agregador o campo “Data_registro”, resultando uma nova tabela agregadora de dados de monitorização e dados de ações de manutenção e destinada à criação e validação do modelo de predição, sendo os dados respetivamente designados por dados de treino (70%) e dados de teste (30%) prevenindo assim o sobreajustamento (*overfitting*), uma vez que um algoritmo treinado em demasia perde capacidade de generalização. O sobreajustamento acontece quando um modelo tem uma elevada capacidade de aprendizagem relativamente à complexidade inerente ao problema e/ou ao número de casos de treino. Esta tabela agrega os registos desde 1 de janeiro de 2012 até 31 de agosto de 2012. Para confrontação dos dados reais com os dados relativos às predições obtidas em todos os cenários validando desta forma o modelo gerado, foi utilizada uma segunda tabela contendo unicamente dados de monitorização relativos ao período de 1 de setembro de 2012 a 31 de dezembro de 2012.

Para provar a validade do sistema procedeu-se à confrontação dos resultados obtidos através dos dados da segunda tabela (dados de monitorização recolhidos entre 1 de setembro de 2012 e

31 de dezembro de 2012) aplicando os diferentes cenários, com as ações de manutenção realmente levadas a cabo, no mesmo período (registo de intervenções desde 01 de setembro de 2012 a 31 de dezembro de 2012), pela empresa e assim validar o funcionamento do sistema, simulando um cenário real de ações de manutenção e monitorização contínua. A Figura 4-7 apresenta uma representação gráfica da metodologia implementada para a realização do caso prático.

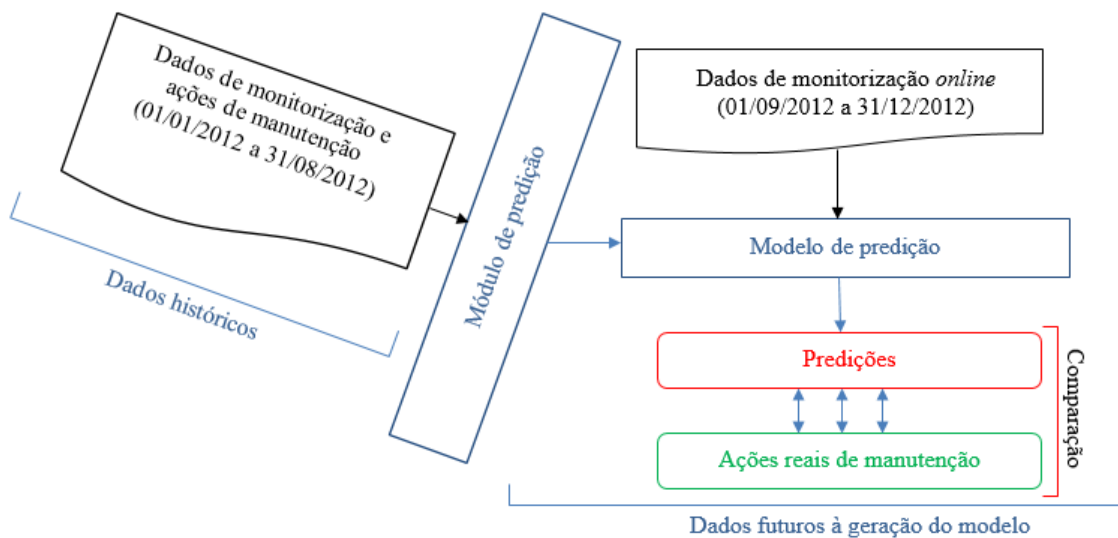


Figura 4-7 - Representação gráfica da metodologia seguida no caso de estudo.

Conhecida a explanação do processo de criação da base de dados e da definição da metodologia de validação dos resultados obtidos é importante salientar a forma de acesso à base de dados e consequentemente às tabelas geradas. O acesso é efetuado pelo *RapidMiner* instalado num sistema informático diferente do servidor onde está implementado o *interface Web* do sistema, no entanto na mesma *Intranet*²⁹, utilizando a mesma gama *IP*³⁰. No entanto o acesso à base de dados poderá ser efetuado remotamente através da utilização de um repositório remoto implementado através da configuração de um acesso VPN, tecnologia referida na subsecção 3.3.1.2 deste documento.

²⁹ Rede de computadores privada que assenta sobre um determinado protocolo *Internet*, de uso exclusivo de um determinado local, como, por exemplo, a rede de uma empresa, podendo somente ser acedida por utilizadores internos.

³⁰ Identificação de um dispositivo (computador, impressora, etc) numa rede local ou pública. Cada computador na internet possui um *IP (Internet Protocol)* único, meio que as máquinas usam para se comunicarem na *Internet*.

4.3.2 Aplicação do Processo 1

À base de dados de treino e teste do modelo foi aplicado o processo 1 em todos os cenários relativos à criação do modelo preditivo, onde numa primeira fase são selecionados os campos relevantes para o estudo. A operação é apresentada na Figura 4-8 relativa à parametrização do operador <Select Attributes>.

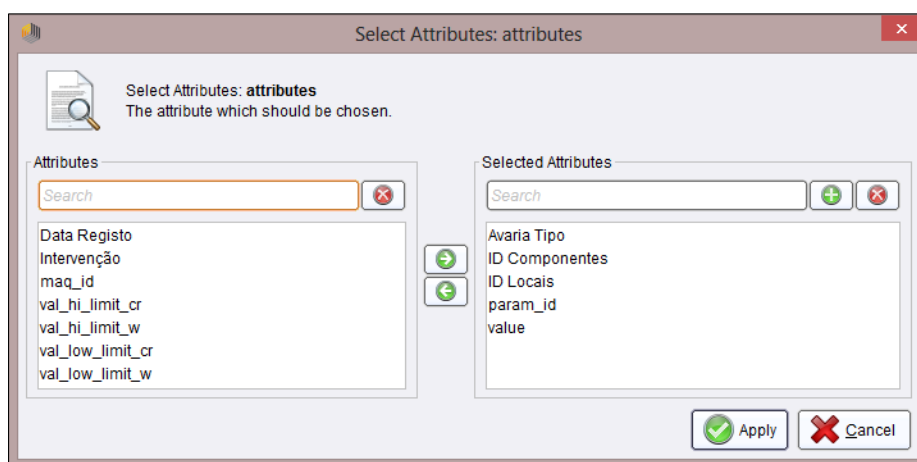


Figura 4-8 - Seleção de atributos para geração do modelo preditivo.

A figura apresenta o processo de seleção que o operador proporciona, mostrando os campos selecionados para estudo na área correspondente, e que consistem na descrição da avaria ocorrida (*Avaria Tipo*), no parâmetro monitorizado aquando da ocorrência dessa mesma avaria (*param_id*), o valor que esse mesmo parâmetro possuía aquando do registo da ocorrência (*value*), a identificação do componente ou item sujeito a intervenção (*ID Componentes*) e a identificação do local da máquina onde ocorreu a intervenção (*ID Locais*).

Seguidamente e tal como é apresentado na Figura 4-3, é definido o *label* ou objeto de estudo onde o atributo “*Avaria Tipo*” é declarado como atributo alvo do modelo a gerar.

O passo seguinte está representado na Figura 4-9, correspondente à configuração do operador <Validation> apresentado na Figura 4-3 e responsável pelo processo de geração e validação do modelo preditivo. Na fase de treino o modelo é criado e refinado usando somente a percentagem da base de dados destinada para o efeito (70%).

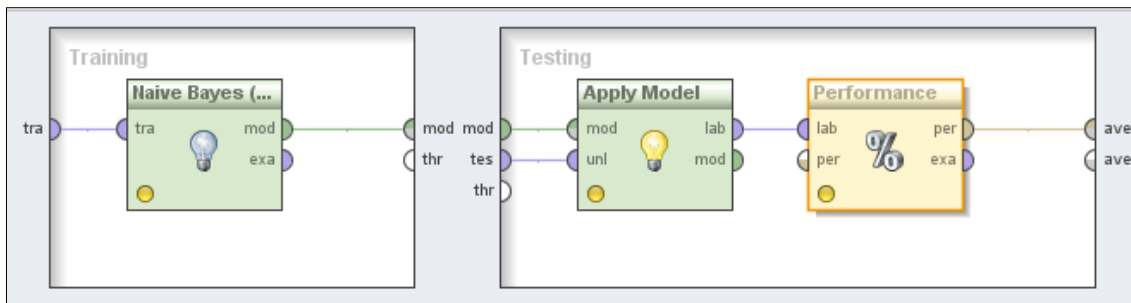


Figura 4-9 - Configuração e aplicação do operador <Validation>.

Após vários ciclos de DCBD obtiveram-se diferentes resultados avaliados pela precisão alcançada pelo modelo gerado. Com a aplicação de diferentes operadores/técnicas de DM aos diferentes cenários de aprendizagem, atingiu-se um apuro dos resultados. A Figura 4-9 apresenta o operador/álgoritmo de DM que melhor resultado atingiu, o operador *Naive Bayes (Kernel)*.

Através da aplicação de redes *Naive Bayes* (operador <*Naive Bayes (Kernel)*>) obtiveram-se resultados com maior certeza (99.98%). Para classificar o modelo gerado, além da medida de certeza do modelo, mediu-se o erro absoluto para estimar o desvio médio absoluto da previsão relativamente ao valor de certeza atingido, sendo este de 0.003 +/- 0.039 %. Outra medida de desempenho obtida consiste na medida de entropia cruzada dos classificadores, atingindo o valor de 0.020. A entropia é uma medida aplicável à partição de um espaço de probabilidade, medindo quanto esse espaço é homogéneo, ou por outro lado, quanto maior a entropia maior a desordem.

Posteriormente foi aplicado o modelo gerado à percentagem de dados para teste (30%), medindo o seu desempenho, possibilitando a verificação do grau de certeza das previsões efetuadas. Através da aplicação desse operador define-se uma previsão de todos os valores relativos aos registos existentes na base de dados relativamente ao campo de estudo (*label*), isto é, permite criar um modelo capaz de prever um valor por defeito para todos os registos existentes na base de dados. A Figura 4-10 mostra a representação da respetiva matriz de confusão obtida com o operador <*Naive Bayes (Kernel)*>, inserido no cenário que melhores resultados apresentou.

O melhor entendimento da matriz apresentada está referido anteriormente na subsecção 2.4.5 deste documento, no entanto será seguidamente abordada a sua constituição.

Reais

→

↓

accuracy: 99.98%						
	true Não	true Mecânico	true Setup	true Eléctrico	true Software	class precision
pred. Não	282398	3	0	0	0	100.00%
pred. Mecânico	0	198	33	2	3	83.90%
pred. Setup	0	8	16	0	0	66.67%
pred. Eléctrico	0	1	0	1	0	50.00%
pred. Software	0	1	0	0	0	0.00%
class recall	100.00%	93.84%	32.65%	33.33%	0.00%	

Figura 4-10 - Matriz de confusão do modelo.

Da Figura 4-10, considere-se a primeira célula de resultados, célula que contém o valor 282398 e que indica o número de verdadeiros positivos para obter o valor “Não”. Como o “Não” indica os casos em que não houve necessidade de efetuar qualquer intervenção sobre a máquina, esta medida indica que o modelo previu o valor correto para não intervenções em 282398 casos. A soma das células abaixo indica o valor 0 na sua totalidade, mostrando o número de falsos negativos ou quantas vezes o modelo previu não ser necessário efetuar uma qualquer ação de manutenção sobre a máquina, e que na realidade foi efetuada. Agrupando as colunas referentes às ações de manutenção podemos depreender os falsos positivos. Em três casos o modelo previu que não seria necessário efetuar uma intervenção, que foi concretizada no plano real. A soma das restantes células fornecem o valor respeitante ao verdadeiro negativo, isto é, a quantidade de casos em que o modelo previu que seria necessário efetuar uma intervenção mas que na realidade não ocorreu. Somando os valores das células que são diagonalmente adjacentes, é possível determinar a exatidão global do modelo. Uma das diagonais apresenta o número de previsões corretas, e a outra diagonal apresenta o número de previsões incorretas.

Os resultados apresentados na matriz de confusão do modelo evidenciam a importância da existência, aquando da criação do modelo, de um maior número de registo de ocorrências de falhas mecânicas em relação às restantes. Existindo um leque mais alargado de parâmetros responsáveis pela deteção de falhas mecânicas, o modelo trabalhou sobre um conjunto maior de registos específicos a este tipo de falha e com informação mais abrangente, sendo desta forma capaz de aprofundar o estudo das previsões relativas a falhas mecânicas e desta forma atingir um nível de precisão elevado relativamente às suas previsões específicas.

Apesar de se apresentar somente a aplicação do operador *Naive Bayes(Kernel)* (Figura 4-9) para geração do modelo preditivo, salienta-se que foram testados diferentes cenários de aprendizagem aplicando operadores distintos (algoritmos de DM) em variados ciclos, tal como descrito anteriormente nesta secção. A Tabela 4-3 apresenta a parametrização dos diferentes

algoritmos aplicados e os melhores resultados obtidos em todos os ciclos de DCBM, considerando o grau de certeza obtido na elaboração do modelo de predição, onde uma maior percentagem de certeza na geração do modelo de predição se traduz num modelo melhor e mais preciso.

Tabela 4-3 - Parametrização e resultados obtidos com diferentes operadores.

Operador	Parametrização	Melhor precisão obtida
Árvores de decisão	<ul style="list-style-type: none"> • Utilização do critério de precisão para a seleção de atributos; • Tamanho mínimo de um nodo para divisão em quatro exemplos em cada subconjunto; • Tamanho mínimo da folha de três exemplos em cada subconjunto; • Ganho mínimo de 0,1; • Profundidade máxima de vinte nodos; • Confiança de 0,25 no cálculo do erro pessimista a quando da elaboração da poda. 	92.51%
Indução de regras	<ul style="list-style-type: none"> • Utilização do critério de precisão para a seleção de atributos; • Proporção da amostra de 0,7 para treino do modelo; • Grau de pureza desejada de 0,9; • Benefício mínimo de poda a ser excedido em 0,25. 	90.3%
Redes de <i>Naive Bayes</i>	<ul style="list-style-type: none"> • Aplicada a correção de Laplace; • Modo de estimativa total; 	99.98%
Redes neuronais artificiais	<ul style="list-style-type: none"> • 300 ciclos máximos de treino; • Taxa de aprendizagem de 0,3 de forma a alterar os pesos em cada etapa de treino; • Impulso de 0,2 de forma a adicionar uma fração de peso à atualização anterior de peso para a atual; • Dados baralhados antes do processo de aprendizagem; • Os dados são normalizados antes de todo o processo de aprendizagem; • Erro de 1.0E-5 de forma a que o processo de aprendizagem termine sempre que o erro no treino do modelo atinja um valor abaixo desse mesmo valor de erro estipulado. 	64.2%

A vantagem do operador *Naive Bayes* é que requer apenas uma pequena quantidade de dados de treino para estimar a média e variância das variáveis necessárias para a classificação da variável em estudo. Como as variáveis independentes são assumidas, apenas as variâncias das variáveis referentes ao campo objeto de estudo precisam de ser determinadas e não de toda a matriz de covariância. Salienta-se que em contraste com o operador *Naive Bayes*, o operador *Naive Bayes Kernel* pode ser aplicado utilizando atributos numéricos. A função *Kernel* consiste numa ponderação utilizada em técnicas não-paramétricas de estimativa. São usados na estimativa de densidade *kernel* de forma a estimar funções de densidade das variáveis aleatórias, ou em regressão *kernel* para estimar a confiança condicional de uma variável aleatória. Durante o processo de configuração deste operador foi aplicada a correção de *Laplace* para melhorar a precisão na estimativa de probabilidades com base na frequência observada dos casos positivos

sobre o número total de casos. Este parâmetro indica se a correção de *Laplace* deve ser utilizada por forma a evitar a alta influência de probabilidade de zero ocorrências de determinado valor sobre o campo objeto de estudo.

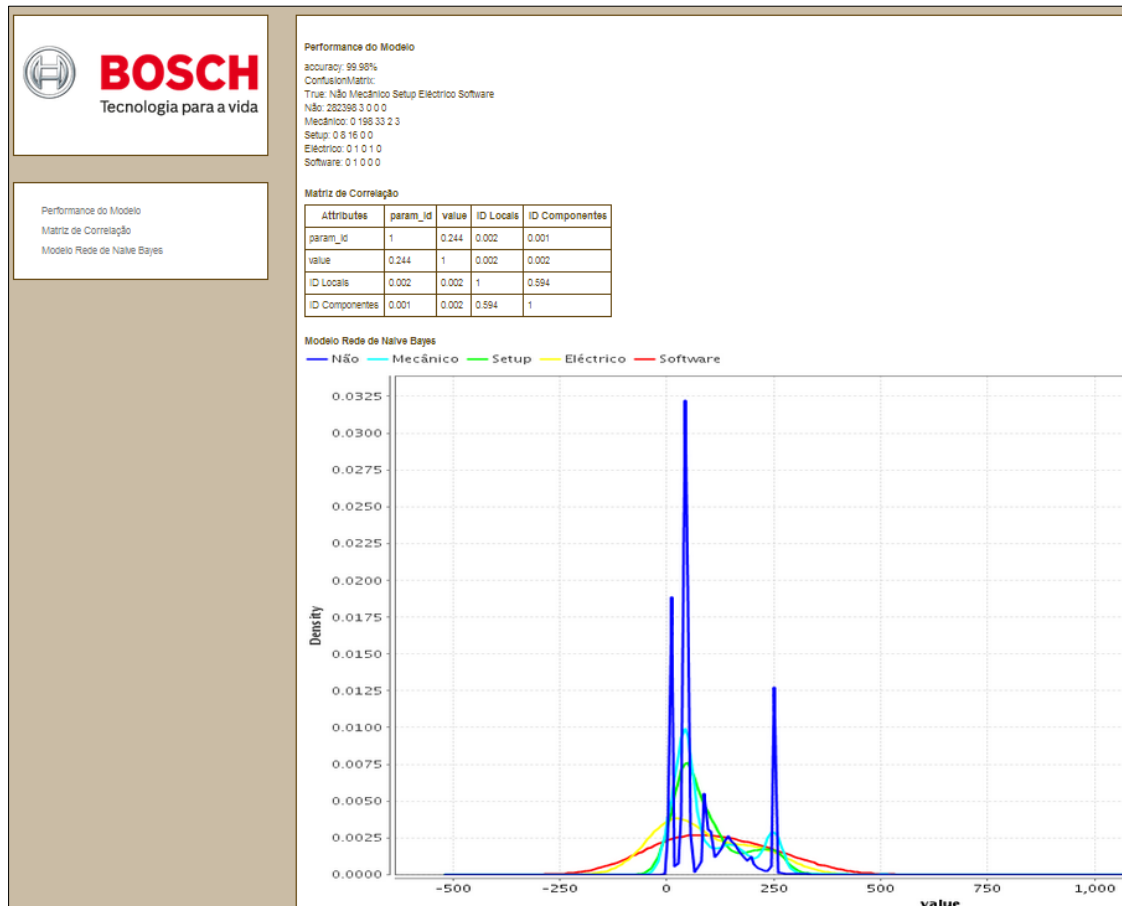


Figura 4-11 - Output gerado pelo 1º processo.

O último passo relativo ao processo 1 consiste na fase de *output* e interpretação dos resultados obtidos. A Figura 4-11 ilustra o *output* gerado para os dados utilizados no cenário onde foram apresentados melhores resultados. É apresentado ao responsável de manutenção informação meramente informativa sobre o desempenho obtido na geração do modelo, bem como uma matriz de correlação dos atributos utilizados na sua geração.

A informação complementa-se com a representação gráfica do modelo gerado, sendo neste caso, representado através do gráfico de densidade de ocorrências, produzido pelo modelo baseado no operador Redes *Naive Bayes(Kernel)*. O gráfico apresenta a densidade (perspetiva para estimar a probabilidade de um valor) alusiva ao valor medido das ocorrências previstas pelo modelo, por cada tipo de intervenção a efetuar, respeitantes a todos os parâmetros medidos

relativos à máquina analisada. Note-se que a natureza das variáveis estudadas pode ser discreta (categórica), por exemplo, com dois estados possíveis, 0 e 1, sendo as suas probabilidades condicionais facilmente representadas em tabelas de probabilidades. Podem também ser de natureza contínua, neste caso representadas por funções de densidade. O gráfico evidencia a possibilidade de visualizar a densidade de ocorrências verificadas aquando da geração do modelo, das diferentes predições e assim a possibilidade de verificar para que valores se atestam maiores números de ocorrências, independentemente do parâmetro analisado.

Da observação do gráfico da Figura 4-11 é possível concluir que o número de ocorrências registadas aquando da geração do modelo, conforme a sua natureza, se traduz numa probabilidade maior de ocorrências, justificado pela diferença de densidade de ocorrências em função do tipo de predição. Com o aumentar de registos aquando da reformulação do modelo, os valores de densidade poderão aumentar, visto existirem mais predições a efetuar por parte do algoritmo durante a sua fase de validação.

Salienta-se que o *output* aqui apresentado não consiste na principal saída do sistema. Os resultados deste processo traduzem-se em mera informação relativa à criação do modelo e transversal a todos os parceiros, podendo desta forma responder à curiosidade do(s) parceiro(s) em visualizar os resultados obtidos na geração do modelo de predição que será posteriormente aplicado de forma a gerar as predições através do processo descrito na subsecção seguinte. Realça-se de igual forma que dependendo do algoritmo de DM aplicado, a visualização gráfica do *output* aqui apresentado varia, existindo algoritmos capazes de apresentar os resultados de forma mais perceptíveis ao utilizador final.

Com o modelo de predição gerado, o seguinte passo consiste na utilização deste mesmo modelo para gerar os alertas de intervenção, utilizando para tal, novos dados de monitorização relativos à máquina utilizada para teste no melhor cenário relativo à criação e validação do modelo preditivo.

4.3.3 Aplicação do Processo 2

Em consonância com o referido em Chrobok et al. (2004), Artuso et al. (2011), Milne et al. (1998), Çiflikli e Kahya-Özyirmidokuz (2010), Yee et al. (2013) e Du et al. (2011), para verificar a funcionalidade do protótipo gerado e validar os resultados alcançados, avaliando o grau de robustez do modelo gerado, utilizaram-se no processo 2, dados de monitorização não utilizados para teste e validação do modelo gerado, dados estes relativos ao período compreendido entre 1

de setembro de 2012 e 31 de dezembro de 2012, isto é, dados futuros considerando os utilizados para geração do modelo de previsão. Foi utilizado o modelo gerado anteriormente sobre o qual são aplicadas novas entradas de valores monitorização, com a finalidade de prever a necessidade de uma intervenção no mesmo equipamento.

Os resultados deste procedimento sob forma de *output* final são apresentados pelo sistema no formato da Figura 4-12.

Bosch
Tecnologia para a vida

Tabela de Predição

param_id	value	confidence(Não)	confidence(Mecânico)	confidence(Setup)	confidence(Elétrico)	confidence(Software)	prediction(Avaria Tipo)
10	200	1.000	0	0.000	0	0	Não
227	83	0.997	0.003	0	0	0	Não
11	200	1.000	0	0.000	0	0	Não
94	2	0.999	0.001	0.001	0	0	Não

param_id	value	confidence(Não)	confidence(Mecânico)	confidence(Setup)	confidence(Elétrico)	confidence(Software)	prediction(Avaria Tipo)
14	200	1.000	0	0.000	0	0	Não
83	248	0.001	0.999	0.000	0.000	0.000	Mecânico
105	40	1.000	0.000	0.000	0.000	0.000	Não
106	44	1.000	0.000	0.000	0.000	0.000	Não
04	149	0.999	0.001	0.000	0.000	0.000	Não
105	40	1.000	0.000	0.000	0.000	0.000	Não
106	44	1.000	0.000	0.000	0.000	0.000	Não

83	248	0.001	0.999	0.000	0.000	0.000	Mecânico
----	-----	-------	-------	-------	-------	-------	----------

Figura 4-12 - Exemplo dos alertas fornecidos pelo sistema.

Através da visualização deste *output*, a equipa de manutenção da empresa verifica a previsão de necessidade de efetuar uma intervenção na máquina em função do parâmetro e do respetivo valor monitorizado. Essa previsão é-lhe fornecida com o respetivo grau de confiança. Analisando a tabela, o responsável de manutenção visualiza, através das duas primeiras colunas, os identificadores monitorizados e o respetivo valor. Analisando a última coluna da tabela verifica-se a previsão obtida relativamente à não necessidade de efetuar uma intervenção preventiva no equipamento ou a previsão de um tipo de intervenção, seja ela, Mecânica, *Setup*, Elétrica ou *Software*. Conforme a previsão, o grau de confiança pode ser verificado na respetiva coluna relativa a cada tipo de previsão (colunas 3 a 7 da tabela).

A previsão evidenciada na Figura 4-12 diz respeito a uma das previsões obtidas para uma possível falha no equipamento utilizado neste caso prático. O responsável de manutenção visualiza o alerta e verifica que para o parâmetro 83, com um valor monitorizado de 248, existe a possibilidade de ocorrência de uma falha mecânica com uma certeza na previsão de 99.9% (grau

de confiança apresentado na coluna relativa à confiança da predição de uma falha mecânica). O modelo apresenta para esta predição uma possibilidade de 0,1% de não ocorrência de qualquer falha. Através da análise desta predição o responsável de manutenção poderá planejar a elaboração de uma ação de manutenção preventiva condicionada mecânica sobre o equipamento, baseada numa predição efetuada pelo módulo de predição, obtida com uma percentagem de certeza elevada e que proporciona informação necessária para atuar antes da possível ocorrência da falha prevista.

A qualquer altura o modelo preditivo poderá ser novamente gerado e refinado executando o processo 1 referente à geração do modelo de predição (Figura 4-3).

4.3.4 Validação e interpretação de resultados obtidos pelo modelo

A última fase do desenvolvimento de um sistema de predição consiste na fase de testes e refinamento, que envolve a validação e verificação do sistema, de modo a assegurar que: i) o sistema funciona corretamente, ii) forneça resultados verdadeiros e iii) satisfaça os requisitos do(s) parceiro(s) de sistema. Para além destes pressupostos, envolve a realização de eventuais alterações nos seus requisitos, realçando-se por exemplo a aquisição eficaz de dados e a disponibilização de resultados da forma mais perceptível possível. Uma das fases mais críticas no desenvolvimento de um sistema de conhecimento com estas características é a respeitante ao desenvolvimento da base de dados funcional, uma vez que a evolução temporal vai implicar alterações no sistema para que este continue a executar de forma eficaz as tarefas que estiveram na origem do seu desenvolvimento.

Caso um modelo seja válido, com base neste é possível efetuar previsões sobre outros conjuntos de dados, podendo os resultados obtidos serem comparados com os resultados reais, caso estes sejam conhecidos. Para validação dos resultados obtidos pelo modelo de predição gerado, realizou-se a contabilização das predições efetuadas pelos diferentes cenários de aplicação de diferentes operadores/ algoritmos de DM utilizados no módulo de predição, confrontando-as com intervenções reais efetuadas pelas equipas de manutenção da empresa à máquina analisada. A Tabela 4-4 apresenta a contabilização e comparação do número de intervenções reais realizadas e o número de intervenções previstas pelo sistema nos diferentes cenários de aprendizagem.

Tabela 4-4 - Contagem do número de ocorrências reais e as previstas pelo sistema.

		Número de Ocorrências				
		Elétrico	Mecânico	Setup	Software	Total
Ocorrências Previstas	Ocorrências Reais	6	136	9	1	152
	Redes <i>Naive Bayes</i>	3	117	7	1	128
	Árvores de Decisão	3	107	7	1	118
	Indução de Regras	5	101	6	1	113
	Redes Neurais Artificiais	2	74	4	0	80

Para o cálculo do número real de ocorrências efetuou-se a contagem das intervenções realizadas pelas equipas de manutenção e registadas na tabela de intervenções contendo registos entre 1 de setembro de 2012 e 31 de dezembro de 2012, enquanto para o cálculo do número de ocorrências previstas foi efetuada uma contagem dos resultados obtidos nos diferentes cenários, onde o melhor resultado é disponibilizados pela saída principal do sistema e apresentada na tabela da Figura 4-12. Os valores obtidos são o resultado da contabilização de todas as intervenções corretivas conforme a sua natureza (ou seja, Mecânica, Setup, Elétrica ou Software). Não considerando que uma ação de manutenção afeta diretamente a eventual necessidade de realizar uma nova intervenção, a Tabela 4-4 conjuntamente com a Figura 4-13 apresenta uma comparação dos dados de intervenções reais e os resultados gerados pelo modelo de previsão nos diferentes cenários.

A previsão obtida pelo modelo indica um menor número de intervenções do que as operadas na realidade. No total de todas as intervenções, a diferença entre a melhor previsão e a realidade é de aproximadamente 15,7%. Esta diferença poderá significar ser um erro de previsão e/ou a existência de um pequeno aumento de intervenções na empresa durante o período analisado, podendo ser fruto de uma degradação da máquina. Este facto poderá também estar relacionado com o intervalo de alerta calculado pelo responsável de manutenção da empresa aquando da definição dos valores máximos e críticos de alerta de monitorização que servem de suporte ao processo decisivo para efetuar uma determinada intervenção preventiva condicionada. Os resultados das predições efetuadas pelo modelo apresentado não considera esses condicionantes e por isso poderá não prever uma intervenção quando efetivamente foi levada a cabo por parte das equipas de manutenção da empresa, e que para o modelo prevê como não necessária.

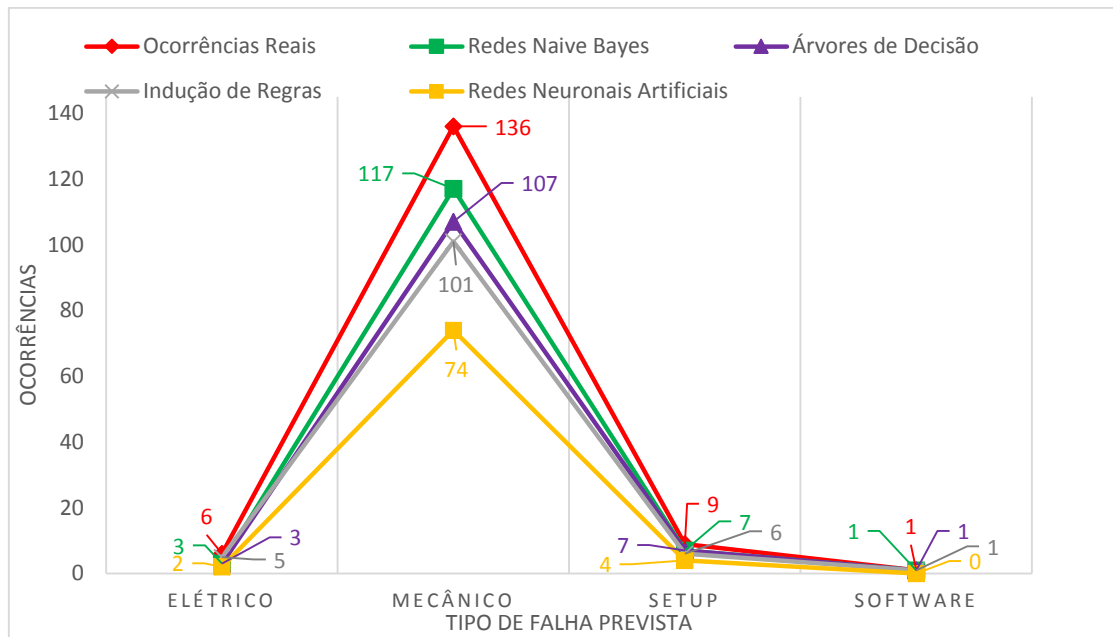


Figura 4-13 - Ocorrências reais efetuadas e as previstas pelo sistema.

Para auxílio na análise dos resultados obtidos com o módulo de predição do sistema apresentado, a Figura 4-14 apresenta uma forma alternativa de visualização da tabela de predição que o *RapidMiner* proporciona. Reporta uma alternativa de visualização dos resultados, *output* principal do sistema gerado com o modelo que melhor precisão obteve, com a vantagem de permitir indicar a existência de relação entre um determinado parâmetro monitorizado e a ocorrência de um determinado tipo de falha, independentemente do seu valor e relativamente à máquina utilizada no estudo.

O gráfico apresenta no eixo horizontal a predição de falha prevista pelo modelo e na coluna vertical a identificação do parâmetro monitorizado. Analisando a figura poder-se-á verificar, por exemplo, que o modelo indica que falhas mecânicas e/ou *software* estão diretamente relacionadas com os valores do parâmetro 84. Analisando os restantes parâmetros sobre os quais o modelo prevê uma possível falha, é possível verificar que todos eles se traduzem unicamente na probabilidade de ocorrência de um tipo de falha. As predições efetuadas apresentam uma relação entre o parâmetro 84 e a ocorrência de falhas mecânicas e de *software* enquanto, por exemplo, o parâmetro 227 se relaciona unicamente com a probabilidade de ocorrência de falhas mecânicas.

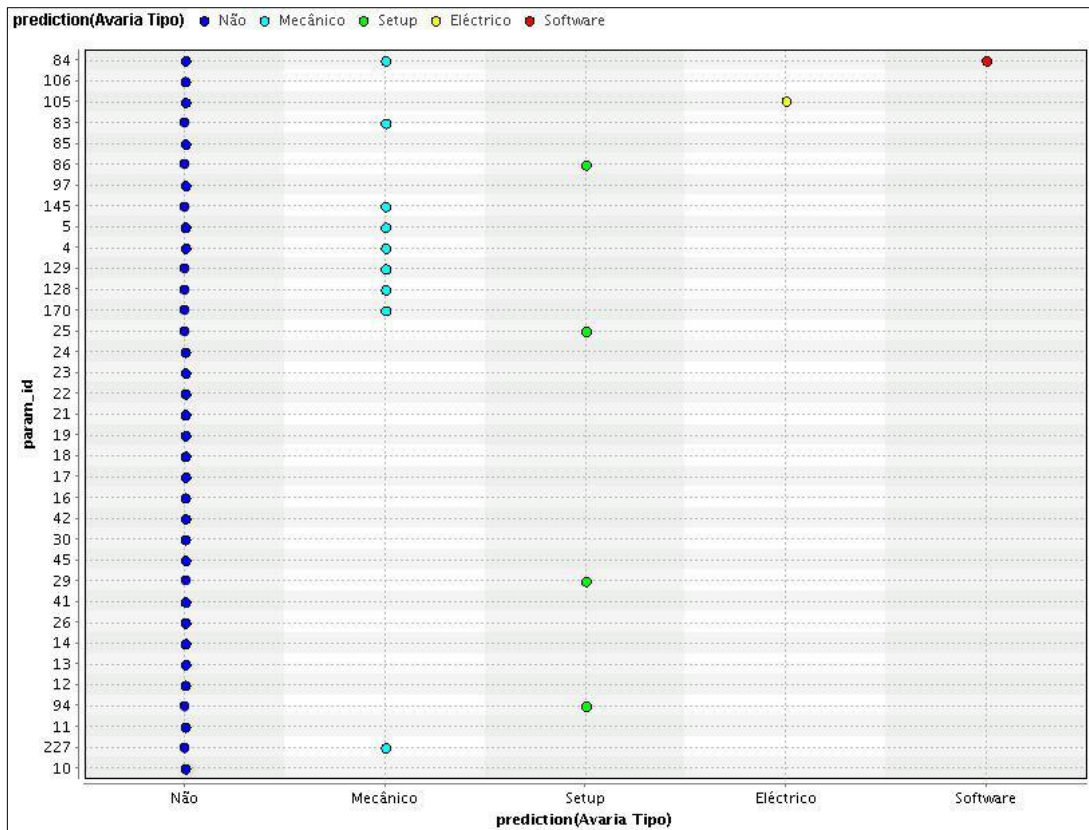


Figura 4-14 - Visualização gráfica da predição efetuada em relação ao parâmetro real monitorizado.

No caso prático aqui apresentado, a informação utilizada relativa ao equipamento analisado baseia-se em dados recolhidos e armazenados através de registo de monitorização contínua e de ações de manutenção que serviram de suporte ao objetivo principal da manutenção preditiva e consequentemente do sistema proposto, a verificação pontual do funcionamento do equipamento antecipando eventuais problemas que possam causar gastos maiores como a manutenção corretiva. Em suma, a construção do modelo de predição assentou em dados relativos à monitorização contínua e nos respetivos instantes em que foram efetuadas intervenções, utilizando o registo do valor do parâmetro responsável pelo alerta real, e assim, incorporar o conhecimento inerente ao instante de atuação. Desta forma, a utilização pelo sistema de uma base de dados inicial contendo esse mesmo apuro e que serviu de suporte à criação do modelo preditivo embecendo todo o *know-how* adquirido, traduziu-se em predições com elevada precisão, e posteriormente validadas através da confrontação das predições obtidas pelo modelo de predição com os acontecimentos reais, utilizando para tal dados relativos a um período de tempo alargado.

Estes resultados não foram considerados para reformulação dos planos de manutenção da empresa, mas com modelos gerados integrando dados atuais, as previsões geradas com um

elevado grau de certeza acerca da ocorrência de uma falha, poder-se-ão traduzir na realização de futuras intervenções.

4.3.5 Disponibilização de resultados por parte do sistema

A interação entre um utilizador e um sistema computacional com essa natureza, define-se como o processo de comunicação entre estes dois elementos e que é gerido por um terceiro elemento preponderante, o *interface* (Pairo et al., 2013). O *interface* assume-se como uma fronteira entre o humano e o computador. Permite representar o comportamento de um sistema computacional interativo e proporciona ao utilizador os meios necessários para controlar, operar ou simplesmente aceder a esse mesmo sistema (Meech, 1999).

Nesta subsecção apresenta-se o *interface* do sistema relativo à atividade A313 – Disseminação de Conhecimento, descrita anteriormente na subsecção 3.2.2 deste documento. Respeitando os princípios hierárquicos das aplicações e das organizações onde operam, também neste trabalho os *outputs* fornecidos pelo sistema de predição estão sujeitos a acessos controlados através do portal de suporte apresentado anteriormente. Transversalmente o acesso realiza-se através de um *interface* responsável pelo *login* por parte de qualquer parceiro integrante do sistema, que após validação estabelece os parâmetros de navegação (Figura 4-15).

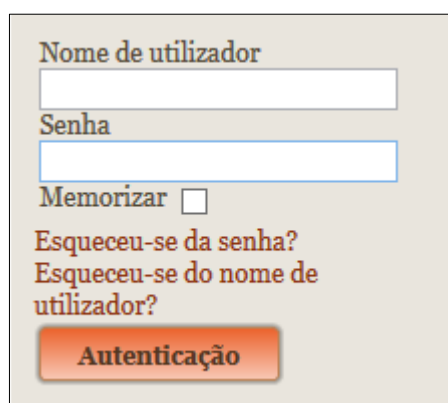
A interface de autenticação do sistema é apresentada num formulário com fundo cinzento claro. No topo, o texto "Nome de utilizador" precede um campo de entrada branco. Abaixo dele, o texto "Senha" precede outro campo de entrada branco. Segue-se o texto "Memorizar" acompanhado de uma caixa de seleção vazia. Abaixo disso, há duas linhas de texto: "Esqueceu-se da senha?" e "Esqueceu-se do nome de utilizador?". No fundo do formulário, há um botão retangular com o texto "Autenticação" em letras brancas sobre um fundo laranja escuro.

Figura 4-15 - Interface de autenticação no sistema

Desta forma, através da autenticação, cada parceiro terá acesso aos *outputs* gerados, assegurando assim acessos controlados à informação inerente a cada interveniente.

A Figura 4-16 apresenta o *interface* que um parceiro obtém após a sua autenticação no sistema, estando aqui exemplificado o *interface* do parceiro de projeto utilizado para este caso prático.



Figura 4-16 - Página obtida após autenticação.

Como observado na figura, cada empresa (por exemplo a Bosch) acede a um *interface* personalizado com facilidade de interação, onde visualiza os modelos gerado pelo módulo de predição e as predições efetuadas pelos mesmos.

4.4 Resultados e discussão

O desenvolvimento do protótipo do sistema SPAMUF obedeceu a um conjunto de etapas coerentes encadeadas da seguinte forma: i) definição dos requisitos do sistema, ii) definição, especificação e esquematização da arquitetura, iii) desenvolvimento da base de conhecimento (i.e., geração dos modelos de predição), e iv) implementação do *interface* com o utilizador.

Foram aplicadas neste trabalho metodologias, técnicas e algoritmos de DM com o objetivo de prever a possibilidade de ocorrências de falhas em equipamentos existentes em linhas de produção de uma unidade fabril incorporada no sistema, sendo nesta fase utilizada apenas uma máquina entre o conjunto de máquinas monitorizadas, garantindo-se uma amostra de dados representativa do universo de estudo.

O desenvolvimento deste protótipo permitiu testar a conceptualização da incorporação do DM num sistema inteligente para predição de falhas em máquinas de unidades fabris globalmente dispersas. Para tal recorreu-se a uma ferramenta de DM (i.e., *RapidMiner*) capaz de proporcionar a concretização e incorporação de um modelo de predição que permite a integração de componentes de DM em aplicações de *software*, independente de plataformas e fontes de dados,

possibilitando uma futura utilização do modelo desenvolvido numa outra plataforma através da especificação PMML.

Os resultados alcançados permitiram validar: i) a abordagem proposta, e ii) a utilização de modelos de aprendizagem automática para a resolução de problemas complexos na área de Engenharia Industrial e Sistemas.

A abordagem proposta integra na fase de modelação duas tarefas: i) segmentação dos dados, e ii) geração do modelo de predição. Através desta abordagem foi possível incrementar o grau de homogeneidade dos casos, o que se traduz na geração de um modelo de previsão com um grau de certeza mais elevado. Considerando que na área em estudo a acuidade preditiva é um fator fulcral, este procedimento assume-se como crítico e capaz de elevar os níveis de desempenho na disponibilização de equipamentos e em consequência de produtividade.

Ultrapassadas as dificuldades de arranque, tal como referido supra nesta secção, a utilização de modelos de aprendizagem automática assentou na geração de um modelo de predição onde foram aplicados e validados diferentes algoritmos para DM preditivo, criando diferentes cenários de aprendizagem com objetivo de criação do respetivo modelo com a maior precisão possível. Verificadas variadas métricas de validação dos modelos gerados mas fundamentalmente assentando na precisão do modelo obtido, optou-se pela geração do modelo final utilizando Redes *Naive Bayes*. Assim o estudo envolveu a elaboração do módulo de predição assente em dois processos do *RapidMiner* com os seguintes objetivos: i) segmentação dos dados e geração do modelo de predição; e ii) predição de conhecimento através do modelo gerado no ponto i).

O modelo desenvolvido no melhor dos cenários testados apresenta índices de confiança elevados, onde a realização de uma etapa inicial de segmentação dos dados que permitiu identificar e criar dois segmentos homogéneos no conjunto de dados disponíveis, gerando regras de auxílio ao processo decisivo baseadas em parâmetros monitorizados de forma contínua, se traduziu de vital importância.

O pragmatismo de trabalhos de natureza académica impõe limites temporais, consonantes e coerentes com os objetivos definidos e recursos, nomeadamente de tempo, disponibilizados para o efeito. A otimização é inerente ao conceito de inacabado e como tal são diversas as perspetivas que emergem quando se folga o aperto das fronteiras do estudo. Nesse sentido salienta-se o facto de não se ter tido a oportunidade de incorporar no estudo dados potencialmente importantes cuja consciencialização se tornou mais evidente com a progressão das etapas e que por exemplo dizem respeito à monitorização contínua de fatores possivelmente relevantes para o trabalho passíveis

de auxiliar a interpretação e validação dos resultados obtidos: i) informação relativa à idade dos equipamentos, ii) tempo de funcionamento dos equipamentos, considerando os tempos de paragem dos mesmos, podendo desta forma estarem sujeitos a um maior número de anomalias ao longo do tempo condicionando desta forma a sua fiabilidade, e iii) o registo de fatores ambientais tais como temperatura ambiente e humidade do ar, passíveis de condicionar de igual forma o normal funcionamento dos equipamentos. Este tipo de informação não foi considerada, visto o parceiro de projeto não efetuar até ao momento a monitorização e registo deste tipo de fatores. Este acréscimo de informação permitiria adicionar mais campos relevantes à geração do modelo preditivo, e como tal, a obtenção de resultados mais consistentes e precisos dos putativos eventos previstos no equipamento. A inclusão deste tipo de dados proporcionaria a inclusão de novos atributos no processo de geração do modelo de predição decorrente do processo 1, traduzindo-se num modelo mais robusto, uma vez que assenta num maior número de parâmetros de entrada, e conseqüentemente em resultados mais abrangentes a nível informativo. A inclusão de um maior número de parâmetros permitirá encontrar novas relações entre parâmetros utilizando um leque mais alargado de atributos, proporcionando aos responsáveis de manutenção alertas de intervenção, conjugadas com informação relativa à forma como se relacionam esses mesmos atributos.

A concretização do caso prático implicou a necessidade de definir uma linha de atuação por forma a atingir resultados relevantes, quer ao nível da elaboração do modelo de predição, quer da posterior geração de predições, com a conseqüente comparação com ocorrências reais, tendo sido necessário tomar decisões relativamente aos seguintes fatores:

- i) o número de equipamentos a incluir no estudo de forma a criar uma base de dados representativa para a criação do modelo e validar posteriormente o mesmo;
- ii) a necessidade de incorporar no estudo um número significativo de registos de ações de manutenção, independentemente da sua natureza, e assim garantir uma base de conhecimento robusta para aplicação dos algoritmos de DM;
- iii) a janela de dados a utilizar de forma a trabalhar com um histórico de dados de monitorização que possuam refinamentos efetuados pelo parceiro no que diz respeito ao valor de parâmetro responsável pelo alarme referente à deteção de uma falha.

Além destes fatores, um condicionalismo afetou de igual forma a linha de atuação, a limitação da capacidade do sistema informático sobre o qual está implementado o módulo preditivo do protótipo do sistema.

Considerando os fatores orientadores do estudo e o condicionalismo apresentado, utilizaram-se dados relativos a um equipamento, obtidos num período específico e relevante, quer para a geração do modelo de predição, quer para as predições geradas e validadas com ocorrências reais. Desta forma os dados utilizados para geração do modelo incorporaram um conjunto alargado de registos contendo um histórico relevante, quer de ações de manutenção, quer de monitorização. A não incorporação de mais equipamentos nesta fase do estudo, foi condicionada pela não existência de um campo agregador na base de dados utilizada. Este campo seria responsável pela identificação do tipo de equipamento e proporcionaria um agrupamento correto na elaboração do correspondente *cluster*. O agrupamento de equipamentos de diferentes tipos ou com características distintas, incorporaria valores iguais de monitorização relativos aos mesmos parâmetros, mas no entanto, indicadores de distintos estados de funcionamento de cada um dos equipamentos/itens.

Também ao nível de melhorias a concretizar, identifica-se a possibilidade da incorporação futura no sistema de um novo equipamento/item, diferente de todos os existentes na atual configuração. Tal como abordado inicialmente neste capítulo, uma gestão cuidada da base de dados de suporte ao processo preditivo é um dos pontos cruciais ao bom desempenho dos processos de DM gerados e com isso a uma atuação adequada dos mesmos no processo preditivo. Assim sendo focou-se a periodicidade de reformulação da base de dados de forma a atualizar mais convenientemente o histórico de intervenções e monitorização aquando da incorporação de mais equipamentos/itens relativos a parceiros novos ou já existentes no sistema, no entanto, sempre que se dá entrada de um novo equipamento/item é fundamental iniciar-se a recolha de dados referentes ao mesmo, não obedecendo inicialmente às periodicidades estabelecidas relativamente à gestão da base de dados já existente e desta forma criar um histórico capaz de ser incorporado no sistema e com isto garantir o seu normal funcionamento.

Como o protótipo do sistema criado é baseado num princípio fundamental, a existência de uma rede colaborativa, este vai embeber no seu seio diferentes conhecimentos e realidades funcionais dos parceiros inseridos no mesmo. Desta forma, o aprimorar de *know-how* de uns parceiros relativamente a outros no processo de deteção da falha e de tomada de decisão referente a tempos de atuação à mesma, proporcionado por estados mais avançados de refinamento na forma de a enfrentar, tronar-se-á uma mais-valia para o sistema. Este comportamento, que podemos considerar como de referência, irá beneficiar e ser beneficiado ao ser embebido na geração dos modelos de predição, já que estes embeberão as melhores práticas e conduzindo a um

desempenho superior. Este é visto como uma ação de *benchmarking*, através do qual um parceiro faz uso do conhecimento de outro(s) na realização de uma função específica a fim de melhorar o seu desempenho. Assim sendo, parceiros que se encontrem em estados mais atrasados no respeitante à deteção da falha e conseqüente tempo de atuação, irão beneficiar do conhecimento mais aprimorado de outros e desta forma permitir que a partilha de informação proporcione modelos de predição mais precisos e eficazes.

No entanto apesar de todos os condicionalismos, os resultados obtidos e confrontados com os registos reais de ações de manutenção corretivas e preventivas condicionadas levadas a cabo pelo parceiro abrem portas ao desenvolvimento de sistemas inteligentes para predição de avarias, embecendo os modelos desenvolvidos para predição de falhas independentemente do equipamento e parâmetro monitorizado.

De igual forma, não faria sentido não tomar em consideração um aspeto crucial em todos os domínios da gestão industrial que são os custos intrínsecos às diferentes atividades diariamente executadas pelas organizações. No contexto deste trabalho, numa manutenção aprimorada, será relevante sintonizar decisões que considerem conhecimento sobre os custos provenientes das atividades de manutenção, não só dos derivados da necessidade de manter as características técnicas e da disponibilidade dos sistemas dentro dos parâmetros aceitáveis, como também, em função da eficiência associada à relação custo benefício, a fim de os serviços prestados se constituírem como um diferencial competitivo para uma qualquer empresa inserida no sistema. Neste sentido, poder-se-á registar que a melhoria da rentabilidade do sector da manutenção não pode prescindir de uma gestão cuidada das intervenções em cada equipamento/item. Por não ter sido possível obter qualquer tipo de informação desta natureza, aquando do acesso à informação, seria oportuno que para melhor desempenho do sistema, se reunissem condições para proceder à recolha deste tipo de dados, com o intuito de, posteriormente, se proceder ao apuramento de indicadores relacionados com os custos da manutenção, devendo-se, neste âmbito, proceder à sua definição, identificação, classificação e respetiva análise.

Finalizando, através da explanação efetuada, focando a variedade de pontos fortes existentes no sistema apresentado, bem como as debilidades presentes e a forma de as superar em futuras abordagens ao sistema proposto, os resultados obtidos abrem portas ao desenvolvimento de sistemas inteligentes para predição de avarias em equipamentos de unidades fabris globalmente dispersas, embecendo modelos desenvolvidos para previsão da ocorrência de falhas em equipamentos/itens sujeitos a monitorização de variados parâmetros.

Capítulo 5

5. Conclusões e perspectivas de trabalho futuro

Ao concluir-se um projeto, é tido por adequado salientar-se as conclusões que foram alcançadas, confrontando os resultados obtidos com os objetivos estabelecidos inicialmente. Considera-se também fundamental uma postura crítica, relativamente às limitações de que entretanto se foi tomando percepção. No sentido de ultrapassar essas limitações, bem como de tentar levantar alguns dos pressupostos fixados nesse mesmo projeto, devem-se apontar perspectivas de trabalho futuro.

Os processos de decisão que visam, em gestão da manutenção, definir práticas proveitosas no sentido de assegurar a eficácia e o controlo adequado dos ativos físicos da produção, dependem de sobremaneira da informação na qual os decisores se baseiam. Muita dessa informação pode ser retirada do registo histórico de avarias e monitorização, a qual depois de transformada eficiente e eficazmente em indicadores relativos à necessidade de intervenções preventivas condicionadas a fim de debelar uma possível falha de um equipamento/item, proporciona conhecimento que ao ser devidamente utilizado, pode contribuir para um acréscimo do desempenho do sector da manutenção. Tal melhoria repercutir-se-á nos resultados do processo produtivo e consequentemente na geração de riqueza para a empresa.

Com este trabalho foi apresentada uma abordagem inovadora, com a convicção de que a via explorada constitui um contributo relevante para o desenvolvimento de sistemas inteligentes para predição de avarias na área da manutenção industrial, utilizando um leque de tecnologias consolidadas e em constante evolução de forma a auxiliar o processo decisivo relativo a capacidade de prever a necessidade de efetuar intervenções de manutenção preventivas.

Foi definida e esquematizada toda uma rede colaborativa capaz de gerir a integração e saída de parceiros no sistema, garantindo confidencialidade nos resultados apurados bem como a segurança nos acessos aos dados que servem de alimento ao sistema. Iniciou-se com a definição do sistema através do estabelecimento de objetivos inerentes ao mesmo, a definição dos requisitos para participação dos potenciais interessados no seu ingresso, bem como na definição estrutural de base do mesmo. Focalizou-se igualmente a esquematização e parametrização do módulo operacional do sistema e responsável pelo *interface* e gestão de comunicação do mesmo, bem como do módulo preditivo.

No decorrer dos trabalhos conducentes à presente tese, ao analisar a informação disponibilizada, enfrentou-se uma contrariedade que dificultou o seu progresso. A informação recolhida no registo histórico de dados apresentava vários tipos de incongruências pois para além de não se encontrar estruturada da melhor forma, enfrentaram-se variadas vezes, situações de falta de dados, informação dúbia, incompleta, inconsistente, ou até mesmo incorreta.

Para que os resultados obtidos possam proporcionar contribuições de maior valor potencial para a tomada de decisão, será indispensável dispor de informação abrangente e colecionada de forma homogénea e criteriosa. Seria necessário e oportuno que, além da coerência e rigor indispensável à inserção da informação no sistema, se efetuasse de igual forma uma menção mais específica ao tipo de falha registada, gerando uma maior granularidade dos dados relativamente à sua identificação e assim tornar mais específica a predição efetuada. Por esta via será possível tecer considerações mais consistentes e específicas acerca dos eventos ocorridos nos equipamentos bem como elaborar previsões mais assertivas.

No entanto os resultados obtidos permitem demonstrar que as técnicas de DM se posicionam como uma ferramenta poderosa e adequada no apoio à tomada de decisão por parte dos responsáveis de manutenção. Com base em dados históricos relativos a ações de manutenção, conjugados com dados de monitorização contínua, é possível prever claramente ações futuras, principal objetivo do sistema de controlo aqui apresentado, através da seleção e utilização de parâmetros específicos de entrada. Os resultados alcançados comprovam a viabilidade da aplicação e desenvolvimento do processo de descoberta de conhecimento em base de dados na área da manutenção industrial, seguindo uma abordagem que privilegia a utilização de regras *Naive Bayes* na previsão da ocorrência de uma falha, previsão esta baseada na criação de um modelo de predição assente em ocorrências previamente conhecidas. A geração do modelo de predição é baseada em critérios de obtenção do melhor grau de certeza, sendo aplicados diferentes algoritmos de DM tais como AD, RNA, IR e RB, optando-se pela criação do modelo de predição com o algoritmo que apresenta melhores resultados.

Em alternativa aos métodos convencionais de gestão da manutenção (apresentados na secção 2.2) e ao sistema implementado e em uso pelo atual parceiro de projeto, surgem os modelos de DM. Neste trabalho foram aplicados diferentes algoritmos de IA aplicados ao DM (todos eles descritos na subsecção 2.4.4) obtendo-se melhores resultados com a aplicação de *Regras Naive Bayes* (descrita em pormenor na subsecção 2.4.4.3). Assim, todo o processo de geração do modelo preditivo assenta sobre a geração de regras aplicando o algoritmo selecionado.

No decorrer do processo de DCBD, foram encontrados alguns obstáculos, relacionados com o grande número de registos relativos a dados de monitorização e o número de casos relativos a intervenções corretivas e preventivas a esses mesmos equipamentos. Contudo, mesmo com um número existente de registos de intervenções, desde que esteja de alguma forma garantida que a amostra é representativa do universo a estudar, é possível induzir um padrão genérico de comportamento. Devido à introdução de novos equipamentos e conseqüentemente às possíveis modificações da infraestrutura de recolha de dados, à instalação de novas aplicações e de novos agentes de recolha de dados e à possibilidade da necessidade de compatibilização com a estrutura funcional da base de dados existente, torna difícil a gestão coerente do sistema.

O modelo gerado no módulo de predição e a incorporação deste no sistema criado, provaram ser eficazes na predição de falhas de um equipamento podendo desta forma ser efetuada uma intervenção antes da possível ocorrência de falha. Comprovou-se que com implementações cuidadas, seguindo a metodologia e abordagem apresentada, o DM pode contribuir para a resolução de problemas complexos na área da manutenção industrial.

Face aos resultados alcançados foi desenvolvida uma primeira versão de um sistema de predição de avarias em máquinas de unidades fabris globalmente dispersas (SPAMUF), englobando dados reais inerentes a ações de monitorização e intervenções de manutenção de uma empresa internacional e parceira de projeto, embebendo no mesmo, modelos de predição baseados em regras geradas através da aplicação de algoritmos *Naive Bayes*.

Este trabalho apresenta soluções e propostas com o principal objetivo de definir mecanismos flexíveis e poderosos de forma a auxiliar a tomada de decisão. As soluções incidiram sobre os principais tipos de manutenção industrial, associadas à utilização de ferramentas, metodologias e técnicas de DM, quer na perspetiva de análise, planeamento, gestão, programação e utilização de um protótipo de um sistema. Os resultados obtidos com este protótipo permitem definir uma base de trabalho para o desenvolvimento futuro de um sistema inteligente de predição de avarias em tempo real, tomando como guião as ferramentas e tecnologias descritas e utilizadas.

5.1 Contribuições

Os contributos deste trabalho podem ser agrupados em duas áreas: a Engenharia de Produção e Sistemas, mais propriamente na área da Manutenção Industrial e nas Tecnologias e Sistemas de Informação (TSI).

Para a área de aplicação da Engenharia de Produção e Sistemas os contributos podem ser enumerados da seguinte forma: (i) geração de modelos de previsão da ocorrência de uma falha em determinado equipamento com níveis elevados de precisão; e (ii) protótipo SPAMUF. O sistema SPAMUF coloca à disposição dos responsáveis de manutenção um ambiente virtual de alerta de intervenções, permitindo obter uma série de alertas de intervenções com a finalidade de reduzir os tempos de quebra produtiva derivada à falha de determinados equipamentos. Através deste ambiente de simulação é possível gerar modelos de predição de forma a sustentar a elaboração de intervenções preventivas condicionadas e não programadas sobre os equipamentos. O protótipo do Sistema SPAMUF desenvolvido neste trabalho abre portas ao desenvolvimento de um sistema de predição em tempo real, englobando um conjunto alargado de organizações globalmente dispersas.

Relativamente à área de TSI podem listar-se os seguintes contributos: (i) aplicação de técnicas de DM a uma nova área (predição de avarias); (ii) criação do sistema SPAMUF – Sistema de Conhecimento baseado em DM para a predição de avarias, com incorporação de técnicas automáticas para aquisição de dados e visualização de resultados.

Este trabalho contribui com a especificação de uma arquitetura de referência original. De uma forma inovadora, propõe-se uma estrutura funcional de suporte às atividades a desenvolver no âmbito do processo preditivo de falha e consequente predição de intervenção. Para tal recorre-se:

- Especificação das atividades a desenrolar em todo o processo;
- Especificação dos fluxos de informação que relacionam cada atividade;
- Especificação dos fluxos de informação que relacionam as atividades com os parceiros potenciais e presentes no sistema.

Embora esta Tese se concentre maioritariamente no processo preditivo inerente à geração e aplicação de modelos de DM, houve a necessidade de especificar todo o processo de formação, formulação e dissolução de parcerias, no sentido de se obter coerência e consistência ao longo de todo o processo de funcionamento do sistema. Em cada uma dessas fases foram desenvolvidos procedimentos que permitem que o processo preditivo se desenrole de acordo com o idealizado.

Ao longo de todo o processo de consulta bibliográfica sobre manutenção preditiva e de elaboração de todo o trabalho prático relativo a esta tese, ficou patente a dificuldade que assiste à procura de parceiros com vista a integrarem sistemas de predição de avarias relativas a diferentes organizações em simultâneo.

Com o objetivo de testar as funcionalidades do sistema especificado e garantir que com base nessas especificações é possível construir um sistema real que vise a integração num sistema de predição de avarias em máquinas de unidades fabris globalmente dispersas, foi construído o protótipo SPAMUF. Este protótipo foi submetido a alguns testes que revelaram a sua funcionalidade, deixando também presentes algumas das suas limitações.

O SPAMUF pretende ser uma ferramenta importante no processo preditivo de avarias, implementada com base nas especificações produzidas no capítulo 3. Sendo essas especificações originais e tendo este protótipo seguido essas especificações, resulta aceitável considerar esta aplicação como sendo um contributo original.

5.2 Desafios e perspetivas de trabalho futuro

O sistema proposto baseia toda a sua funcionalidade num adequado desempenho de todos os parceiros integrantes e estruturas funcionais previstas. Nesse sentido, e considerando que a construção e aplicação do protótipo do sistema incidiu apenas sobre uma empresa parceira de projeto e uma máquina analisada, será de todo conveniente e interessante alargar o teste prático a todas as máquinas integrantes do sistema produtivo da empresa e numa fase posterior integrar mais parceiros no esquema funcional do projeto. Nomeadamente, afigura-se muito interessante a tarefa de implementar a base de dados relacional apresentada e adotada neste trabalho, tornando-a transversal a todas as organizações interessadas em incorporar os seus dados no sistema evitando desta forma criar modelos de DM individualizados e partindo desta forma para um sistema global e uniformizado.

Um dos objetivos com que se partiu para a realização deste trabalho consistia em conseguir-se lidar com o dinamismo inerente aos diferentes esquemas funcionais de manutenção passíveis de incorporar no sistema e à conseqüente necessidade de modificações constantes na estrutura do protótipo de predição. Na elaboração deste trabalho era necessário ter em conta estes dois fatores, no entanto devido ao facto de se ter integrado unicamente dados referentes a atividades de monitorização e manutenção de uma só empresa, os problemas inerentes à uniformização dos dados analisados não foi tido em conta.

Presentemente, no contexto do criado e desenvolvido, o sistema elaborado com este trabalho foi testado e validado recorrendo à utilização de uma máquina, sujeita a um processo de monitorização contínua. Foram monitorizados um leque alargado de parâmetros, traduzindo-se

com isso, na simulação de atuação do sistema gerado num cenário operativo real, confrontando desta forma os resultados atingidos com intervenções reais levadas a cabo por parte das equipas de manutenção ao equipamento.

A melhoria do sistema passa pela integração da base de dados projetada e desta forma comparar os resultados obtidos com as reais ocorrências, tal como o efetuado durante o trabalho aqui apresentado.

Sendo mantidas a precisão das predições obtidas, será necessário dar o próximo passo relativo à integração de máquinas similares à utilizada neste caso prático e confrontar com as reais ações de manutenção levadas a cabo por parte do parceiro de projeto.

A ambição de criar um sistema global mantém-se no horizonte de curto prazo onde o trajeto implementado está a criar condições para que isso se torne possível e sustentável.

Um estudo interessante poderá passar pela especificação pormenorizada das bases de dados a incorporar no sistema por parte de diferentes organizações contribuindo desta forma para o incremento da robustez da solução global apresentada.

Com a possível inclusão de novos atributos de estudo, a aplicação de diferentes algoritmos de DM deverá ser tomada em conta. Dependendo do tipo de atributos a incorporar, outros algoritmos de DM preditivo poderão produzir resultados ainda mais exatos. Assim, a inclusão de novos atributos, quer relativos à monitorização quer relativos a ações de manutenção corretivas e preventivas, permitirão o desenvolvimento de novos modelos de predição de avarias, bem como criação de modelos para reconhecimento de padrões de dados.

A evolução da base de dados a incorporar, bem como dos diferentes atributos incorporados poderá servir de impulso para um novo tipo de estudo onde a predição de falha poderá ser baseada no estudo de vários atributos em simultâneo, através do estudo de relações existente entre eles e assim evoluir de uma predição por atributo para uma predição por grupo de atributos.

Conforme se pode constatar existe um longo caminho a percorrer. Além das limitações, foram também apresentadas perspetivas e sugestões de novos desafios, alguns dos quais poderão adquirir dimensão e abrangência para novos trabalhos de Doutoramento ou Mestrado, permitindo assim que este trabalho progrida, dando origem a um sistema cada vez mais completo, refinado e útil para as organizações.

A concretização do objetivo proposto não foi um processo imediato. No seu prosseguimento foram alcançadas contribuições parcelares ou colaterais que de seguidamente são apresentadas:

Bastos, Pedro; Lopes, Isabel; Pires, Luís (2013) - Application of data mining in a maintenance system for failure prediction. In Advances in Safety, Reliability and Risk Management. ESREL 2013 29 Set – 2 Oct, Amsterdam, Holland. Safety, Reliability and Risk Analysis: Beyond the Horizon – Steenbergen et al. (Eds), Taylor & Francis Group, p. 933-970, ISBN 978-1-138-00123-7

Bastos, Pedro; Lopes, Isabel; Pires, Luís (2012) - A Maintenance Prediction System using Data Mining Techniques. Proceedings of the World Congress on Engineering 2012 Vol III. WCE 2012, July 4 - 6, 2012, London, U.K. ISBN: 978-988-19252-2-0

Bastos, Pedro; Lopes, Isabel; Pires, Luís (2011) - SPAMUF: a behaviour-based maintenance prediction system. In Advances in Safety, Reliability and Risk Management: 20th European Safety and Reliability. ESREL 2011- Troyes, France. London: Taylor & Francis. p. 977-982. ISBN 978-0-415-68379-1

Lopes, Isabel; Pires, Luís; **Bastos, Pedro;** "A decentralized predictive maintenance system based on data mining concepts"; First International Conference on Business Sustainability. Management, Technology and Learning for Individuals, Organizations and Society in Turbulent Markets, June 25-27, (2008), Ofir, Esposende – Portugal

Bibliografia

- Adair-Heeley, C. (1989) 'The JIT challenge for maintenance', *Production & Invention Management Review with APICS News*, 9(9), 34-35.
- Akthar, F. and Hahne, C. (2012) *RapidMiner 5 Operator Reference*.
- Alves, P. A. V. (2007) *E-generation: especificação de uma arquitetura para Intranets educacionais baseada em agentes*, unpublished thesis Universidade do Minho.
- Apte, C., Liu, B., Pednault, E. P. D. and Smyth, P. (2002) 'Business Applications of Data Mining', *ACM*, 45(8), 49-53.
- Artuso, P., Gammon, R., Orecchini, F. and Watson, S. J. (2011) 'Alkaline electrolyzers: Model and real data analysis', *International Journal of Hydrogen Energy*, 36, 7956-7962.
- Arunraj, N. S. and Maiti, J. (2006) 'Risk-based maintenance - Techniques and applications', *Journal of Hazardous Materials*, 142(3), 653-661.
- Ascher, H. and Feingold, H. (1984) *Repairable Systems Reliability: Modeling, Inference, Misconceptions and their Causes*, Texas: Marcel Dekker.
- Banks, J., Reichard, K. and Drake, M. (2008) *System Reliability and Condition Based Maintenance*, translated by Las Vegas: 423 - 427
- Bansal, D., Evans, D. and Jones, B. (2005a) 'Application of a real time predictive maintenance system to a production machine system.', *International Journal of Machine tools and manufacture*, 1210-1221.
- Bansal, D., Evans, D. and Jones, B. (2005b) 'A reverse algorithm for the real time predictive maintenance system.', *International Journal of Machine tools and manufacture*, 1-11.
- Bansal, D., Evans, D. J. and Jones, B. (2004) 'A real-time predictive maintenance system to a production machine systems.', *International Journal of Machine tools and manufacture*, 44, 759-766.
- Bansal, D., Evans, D. J. and Jones, B. (2006) 'BJEST: A reverse algorithm for the real time predictive maintenance system.', *International Journal of Machine tools and manufacture*, 46, 1068-1078.
- Barlow, R. E. and Hunter, L. (1960) 'Optimum preventive maintenance policies', *Operations Research*, 8, 90-100.
- Barlow, R. E. and Proschan, F. (1965) *Mathematical theory of reliability*, New York: John Wiley and Sons.
- Batanov, D., Nagarur, N. and Nitikhumkasem, P. (1993) 'Expert-MM: A Knowledge Based System for Maintenance Management', *Artif. Intell. Eng.*, 8, 283-291.
- Bateman, J. (1995) 'Preventive maintenance Stand alone manufacturing compared with cellular manufacturing', *Industrial Management*, 37(1), 19-21.
- Bellifemine, F. L., Caire, G. and Greenwood, D. (2007) *Developing Multi-Agent Systems with JADE*, ISBN:978-0-470-05747-6.

- Berry, M. J. A. and Linoff, G. (1993) *Mastering Data Mining: The Art and Science of Customer Relationship Management*, USA: John Wiley & Sons, Inc.
- Berson, A., Smith, S. and Thearling, K. (2000) *Building Data Mining Applications for CRM*, USA: McGraw-Hill.
- Bevilacqua, M., Ciarapica, F. E. and Paciarotti, C. (2012) 'Business Process Reengineering of emergency management procedures: A case study', *Safety Science*, 50, 1368-1376.
- Blackmore, J. W. (2001) 'Keys to CMMS success at Chelan county PUD', *Power Engineering Society Winter Meeting*, 2, 573-576.
- Bloch, H. P. and Geitner, F. K. (1997) *Machinery Failure Analysis and Troubleshooting*, Houston: Gulf Publishing Co., ISBN:0884156621
- Bordini, R. H., Hübner, J. F. and Wooldridge, M. (2007) *Programming Multi-Agent Systems in AgentSpeak using Jason*, ISBN:978-0-470-02900-8.
- Bradshaw, J. M. (1997) *An introduction to software agents*, ISBN:0-262-52234-9
- Bray, T., Paoli, J., Sperberg-McQueen, C. M. and Maler, E. (2000) Extensible Markup Language (XML) 1.0 (Second Edition), email to [accessed
- Browne, J. (1996) 'Guest Editorial - Extended Enterprises, INTRANETS and SMES.', *Production Planning and Control*, 8, 207.
- Buntine, W. L. (1994) 'Operations for learning with graphical models,'" , *Journal of Artificial Intelligence Research*, 2, 159-225.
- Cai, Z., Sun, S., Si, S. and Yannou, B. (2011) 'Identifying product failure rate based on a conditional Bayesian network classifier', *Expert Systems with Applications*, 38, 5036-5043.
- Campbell, J. D. (1999) *The Reliability Handbook – Plant engineering and maintenance*, Clifford/Elliott Publications.
- Carnero, M. C. and Novés, J. L. (2006) 'Selection of computerised maintenance management system by means of multicriteria methods', *Production Planning & Control*, 17(5), 335-354.
- Carvalho, S. (2008) *Recolha e Distribuição Automática de Informação/Dados de Suporte a um Sistema de Alerta Precoce de Catástrofes*, unpublished thesis Universidade de Trás-os-Montes e Alto Douro.
- Chan, F. T. S., Lau, H. C. W. and Ip, R. W. L. (2005) 'Implementation of total productive maintenance: A case study.', *International Journal of Production Economics*, 95(1), 71-94.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000) *CRISP-DM 1.0 - Step-by-step Data mining guide*, CRISP-DM Consortium.
- Chester, M. (1993) *Neural Network - A tutorial*, USA: PTR Prentice-Hall, Inc.
- Chidambaram, S., Whitman, L. and Cheraghi, S. H. (1999) *A supply chain transformation methodology*, translated by San Antonio, Texas, USA.
- Cho, H. and Lee, I. (1999) 'Integrated framework of IDEF modelling methods for structured design of shop floor control systems', *Int. J. Computer Integrated Manufacturing*, 12(2), 113-128.

- Chrobok, R., Kaumann, O., Wahle, J. and Schreckenber, M. (2004) 'Different methods of traffic forecast based on real data', *European Journal of Operational Research*, 155, 558–568.
- Chu, C., Proth, J. M. and Wolff, P. (1998) 'Predictive maintenance: The one-unit replacement model.', *Int. J. Production Economics*, 285-295.
- Çiflikli, C. and Kahya-Özyirmidokuz, E. (2010) 'Implementing a data mining solution for enhancing carpet manufacturing productivity', *Knowledge-Based Systems*, 23, 783–788.
- Cios, K. j., Pedrytcz, W. and Swiniarski, R. W. (1998) *Data Mining Methods for Knowledge Discovery*, USA.
- Cléroux, R., Dubuc, S. and Tilquin, C. (1979) 'The age replacement problem with minimal repair and random repair costs', *Operations Research*, 27, 1158-1167.
- Cléroux, R. and Hanscom, M. (1974) 'Age replacement with ajustement and depreciation costs and interest Charges', *Technometrics*, 16, 235-239.
- A SAS Institute White Paper (1999) *Finding the solution to Data Mining - A map of features and components of SAS Enterprise Miner Software version 3* available: www.sas.com [accessed].
- Cooke, F. L. (2003) 'Plant maintenance strategy: evidence from four British manufacturing firms', *Journal of Quality in Maintenance Engineering*, 9(3), 239–249.
- Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. (2009) *Introduction to algorithms*, ISBN:9780262033848.
- Cortez, P. (2002) *Modelos inspirados na natureza para a previsão de séries temporais*, unpublished thesis Universidade do Minho.
- Costa, J. R. R. d. (2012) *Sistema de gestão da manutenção dos equipamentos das linhas de montagem da Bosch Car Multimédia*, unpublished thesis Universiade do Minho.
- Delen, D., Walker, G. and Kadam, A. (2005) 'Predicting breast cancer survivability: a comparison of three data mining methods', *Artificial Intelligence in Medicine*, 34(2), 113-127
- Drinkwater, R. W. and Hastings, N. A. J. (1967) 'An economic replacement policy', *Operational Research Quarterly*, 18, 121-138.
- Du, J., Ling, C. X. and Zhou, Z.-H. (2011) 'When Does Cotraining Work in Real Data?', *IEEE Transactions on Knowledge and Data Engineering*, 23(5), 788-799.
- Edwards, D. J., Holt, G. D. and F.C.Harris (1998) 'Predictive maintenance techniques and their relevance to construction plant', *Journal of Quality in Maintenance Engineering*, 4(1), 25-37.
- Efthymiou, K., Papakostas, N., Mourtzis, D. and Chryssolouris, G. (2012) *On a Predictive Maintenance Platform for Production System*, translated by Athens.
- Fayyad, U. and Piatetsky-Shapiro, G. (1996b) *Knowledge Discovery and Data Mining: Towards a Unifying Framework*, translated by Oregon.
- Fayyad, U., Piatetsky-Shapiro, G., Smith, P. and Uthurusamy, R. (1996a) 'Advances in Knowledge Discovery and Data Mining', *AAAI Press/ The MIT Press*.
- Feldmann, C. and Tieso, J. V. (1998) *The Practical Guide to Business Process Reengineering Using Idef0*, New York: Dorset House, ISBN:978-0932633378.

- Ferber, J. and Gasser, L. (1991) *Intelligence artificielle distribuée*, Avignon, France.
- Ferguson, P. and Huston, G. (1998) *What is a VPN?*
- FIPS-PUBS (1993) *Integration Definition Function Modeling (IDEFO)*,
- Fore, S. and Msipha, A. (2010) 'Preventive maintenance using reliability centred maintenance (rcm): a case study of a ferrochrome manufacturing company', *South African Journal of Industrial Engineering*, 21(1), 207-233.
- Fornaro, O., Magro, M. C. and Pinceti, P. (2004) *Diagnostics for measure transmitters (predictive maintenance)*, translated by Internacional, I., Como, Italy.
- Fox, B. (1966) 'Age replacement with discounting', *Operations Research*, 10, 210-220.
- François, D. (2008) *Methodology and standards for data analysis with machine learning tools*, translated by Bruges (Belgium): ESANN'2008 Proceedings, 239-246.
- Fu, C., Ye, L., Liu, Y., Yu, R., lung, B., Cheng, Y. and Zeng, Y. (2004) 'Predictive Maintenance in Intelligent-Control-Maintenance-Management System for Hydroelectric Generating Unit', *IEEE Transactions on Energy Conversion*, 19 n°1, 179-186.
- Gago, P. M. C. (2001) *Métodos para a Seleccção das Regras Mais Promissoras Em Extracção de Conhecimento de Bases de Dados*, unpublished thesis Universidade de Coimbra.
- Gershenson, C. (2003) *Artificial Neural Networks for Beginners*.
- Gharbi, A. and Kenné, J. P. (2005) 'Maintenance scheduling and production control of multi machine manufacturing systems.', *Computers and Industrial engineering*, 693-707.
- Gilbert, D., Aparicio, M., Atkinson, B., Brady, S., Ciccarino, J., Grosf, B., O'Connor, P., Osisek, D., Pritko, S., Spagna, R. and Wilson, L. (1995) *IBM Intelligent Agent Strategy*, IBM Corporation.
- Glasser, G. J. (1967) 'The age replacement problem', *Technometrics*, 9, 83-91.
- Goebel, M. and Gruenwald, L. (1999) 'A survey of Data Mining and Knowledge Discovery Software Tools', *SIGKDD*, June, 20-33.
- Gonçalves, B. (2009) *Sistema de Planeamento e Controlo de Produção Dinâmico Baseado em Agentes*, unpublished thesis (Master), Universidade do Minho.
- Grabenstetter, D. H. and Usher, J. M. (2013) 'Determining job complexity in an engineer to order environment for due date estimation using a proposed framework', *International Journal of Production Research*, 1-13.
- Groth, R. (2000) *Data Mining: Building Competitive Advantage*, USA: Prentice Hall PTR.
- Guazzelli, A. (2010) *Representing predictive solutions in PMML - Move from raw data to predictions*.
- Guazzelli, A., Zeller, M., Lin, W.-C. and Williams, G. (2009) 'PMML: An Open Standard for Sharing Models', *The R Journal*, 1(1), 60-65.
- Haider, A. and Koronios, A. (2006) 'E-Prognostics: A step towards E-Maintenance of Engineering Assets ', *Journal of Theoretical and Applied Electronic Commerce Research*, 1(1), 42-55.

- Han, J. and Kamber, M. (2001) *Data Mining: Concepts and Techniques*, USA: Morgan Kaufmann Publishers.
- Han, J. and Kamber, M. (2011) *Data Mining: Concepts and Techniques, Third Edition*, Elsevier, ISBN: 978-0123814791.
- Harding, J., Shahbas, M., Srinivas, S. and Kusiak, A. (2006) 'Data Mining in manufacturing: A review', *Journal of Manufacturing Science and Engineering*, 128, 969-976.
- Hartley, J. (2004) 'Case study research' in Cassell, C. and Symon, G., eds., *Essential Guide to Qualitative Methods in Organizational Research*, London: Sage Publications Ltd, 323–333.
- Hashemian, H. M. and Bean, W. C. (2011) 'State-of-the-Art Predictive Maintenance Techniques', *IEEE Transactions on Instrumentation and Measurement*, 60(10), 3840 - 3492.
- Hastings, N. A. (1969) 'The repair limit replacement method', *Operational Research Quarterly*, 20, 337-349.
- Heckerman, D., Geiger, D. and Chickering, D. M. (1995) 'Learning Bayesian networks: The combination of Knowledge and statistical', *Machine Learning*, 20, 197-243.
- Heckerman, D. and Wellman, M. P. (1995) 'Bayesian Networks', *Communications of the ACM*, 38(3).
- Heitlinger, P. (2001) *O Guia Prático da XML, conceitos, exemplos, prática e aplicações da linguagem universal*, V. N. Famalicão: ISBN:972-8426-45-3.
- Hsu, L. and Kuo, S. (1995) 'Design of optimal maintenance policies based on on-line sampling plans', *European Journal of Operational Research*, 86, 345-357.
- Huo, Z. and Zhang, Z. (2003) *CMMS and Its Application in Power Systems*, translated by Internacional, I., Washington, DC.
- IEC-60300-3-11 (2009) 'Dependability management – Part 3-11: Application guide – Reliability centred maintenance',
- ISO-55001 (2014) 'Asset management - Management systems - Requirements',
- Jardine, A. K. S., Lin, D. and Banjevic, D. (2006) 'A review on machinery diagnostics and prognostics implementing condition-based maintenance', *Mechanical Systems and Signal Processing*, 20, 1483–1510.
- Jeon, H., Petrie, C. and Cutkosky, M. R. (2000) 'JATLite: A Java Agent Infrastructure with Message Routing', *IEEE Internet Computing*, 4(2), 87-96.
- Jeon, J., Kim, C. and Lee, H. (2011) 'Measuring efficiency of total productive maintenance (TPM): a three-stage data envelopment analysis (DEA) approach', *Total Quality Management*, 22(8), 911–924.
- John, G. and Langley, P. (1995) *Estimating continuous distributions in Bayesian classifiers*, translated by Citeseer: 338–345.
- Kadav, A., Kawale, J. and Mitra, P. (2003) *Data Mining Standards*.
- Kans, M. (2008) 'An approach for determining the requirements of computerised maintenance management systems', *Computers in Industry*, 59, 32-40.

- Khan, F. I. and Haddara, M. (2004) 'Risk-based maintenance (RBM): a new approach for process plant inspection and maintenance', *Process Safety Progress*, 24(4), 252-265.
- Kohavi, R. and Provost, F. (1998) 'Glossary of Terms Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, Machine Learning, 30', 271-274.
- Kufrin, R. (1997) *Generating C4.5 Production Rules in Parallel*, translated by Providence, Rhode Island.
- Lampinen, J. and Vehtari, A. (2001) 'Bayesian Approach for Neural Networks – Review and Case Studies', *Neural Networks and Soft Computing*, 14(3), 7-24.
- Lee, H. D., Monard, M. C. and A. Baranauskas, J. (1999) *Empirical Comparison of Wrapper and Filter Approaches for Feature Subset Selection*, Technical Report, ICMC-USP.
- Lee, J. (2010) 'Design of Self-Maintenance and Engineering Immune Systems for Smarter Machines and Manufacturing Systems', in Lisboa, 1 e 2 de Julho, 1st IFAC Workshop on Advanced Maintenance Engineering, Services and Technology,
- Lemm, J. C. (1999) *Bayesian field theory*, Universitat Munster, Institut fur Theoretische Physik.
- Létourneau, S., Famili, F. and Matwin, S. (1999) 'Data Mining to Predict Aircraft Component Replacement', *IEEE Intell. Syst.*, November/December 1999, 59-66.
- Lewis, E. E. (1996) *Introduction to Reliability Engineering*, Second Edition ed., John Wiley and Sons Inc.
- Lin, C.-C. and Tseng, H.-Y. (2005) 'A neural network application for reliability modelling and condition-based predictive maintenance', 174-179.
- Liu, W.-Y. and Fang, W. (2006) *Using IDEF0/Petri net for Ontology-Based Task Knowledge Analysis: The Case of Emergency Response for Debris-Flow*, translated by Hawaii: IEEE, 1-10.
- Lopes, I. (2007) *Técnicas Quantitativas no Apoio à Decisão em Sistemas de Manutenção*, unpublished thesis Universidade do Minho - Escola de Engenharia.
- Lu, S., Tu, Y.-C. and Lu, H. (2007) 'Predictive Condition-based Maintenance for Continuously Deteriorating Systems', *Quality and Reliability Engineering International*, 23, 71-81.
- Luce, S. (1999) 'Choice Criteria in Conditional Preventive Maintenance', *Mechanical Systems and Signal Processing*, 13(1), 163-168.
- Luck, M., Griffiths, N. and d'Inverno, M. (1997) *From Agent Theory to Agent Construction: A Case Study*, translated by Springer-Verlag, 49-63.
- Luck, M., McBurney, P. and Preist, C. (2003) *Agent Technology: Enabling Next Generation Computing - A Roadmap for Agent Based Computing*, ISBN:0854 327886.
- Machado, M., Rubião, J. and Souza, R. (2005) 'Previsão de séries temporais de falhas em manutenção industrial usando redes neurais', *ENGEVISTA*, 7(2), 4-18.
- Makabe, H. and Morimura, H. (1963) 'On some preventive maintenance policies', *Journal of Operations Research Society of Japan*, 6, 17-47.
- Marca, D. and McGowan, C. (1987) *Structured Analysis and Design Technique*, McGraw-Hill, ISBN:0-07-040235-3.

- Marques, R., Lourenço, L. and Barros, J. (2009) *Data Mining Techniques and Ultrasonic Pulse Velocity Tests for the Assessment of Damage Levels in Concrete exposed to High Temperatures and subject to Compression*, translated by Topping and Tsompanakis, Y., Funchal, Portugal: Civil-Comp Press, 1-11.
- Mauil, R., Smart, A. and Liang, L. (2013) 'A process model of product service supply chains', *Production Planning & Control: The Management of Operations Research*, 1-16.
- Meech, J. F. (1999) *Contextualizing User Interfaces for Complex Systems*, unpublished thesis University of Birmingham.
- Michalsky, R. S., Bratko, I. and Miroslav, K. (1998) *Machine Learning and Data Mining Methods and Applications*, England: John Wiley & Sons, Inc.
- Milne, R., Drummond, M. and Renoux, P. (1998) 'Predicting paper making defects on-line using data mining', *Knowledge-Based Systems*, 11, 332-338.
- Mitchel, T. M. (1998) *Machine Learning*, Boston, USA: McGraw-Hill.
- Morais, M. (2007) *Fiabilidade e Controlo de Qualidade*, Lisboa: Instituto Superior Técnico.
- Morimura, H. (1970) 'On some preventive maintenance policies for IFR', *Journal of Operations Research Society of Japan*, 12, 94-125.
- Moubray, J. (1997) *Reliability-Centered Maintenance II, 2^a edition*, Oxford: Butterworth-Heinemann.
- Muller, A., Marquez, A. C. and lung, B. (2008) 'On the concept of e-maintenance: Review and current research', *Reliability Engineering and System Safety*, 93, 1165-1187.
- Muth, E. J. (1977) 'An optimal decision rule for repair vs replacement', *IEEE Transactions*, R-26, 179-181.
- Nakagawa, T. (1980) 'Replacement models with inspection and preventive maintenance', *Microelectronics and Reliability*, 20, 427-433.
- Nakagawa, T. and Mizutani, S. (2009) 'A summary of maintenance policies for a finite interval', *Reliability Engineering and System Safety*, 94, 89-96.
- Netgear, ed. (2005) *Virtual Private Networking Basics*, Santa Clara, USA: Netgear, Inc.
- NP-EN-13306 (2007) 'Norma Portuguesa - Terminologia da Manutenção',
- Nwana, H. (1996) 'Software agents: An overview', *The Knowledge Engineering Review*, 11(3), 1-40.
- O'Connor, P. and Kleyner, A. (2012) *Practical Reliability Engineering, 5th Edition*, ISBN:978-0-470-97982-2.
- Oliveira, J. and Rosa, L. d. (2010) 'Modeling of processes IDEF: descriptive model of the productive chain of biodiesel', *Revista Gestão Industrial*, 6(2), 159-174.
- McGraw Hill (1999) *Maintenance planning and scheduling handbook*. available: [accessed].
- Park, K. S. (1979) 'Optimal number of minimal repairs before replacement', *IEEE Transactions*, R-28, 137-140.

- Pereira, F. J. D. and Sena, F. M. V. (2012) *Fiabilidade e sua Aplicação à Manutenção*, Publindústria, ISBN:9789728953997.
- Piairo, J., Madureira, A., Pereira, J. P. and Pereira, I. (2013) 'Desenvolvimento e avaliação de um interface com o utilizador para um sistema de escalonamento', *Revista Ibérica de Sistemas e tecnologia de Informação*, 11, 77-91.
- Pinjala, S. K., Pintelona, L. and Vereecke, A. (2006) 'An empirical investigation on the relationship between business and maintenance strategies', *Int. J. Production Economics*, 214-229.
- Pinto, F. (2005) *A Descoberta de Conhecimento em Bases de Dados como Suporte à Actividades de Business Intelligence - Aplicação na Área de Database Marketing*, unpublished thesis Escola de Engenharia - Universidade do Minho.
- Pires, L. (2004) *Desenvolvimento de um sistema de planeamento e controlo da produção para empresas distribuídas virtuais*, unpublished thesis Universidade do Minho.
- Quinlan, J. R. (1986) 'Induction of Decision Trees', *Machine Learning*, 1, 81-106.
- Quinlan, J. R. (1998) *C4.5 Programs for Machine Learning*, Morgan Kaufmann Publishers, Inc.
- Ran, A. and Rosenlund, S. I. (1976) 'Age replacement with discounting for a continuous maintenance cost model', *Technometrics*, 18, 459-465.
- Rausand, M. (1998) 'Reliability Centred Maintenance', *Journal of Reliability Engineering and System Safety*, 60(2), 121-132.
- Reffat, R., Gero, J. and Peng, W. (2004) *Using data mining on building maintenance during the building life cycle*, translated by Tasmania, Australia: 91-97.
- Rezende, O. S., ed. (2003) *Sistemas Inteligentes Fundamentos e Aplicações*, Editora Manole, Lda ed., Brasil.
- Rich, E. and Knight, K. (1991) *Artificial Intelligence*, Singapore: International Edition, MacGraw Hill.
- Rigdon, S. E. (2008) 'Repairable Systems Reliability' in *Encyclopedia of Quantitative Risk Analysis and Assessment*, John Wiley & Sons ed.
- Romanowski, C. and Nagi, R. (2001) *Analyzing maintenance data using data mining methods, Data mining for design and manufacturing*, MA, USA: Dan Braha Kluwer Academic Publishers Norwell, ISBN:1-4020-0034-0
- Rud, O. P. (2001) *Data Mining Cookbook*, USA: John Wiley & Sons, Inc.
- Ruiz, P. P., Foguem, B. K. and Grabot, B. (2014) 'Generating Knowledge in Maintenance from Experience Feedback', *Knowledge-Based Systems*.
- Sahoo, T. and Liyanage, J. P. (2008) 'Computerized Maintenance Management Systems For Effective Plant Performance.', *Chemical Engineering*, 115(1), 38-41.
- Santos, M. F. and Azevedo, C. (2005) *Data Mining - Descoberta de conhecimento de bases de dados*, FCA - Editora de Informática, ISBN:972-722-509-8.
- Schaeffer, R. L. (1971) 'Optimum age replacement policies with increasing cost factor', *Technometrics*, 3, 139-144.
- Schippers, W. A. J. (2001) 'An integrated approach to process control', *International Journal of Production Economics*, 69(1), 93-105.

- Schmidt, R. (2003) *Busca e seleção de parceiros para empresas virtuais: Uma abordagem baseada em agentes móveis*, unpublished thesis (Master), Universidade Federal de Santa Catarina.
- Scott, C., Wolfe, P. and Erwin, M. (1999) *Virtual Private Networks*, O'Reilly, ISBN:1-56592-529-7.
- Serifi, V., Daačić, P., Jecmenica, R. and Labović, D. (2009) 'Functional and Information Modeling of Production Using IDEF Methods', *Journal of Mechanical Engineering*, 55(2), 131-140.
- Serrano, L., Alcobia, C., Mateus, M. and Silva, M. (2004) *Sistemas de Aquisição, Processamento e Armazenamento de Dados*, translated by IPQ- Caparica.
- Serrão, C. and Marques, J. (2000) *Programação com PHP*, FCA, ISBN:972-722-201-3.
- Shang, Z., Qiu, C. and Zhu, S. (2011) 'A Model of Application System for Man-Machine-Environment System Engineering in Vessels Based on IDEF0 ', *J. Marine Sci. Appl.*, 10, 347-357.
- Shehab, E., Fowler, C., Gil, A. R., Abdalla, H., Darwish, M., Abdulhafed, H., Ahmed, A., Ahouie, H., Alechnovic, A., Paumes, C., Tacchini, E., Urbane, I. and Richardson, M. (2013) 'Enhancement of product information collaboration and access in the aerospace industry', *International Journal of Production Research*, 51(11), 3225–3240.
- Siqueira, I. Q. (2005) *Manutenção Centrada na Confiabilidade*, Qualitymark.
- Sivazlian, B. D. (1973) 'On a discounted replacement problem with arbitrary repair time distribution', *Management Science*, 19, 1301-1309.
- Smith, A. M. (1993) *Reliability Centred Maintenance*, New York: McGraw Hill Inc.
- Su, J., Jelber, S. S., Matwin, S. and Huang, J. (2009) 'Active Learning with Automatic Soft Labeling for Induction of Decision Trees ', *Lecture Notes in Computer Science*, 5549/2009, 241-244.
- Swanson, L. (2001) 'Linking maintenance strategies to performance', *Int. J. Production Economics*, 237-244.
- Sylvain, L., Fazel, F. and Stan, M. (1999) 'Data Mining to Predict Aircraft Component Replacement', *IEEE Intelligent Systems*, 14(6), 59-65.
- Tan, P.-N., Steinbach, M. and Kumar, V. (2006) *Introduction to Data Mining*, Pearson Addison Wesley, ISBN:978-0321321367.
- Teixeira, B. (2008) *Análise de confiabilidade aplicada ao projecto de sistemas mecânicos*, unpublished thesis Universidade de São Paulo.
- Teixeira, H. A. d. C. Q. (2005) *Sistemas de conhecimento baseados em data mining : aplicação à análise da estabilidade de estruturas metálicas*, unpublished thesis Universidade do Minho - Escola de Engenharia.
- Toloo, M., Sohrabi, B. and Nalchigar, S. (2009) 'A new method for ranking discovered rules from data mining by DEA', *Expert Systems with Applications*, 36, 8503 - 8508.
- Torell, W. and Avelar, V. (2004) *Tempo Médio entre Falhas: Explicação e Padrões*.
- Tsang, A. H. C. (1998) 'A strategic approach to managing maintenance performance', *Journal of Quality in Maintenance Engineering*, 4(2), 87-94.

- Vrignat, P., Avila, M., Duculty, F., Aupetit, S., Slimane, M. and Kratz, F. (2012) 'Maintenance policy: degradation laws versus Hidden Markov Model availability indicator', *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 226(2), 137-155.
- Waltman, W. D. and Presley, A. (1993) 'Reading and Critiquing and IDEF0 Model', 2012(2012)available: <http://www.webs.twsu.edu/enteng/papers/howidef0.pdf> [accessed 2012].
- Weinstein, L. and Chung, C.-H. (1999) 'Integrating maintenance and production decisions in a hierarchical production planning environment', *Computers & Operations Research*, 26, 1059 - 1074.
- Wilder, P. and Cannon, H. (1993) *Advantages of a computerized maintenance management system in managing plant operations* translated by Chicago, USA: IEEE, 501-512.
- Wireman, T. (2004) *Total Productive Maintenance*, Industrial Press ed., ISBN:978-0-8311-0210-4.
- Witten, I. H. and Frank, E. (2005) *Data Mining. Practical Machine Learning Tools and Techniques, Second Edition.*, Morgan Kaufmann Publishers, ISBN:0-12-088407-0.
- Yao, X., Fernández-Gaucherand, E., Fu, M. C. and Marcus, S. I. (2004) 'Optimal Preventive Maintenance Scheduling in Semiconductor Manufacturing', *IEEE Transactions On Semiconductor Manufacturing*, 17(3), 345-356.
- Yee, P., Fischer, M. and Kam, C. (2013) 'Prospective Validation of Virtual Design and Construction Methods', *Journal of Information Technology in Construction*, 18, 214-239.
- Zerbinatti, L. and Júnior, L. M. (2006) 'Fuzzy e Bayes no Apoio à Decisão Médica', *Exacta*, 4(2), 309-315.
- Zhang, P. and Su, W. (2012) 'Statistical Inference on Recall, Precision and Average Precision under Random Selection', in *9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012)* Chongqing, Sichuan, China, IEEE, 1348-1352.
- Zhang, Z. and Kusiak, A. (2012) 'Monitoring Wind Turbine Vibration Based on SCADA Data', *Journal of Solar Energy Engineering*, 134, 1-12.
- Zio, E. and Compare, M. (2013) 'Evaluating maintenance policies by quantitative modeling and analysis', *Reliability Engineering and System Safety*, 109, 53-65.

Anexos

Anexo A

Anexo representativo do *Interface* de autenticação *Back-end* do portal. O processo de autenticação é efetuado através da introdução do nome de utilizador referente ao administrador do portal *web* e a respetiva palavra-passe.



Iniciar sessão para administração de Joomla!

Utilize um nome de utilizador e senha válidos para aceder à zona de administração do site.

[Voltar à página principal do site](#)

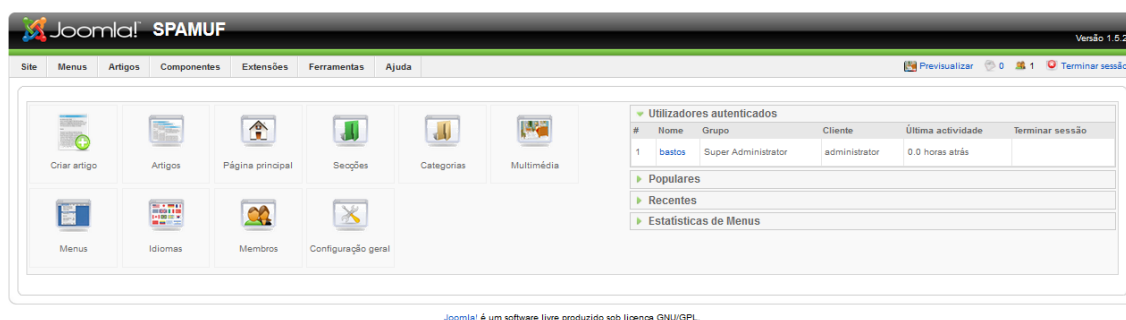
Nome de utilizador

Senha

Idioma

Autenticar 

Após o processo de autenticação, o administrador acede ao interface *Back-end* de gestão do portal *web* e apresentado na figura seguinte.



Joomla! SPAMUF Versão 1.5.22

Site Menus Artigos Componentes Extensões Ferramentas Ajuda

Previsualizar 0 1 Terminar sessão

Utilizadores autenticados

#	Nome	Grupo	Cliente	Última actividade	Terminar sessão
1	bastos	Super Administrator	administrador	0.0 horas atrás	

Populares
Recentes
Estatísticas de Menus

Joomla! é um software livre produzido sob licença GNU/GPL

Todo o processo gestão de artigos (páginas *web*), instalação de *templates*, módulos e *plugins* é efetuado através dos menus existentes fornecendo acessos a diferentes áreas de gestão associadas ao objeto em causa. Toda a parametrização dos diferentes objetos instalados é efetuada em áreas específicas.

Além do interface gráfico de programação a área de *Back-end* fornece acesso ao código fonte de programação, podendo assim ser manipulado a mais baixo nível.

Seguidamente é listado o código *PHP* do *template* (modelo do site *web*) elaborado e aplicado ao interface *Front-end* do portal.

```
<?php
defined('_JEXEC') or die('Restricted access'); // no direct access
require_once dirname(__FILE__) . DIRECTORY_SEPARATOR . 'functions.php';
$document = null;
if (isset($this))
    $document = & $this;
$baseUrl = $this->baseUrl;
$templateUrl = $this->baseUrl . '/templates/' . $this->template;
artxComponentWrapper($document);
?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="<?php echo $this->language; ?>"
lang="<?php echo $this->language; ?>" >
<head>
<jdoc:include type="head" />
<link rel="stylesheet" href="<?php echo $this->baseUrl; ?>/templates/system/css/system.css"
type="text/css" />
<link rel="stylesheet" href="<?php echo $this->baseUrl; ?>/templates/system/css/general.css"
type="text/css" />
<link rel="stylesheet" type="text/css" href="<?php echo $templateUrl; ?>/css/template.css"
media="screen" />
<!--[if IE 6]><link rel="stylesheet" href="<?php echo $templateUrl; ?>/css/template.ie6.css"
type="text/css" media="screen" /><![endif]-->
<!--[if IE 7]><link rel="stylesheet" href="<?php echo $templateUrl; ?>/css/template.ie7.css"
type="text/css" media="screen" /><![endif]-->
<script type="text/javascript" src="<?php echo $templateUrl; ?>/script.js"></script>
</head>
<body>
<div id="art-page-background-simple-gradient">
    <div id="art-page-background-gradient"></div>
</div>
<div id="art-main">
<div class="art-sheet">
    <div class="art-sheet-tl"></div>
    <div class="art-sheet-tr"></div>
    <div class="art-sheet-bl"></div>
    <div class="art-sheet-br"></div>
    <div class="art-sheet-tc"></div>
```

```

<div class="art-sheet-bc"></div>
<div class="art-sheet-cl"></div>
<div class="art-sheet-cr"></div>
<div class="art-sheet-cc"></div>
<div class="art-sheet-body">
<div class="art-header">
  <div class="art-header-jpeg"></div>
<div class="art-logo">
  <h1 id="name-text" class="art-logo-name"><a href="<?php echo $baseUrl;
?>/">SPAMUF</a></h1>
  <div id="slogan-text" class="art-logo-text">Sistema de Prediço de Avarias em Unidades Fabris
Globalmente Dispersas</div>
</div>
</div>
<jdoc:include type="modules" name="banner3" />
<jdoc:include type="modules" name="banner1" style="artstyle" artstyle="art-nostyle" />
<?php echo artxPositions($document, array('top1', 'top2', 'top3'), 'art-block'); ?>
<div class="art-content-layout">
  <div class="art-content-layout-row">
    <?php if (artxCountModules($document, 'left')) : ?>
    <div class="art-layout-cell art-sidebar1">
      <jdoc:include type="modules" name="left" style="xhtml" />
    </div>
    <?php endif; ?>
    <div class="art-layout-cell art-?>?php echo artxGetContentCellStyle($document); ?>">
    <?php
      echo artxModules($document, 'banner2', 'art-nostyle');
      if (artxCountModules($document, 'breadcrumb'))
        echo artxPost(null, artxModules($document, 'breadcrumb'));
      echo artxPositions($document, array('user1', 'user2'), 'art-article');
      echo artxModules($document, 'banner3', 'art-nostyle');
    ?>
    <?php if (artxHasMessages()) : ?><div class="art-post">
      <div class="art-post-body">
        <div class="art-post-inner">
          <div class="art-postcontent">
            <!-- article-content -->
            <jdoc:include type="message" />
            <!-- /article-content -->
          </div>
          <div class="cleared"></div>
        </div>
        <div class="cleared"></div>
      </div>
    </div>
    <?php endif; ?>
    <jdoc:include type="component" />
    <?php echo artxModules($document, 'banner4', 'art-nostyle'); ?>

```

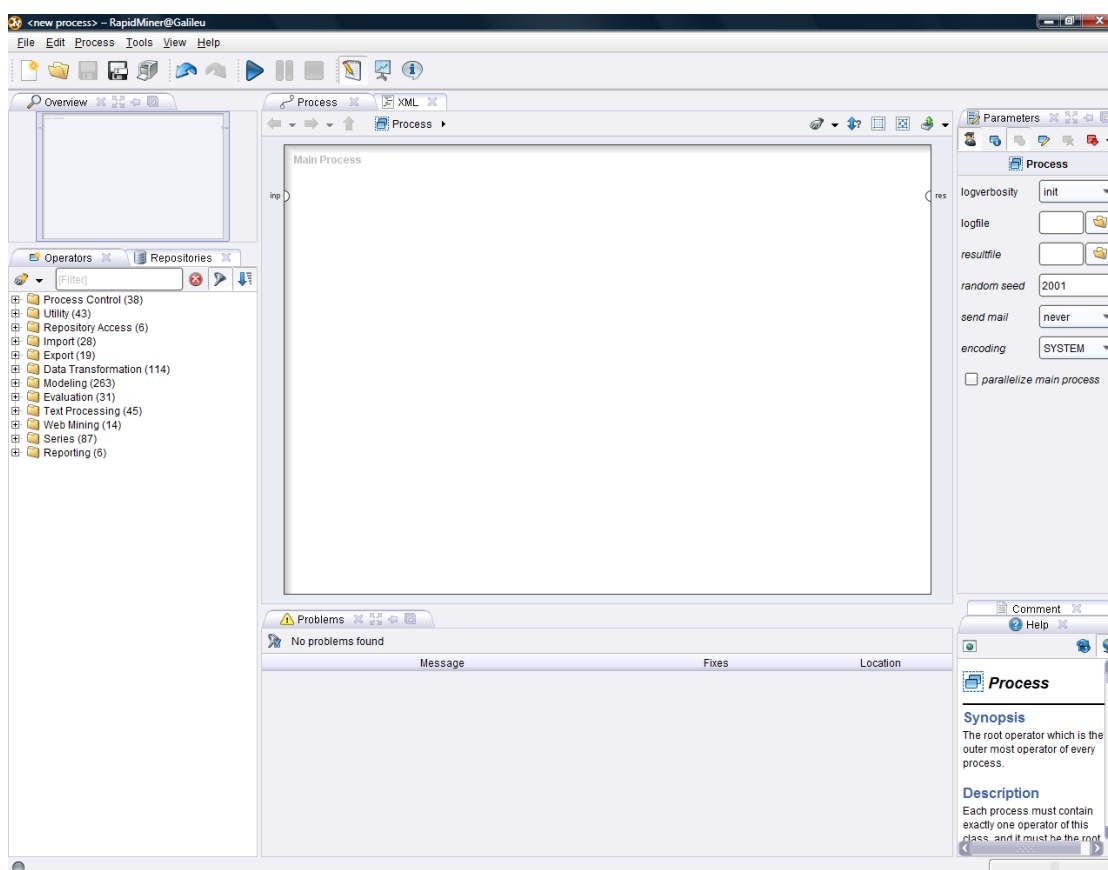
```

<?php echo artxPositions($document, array('user4', 'user5'), 'art-article'); ?>
<?php echo artxModules($document, 'banner5', 'art-nostyle'); ?>
</div>
<?php if (artxCOUNTModules($document, 'right')) : ?>
<div class="art-layout-cell art-sidebar2"><?php echo artxModules($document, 'right', 'art-block');
?>
</div>
<?php endif; ?>
</div>
<div class="cleared"></div>
<?php echo artxPositions($document, array('bottom1', 'bottom2', 'bottom3'), 'art-block'); ?>
<jdoc:include type="modules" name="banner6" style="artstyle" artstyle="art-nostyle" />
<div class="art-footer">
  <div class="art-footer-t"></div>
  <div class="art-footer-b"></div>
  <div class="art-footer-body">
    <?php echo artxModules($document, 'syndicate'); ?>
    <div class="art-footer-text">
      <?php if (artxCOUNTModules($document, 'copyright') == 0): ?>
<p>Copyright &copy; 2011 Pedro Bastos.<br />
Todos os direitos reservados.</p>
      <?php else: ?>
      <?php echo artxModules($document, 'copyright', 'art-nostyle'); ?>
      <?php endif; ?>
    </div>
    <div class="cleared"></div>
  </div>
</div>
  <div class="cleared"></div>
</div>
  <div class="cleared"></div>
</div>
</div>
<div class="cleared"></div>
</div>
</body>
</html>

```

Anexo B

Anexo representativo do *interface* proporcionado pelo *RapidMiner*. Através da figura seguinte são visualizadas e descritas as diferentes abas seleccionáveis durante a implementação de um processo de *data mining*:



- i) *Overview*: apresenta o esquema de implementação global de todas as ferramentas activas;
- ii) *Operators*: lista de operadores disponíveis agrupados por categorias e subcategorias de tarefas e funções;
- iii) *Repositories*: espaço de armazenagem das bases de dados e dos processos (projectos de *data mining*) construídos;
- iv) *Process*: interface onde é construído o processo de forma simples, arrastando e conectando operadores;
- v) *Tree*: interface que apresenta o processo no formato de árvore, permitindo também a construção do mesmo tipo de visualização;

- vi) *XML*: linha de comandos da linguagem XML, utilizada pelo programa para desenhar o processo, possibilitando uma forma alternativa de construção do mesmo.
- vii) *Problems*: aba que reúne todos os erros identificados pelo software;
- viii) *Log*: Registo de eventos de todas as operações realizadas e respectiva descrição;
- ix) *Parameters*: vista que possibilita a definição e configuração de parâmetros disponíveis em cada operador;
- x) *Help*: breve descrição de cada um dos operadores, com a descrição dos métodos algoritmos que implementa, dos parâmetros de entrada e dos outputs fornecidos;
- xi) *Comment*: permite a introdução de comentários por parte do utilizador.

Anexo C

Código XML referente ao processo 1 implementado em *RapidMiner*.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.3.015">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.3.015" expanded="true"
name="Process">
    <process expanded="true">
      <operator activated="true" class="retrieve" compatibility="5.3.015" expanded="true"
height="60" name="Retrieve Dados Modelo" width="90" x="45" y="390">
        <parameter key="repository_entry" value="Dados Modelo"/>
      </operator>
      <operator activated="true" class="replace_missing_values" compatibility="5.3.015"
expanded="true" height="94" name="Replace Missing Values" width="90" x="45" y="255">
        <parameter key="default" value="none"/>
        <list key="columns"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="5.3.015"
expanded="true" height="76" name="Select Attributes" width="90" x="45" y="165">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="Avaria Tipo | ID Componentes | ID
Locais | param_id | value"/>
      </operator>
      <operator activated="true" class="set_role" compatibility="5.3.015" expanded="true"
height="76" name="Set Role" width="90" x="45" y="30">
        <parameter key="attribute_name" value="Avaria Tipo"/>
        <parameter key="target_role" value="label"/>
        <list key="set_additional_roles"/>
      </operator>
      <operator activated="true" class="correlation_matrix" compatibility="5.3.015"
expanded="true" height="94" name="Correlation Matrix" width="90" x="179" y="75"/>
      <operator activated="true" class="split_validation" compatibility="5.3.015" expanded="true"
height="112" name="Validation" width="90" x="246" y="255">
        <process expanded="true">
          <operator activated="true" class="naive_bayes_kernel" compatibility="5.3.015"
expanded="true" height="76" name="Naive Bayes (Kernel)" width="90" x="112" y="30">
            <parameter key="estimation_mode" value="full"/>
            <parameter key="use_application_grid" value="true"/>
          </operator>
```

```

    <connect from_port="training" to_op="Naive Bayes (Kernel)" to_port="training set"/>
    <connect from_op="Naive Bayes (Kernel)" from_port="model" to_port="model"/>
    <portSpacing port="source_training" spacing="0"/>
    <portSpacing port="sink_model" spacing="0"/>
    <portSpacing port="sink_through 1" spacing="0"/>
  </process>
  <process expanded="true">
    <operator activated="true" class="apply_model" compatibility="5.3.015"
expanded="true" height="76" name="Apply Model" width="90" x="45" y="30">
      <list key="application_parameters"/>
    </operator>
    <operator activated="true" class="performance_classification" compatibility="5.3.015"
expanded="true" height="76" name="Performance" width="90" x="179" y="30">
      <parameter key="skip_undefined_labels" value="false"/>
      <list key="class_weights"/>
    </operator>
    <connect from_port="model" to_op="Apply Model" to_port="model"/>
    <connect from_port="test set" to_op="Apply Model" to_port="unlabelled data"/>
    <connect from_op="Apply Model" from_port="labelled data" to_op="Performance"
to_port="labelled data"/>
    <connect from_op="Performance" from_port="performance" to_port="averagable 1"/>
    <portSpacing port="source_model" spacing="0"/>
    <portSpacing port="source_test set" spacing="0"/>
    <portSpacing port="source_through 1" spacing="0"/>
    <portSpacing port="sink_averagable 1" spacing="0"/>
    <portSpacing port="sink_averagable 2" spacing="0"/>
  </process>
</operator>
  <operator activated="true" class="multiply" compatibility="5.3.015" expanded="true"
height="94" name="Multiply" width="90" x="447" y="30"/>
  <operator activated="true" class="reporting:generate_report" compatibility="5.3.000"
expanded="true" height="112" name="Generate Report" width="90" x="447" y="255">
    <parameter key="report_name" value="Bosch"/>
    <parameter key="format" value="HTML"/>
    <parameter key="html_output_directory" value="C:\Users\bastos\Desktop\regras"/>
    <parameter key="html_logo_file"
value="C:\Users\bastos\Desktop\bosch_logo_portugese.png"/>
  </operator>
  <operator activated="true" class="reporting:report" compatibility="5.3.000"
expanded="true" height="60" name="Report (2)" width="90" x="648" y="300">
    <parameter key="report_item_header" value="Performance do Modelo"/>
    <parameter key="specified" value="true"/>
    <parameter key="reportable_type" value="Performance Vector"/>
    <parameter key="renderer_name" value="Table / Plot View"/>
    <list key="parameters"/>
  </operator>
  <operator activated="true" class="reporting:report" compatibility="5.3.000"
expanded="true" height="60" name="Report (3)" width="90" x="648" y="390">

```

```

<parameter key="report_item_header" value="Matriz de Correlação"/>
<parameter key="specified" value="true"/>
<parameter key="reportable_type" value="Numerical Matrix"/>
<parameter key="renderer_name" value="Table View"/>
<list key="parameters">
  <parameter key="min_row" value="1"/>
  <parameter key="max_row" value="2147483647"/>
  <parameter key="min_column" value="1"/>
  <parameter key="max_column" value="2147483647"/>
</list>
</operator>
<operator activated="true" class="reporting:report" compatibility="5.3.000"
expanded="true" height="60" name="Report" width="90" x="648" y="210">
  <parameter key="report_item_header" value="Modelo Rede de Naive Bayes"/>
  <parameter key="specified" value="true"/>
  <parameter key="reportable_type" value="Kernel Bayes Model"/>
  <parameter key="renderer_name" value="Plot View"/>
  <list key="parameters">
    <parameter key="range_" value="."/>
    <parameter key="rotate_labels" value="false"/>
  </list>
  <parameter key="image_width" value="1280"/>
</operator>
<operator activated="true" class="write_model" compatibility="5.3.015" expanded="true"
height="60" name="Write Model" width="90" x="648" y="30">
  <parameter key="model_file" value="D:\Doutoramento\Tese\Dados Finais\modelo rede
bayes"/>
  <parameter key="output_type" value="Binary"/>
</operator>
<connect from_op="Retrieve Dados Modelo" from_port="output" to_op="Replace Missing
Values" to_port="example set input"/>
<connect from_op="Replace Missing Values" from_port="example set output" to_op="Select
Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Set Role"
to_port="example set input"/>
<connect from_op="Set Role" from_port="example set output" to_op="Correlation Matrix"
to_port="example set"/>
<connect from_op="Correlation Matrix" from_port="example set" to_op="Validation"
to_port="training"/>
<connect from_op="Correlation Matrix" from_port="matrix" to_op="Generate Report"
to_port="through 3"/>
<connect from_op="Validation" from_port="model" to_op="Multiply" to_port="input"/>
<connect from_op="Validation" from_port="averagable 1" to_op="Generate Report"
to_port="through 2"/>
<connect from_op="Multiply" from_port="output 1" to_op="Write Model" to_port="input"/>
<connect from_op="Multiply" from_port="output 2" to_op="Generate Report"
to_port="through 1"/>

```

```

    <connect from_op="Generate Report" from_port="through 1" to_op="Report"
to_port="reportable in"/>
    <connect from_op="Generate Report" from_port="through 2" to_op="Report (2)"
to_port="reportable in"/>
    <connect from_op="Generate Report" from_port="through 3" to_op="Report (3)"
to_port="reportable in"/>
    <connect from_op="Report (2)" from_port="reportable out" to_port="result 4"/>
    <connect from_op="Report (3)" from_port="reportable out" to_port="result 3"/>
    <connect from_op="Report" from_port="reportable out" to_port="result 2"/>
    <connect from_op="Write Model" from_port="through" to_port="result 1"/>
    <portSpacing port="source_input 1" spacing="0"/>
    <portSpacing port="sink_result 1" spacing="0"/>
    <portSpacing port="sink_result 2" spacing="0"/>
    <portSpacing port="sink_result 3" spacing="0"/>
    <portSpacing port="sink_result 4" spacing="0"/>
    <portSpacing port="sink_result 5" spacing="0"/>
  </process>
</operator>
</process>

```

Anexo D

Código XML referente ao processo 2 implementado em *RapidMiner*.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.3.015">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.3.015" expanded="true"
name="Process">
    <process expanded="true">
      <operator activated="true" class="read_model" compatibility="5.3.015" expanded="true"
height="60" name="Read Model" width="90" x="45" y="30">
        <parameter key="model_file" value="D:\Doutoramento\Tese\Dados Finais\modelo
rede bayes"/>
      </operator>
      <operator activated="true" class="retrieve" compatibility="5.3.015" expanded="true"
height="60" name="Retrieve Dados para Predição" width="90" x="45" y="390">
        <parameter key="repository_entry" value="Dados para Predição"/>
      </operator>
      <operator activated="true" class="replace_missing_values" compatibility="5.3.015"
expanded="true" height="94" name="Replace Missing Values" width="90" x="45" y="255">
        <parameter key="default" value="none"/>
        <list key="columns"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="5.3.015"
expanded="true" height="76" name="Select Attributes" width="90" x="45" y="120">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="param_id|value"/>
      </operator>
      <operator activated="true" class="apply_model" compatibility="5.3.015"
expanded="true" height="76" name="Apply Model" width="90" x="179" y="30">
        <list key="application_parameters"/>
      </operator>
      <operator activated="true" class="multiply" compatibility="5.3.015" expanded="true"
height="94" name="Multiply" width="90" x="313" y="30"/>
      <operator activated="true" class="write_csv" compatibility="5.3.015" expanded="true"
height="76" name="Write CSV" width="90" x="514" y="30">
        <parameter key="csv_file" value="D:\Doutoramento\Tese\Dados
Finais\resultados.csv"/>
      </operator>
    </process>
  </operator>
</process>
```

```

    <operator activated="true" class="reporting:generate_report" compatibility="5.3.000"
expanded="true" height="76" name="Generate Report" width="90" x="313" y="210">
    <parameter key="report_name" value="Bosch"/>
    <parameter key="format" value="HTML"/>
    <parameter
                                                                    key="html_output_directory"
value="C:\Users\bastos\Desktop\Predicao"/>
    <parameter
                                                                    key="html_logo_file"
value="C:\Users\bastos\Desktop\bosch_logo_portugese.png"/>
    <parameter key="html_image_format" value="gif"/>
</operator>
    <operator activated="true" class="reporting:report" compatibility="5.3.000"
expanded="true" height="60" name="Report" width="90" x="447" y="210">
    <parameter key="report_item_header" value="Tabela de Predição"/>
    <parameter key="specified" value="true"/>
    <parameter key="reportable_type" value="Data Table"/>
    <parameter key="renderer_name" value="Data View"/>
    <list key="parameters">
    <parameter key="attribute_filter_type" value="all"/>
    <parameter key="use_except_expression" value="false"/>
    <parameter key="value_type" value="attribute_value"/>
    <parameter key="use_value_type_exception" value="false"/>
    <parameter key="except_value_type" value="time"/>
    <parameter key="block_type" value="attribute_block"/>
    <parameter key="use_block_type_exception" value="false"/>
    <parameter key="except_block_type" value="value_matrix_row_start"/>
    <parameter key="invert_selection" value="false"/>
    <parameter key="include_special_attributes" value="true"/>
    <parameter key="min_row" value="1"/>
    <parameter key="max_row" value="474465"/>
    </list>
</operator>
    <connect from_op="Read Model" from_port="output" to_op="Apply Model"
to_port="model"/>
    <connect from_op="Retrieve Dados para Predição" from_port="output" to_op="Replace
Missing Values" to_port="example set input"/>
    <connect from_op="Replace Missing Values" from_port="example set output"
to_op="Select Attributes" to_port="example set input"/>
    <connect from_op="Select Attributes" from_port="example set output" to_op="Apply
Model" to_port="unlabelled data"/>
    <connect from_op="Apply Model" from_port="labelled data" to_op="Multiply"
to_port="input"/>
    <connect from_op="Multiply" from_port="output 1" to_op="Write CSV" to_port="input"/>
    <connect from_op="Multiply" from_port="output 2" to_op="Generate Report"
to_port="through 1"/>
    <connect from_op="Write CSV" from_port="through" to_port="result 2"/>
    <connect from_op="Generate Report" from_port="through 1" to_op="Report"
to_port="reportable in"/>
    <connect from_op="Report" from_port="reportable out" to_port="result 1"/>

```

```
<portSpacing port="source_input 1" spacing="0"/>  
<portSpacing port="sink_result 1" spacing="0"/>  
<portSpacing port="sink_result 2" spacing="0"/>  
<portSpacing port="sink_result 3" spacing="0"/>  
</process>  
</operator>  
</process>
```