Universidade do Minho
Escola de Engenharia

Olívia Raquel Ferreira Oliveira

Extração de Conhecimento nas Listas de Espera para Consulta e Cirurgia

Outubro de 2012

Olívia Raquel Ferreira Oliveira **Extração de Conhecimento nas Listas de Espera para Consulta e Cirurgia**

UMinho | 2012

Universidade do Minho
Escola de Engenharia

Olívia Raquel Ferreira Oliveira

Extração de Conhecimento nas Listas de Espera para Consulta e Cirurgia

Dissertação de Mestrado
Mestrado Integrado em Engenharia Biomédica
Área de especialização Informática Médica

Trabalho efetuado sob a orientação do
Professor Doutor José Machado

Outubro de 2012

Os verdadeiros campeões são aqueles que caem e se levantam,
e que voltam a cair e se levantam novamente...
E não aqueles que nunca caíram!

Agradecimentos

Agradeço ao Professor José Machado por ter sido um orientador presente, demonstrando inteira disponibilidade para resolver todos os problemas que foram surgindo ao longo do projeto.

Agradeço à minha Mãe, Irmão e Prima Ana, primos e tios, e a toda a família por todo o apoio, carinho e compaixão prestados ao longo do projeto.

Agradeço aos meus amigos, em especial a Mariana, a Nídia e o Elias, pelos momentos de incentivo e motivação quando tudo parecia mais difícil e improvável de concretizar. Um especial agradecimento ao Tiago pelo livro que me emprestou e que me ajudou no esclarecimento de ideias fundamentais.

Agradeço aos meus colegas de curso, em especial à companheira de trabalho Marta, pelos momentos partilhados e de entre-ajuda ao longo de todo o percurso.

Resumo

O desenvolvimento industrial, a constante inovação e a necessidade de melhoria contínua por parte das organizações originou um crescimento exponencial do volume de dados armazenados existindo, por isso, uma maior quantidade de informação relacionada com cada instituição. Cada vez mais as instituições dependem do tipo de dados que armazenam e acumulam.

Atualmente, uma organização tem de fazer uma gestão eficiente das suas bases de dados de modo a extrair o máximo de conhecimento possível para apoiar no processo de tomada de decisão e assim garantir competitividade no mundo dos negócios e dos mercados. A necessidade de implementação de um sistema de apoio à decisão no seio de uma instituição fez emergir o conceito de *Business Intelligence*: processo responsável pela transformação dos dados em informação útil e organizada, e a subsequente conversão dessa informação em conhecimento valioso para a tomada de decisão.

A área da saúde é bastante susceptível, tanto a nível clínico como administrativo, à qualidade e rapidez das decisões tomadas, uma vez que estas decisões colocam sempre em causa a vida humana. Assim, é de extrema importância a utilização de sistemas de apoio à decisão nas unidades de saúde.

O propósito deste projeto prende-se, essencialmente, com a exploração e aplicação de uma ferramenta de BI aplicada no contexto da saúde. Por outro lado, pretende-se também avaliar a aplicabilidade de uma ferramenta *open source* em sistemas complexos e integrados como os das instituições hospitalares. Neste sentido, a ferramenta explorada e avaliada foi o Pentaho. Foi realizada uma monitorização e simulação dos dados clínicos relativos às listas de espera para consulta e cirurgia e à ocupação das salas do bloco operatório de um hospital no norte de Portugal.

Verificou-se que o Pentaho, enquanto ferramenta *open source*, é inteiramente capaz de ser implementada e integrada numa instituição hospitalar, com a potencialidade de uma ferramenta proprietária. Sendo assim conclui-se que o Pentaho é uma ferramenta de BI bastante eficiente, capaz de apresentar soluções válidas e atrativas para a resolução de problemas e o para suporte à tomada de decisões.

Abstract

Nowadays organizational environment is entirely dependent on several types of data such as, operational data, suppliers and customers. With the industrial development, the constant innovation and the need for continuous improvement by the organizations, the volume of stored data has grown exponentially, and therefore, there is a greater amount of information regarding each institution.

Presently, an organization has to manage efficiently its databases in order to turn all the available information into possible valuable knowledge to support the decision making process and to be competitive in the business and markets worlds. The need of a system to support decision making within an organization, created the concept of Business Intelligence (BI): process responsible for turning data into usefull and organized information, and convert it to valuable knowledge for decison making.

Healthcare is a field which is very susceptible to the speed and quality of decision making, both at the clinical and administrative levels, since this decisions often put human life at risk. Therefore, it is of extreme importance the implementation of decision support systems in healthcare facilities.

The main purpose of this project is implementing a BI tool in the healthcare context. Moreover, it also aims to evaluate the performance of an open source tool when applied to complex and integrated systems, such as hospital units. For that, the tool explored and evaluated was Pentaho. The clinical data analysed were the waiting lists for doctor's appointments and surgery, and occupational rate of the operating room of an hospital in the north of Portugal.

Pentaho, as an open source tool, is fully capable of being implemented and integrated in an hospital organization, with the potentiality of a proprietary tool. It was concluded that Pentaho is a very efficient BI tool, capable of presenting valid and attractive solutions for problem solving and support of decision making.

Conteúdo

Lista de Figuras

Lista de Tabelas

Glossário

BD - Base de Dados
EC - Extração de Conhecimento
BI - Business Intelligence
DW - Data Warehouse
ETL - Extract, Transform and Load
DM - Data Mining
TI - Tecnologias de Informação
OLAP - Online Analytical Processing
OLTP - Online Transactional Processing
KPI - Key Performance Indicator
SIGIC - Sistema Integrado de Gestão de Inscritos para Cirurgia
Pentaho CE - Pentaho Community Edition
Pentaho EE - Pentaho Enterprise Edition
PRD - Pentaho Report Designer
PDI - Pentaho Data Integration
CDE - Community Dashboard Edition

Capítulo 1

Introdução

A informação é considerada um valor acrescentado necessário para planear e controlar as atividades de uma organização de modo eficaz. Os sistemas de informação, responsáveis pela transformação dos dados, têm a difícil tarefa de extrapolar a informação realmente importante e valiosa para o processo de tomada de decisão: os indicadores estratégicos. Estes são explorados através dos dados operacionais contidos na Base de Dados (BD) [6, 5].

Os sistemas de apoio à decisão emergiram na década de 80 e oferecem técnicas e meios de extração de conhecimento a partir de um conjunto de dados. A implementação de um sistema de Extração de Conhecimento (EC) nas organizações é, assim, crucial no sentido de suportar a tomada de decisão. A importância de tratamento dos dados aumenta continuamente e a necessidade das instituições tomarem rápidas decisões baseadas no "melhor" conhecimento possível fez surgir um novo conceito: *Business Intelligence* (BI). A BI pode ser definida como o processo de transformação de dados em informação e sua posterior transformação em conhecimento [6, 5].

Ao longo de muitos anos, o domínio médico tem sido alvo de uma vasta investigação por parte dos peritos em informática. A dinâmica e elevada quantidade de informação aumenta diariamente, obrigando as instituições de saúde a providenciar uma gestão eficiente do conhecimento. Deste modo, torna-se possível melhorar substancialmente os serviços de saúde disponíveis, diminuindo os custos e a quantidade de erros médicos ocorridos, tentando assim, tornar toda a informação automática, convertendo o máximo de informação possível em conhecimento útil para a instituição e, desta forma, suportar o processo de tomada de decisão [7].

As fontes de informação nas unidades de saúde, devido à infinita quantidade de informação, são distribuídas, extensas, heterogéneas e complexas. Deste modo, torna-se preciosa a integração dos diversos sistemas hospitalares colaborando para uma maior homogeneidade entre os sistemas clínico, médico e administrativo. Assim, a informação pode fluir, circular e difundir dinamicamente entre os diversos sistemas e deste modo contribuir para um melhoramento dos sistemas de informação na saúde [8].

A evolução demográfica e, em particular, o envelhecimento da população estão a alterar os padrões das patologias e a ameaçar a sustentabilidade dos sistemas de saúde. Nesse sentido, devem ser tomadas medidas, recorrendo à criação de máximo conhecimento possível para apoiar o processo de tomada de decisões. Atualmente a tomada de decisão é considerada uma das principais estratégias das organizações modernas [9].

Uma grande parte dos dados que documentam diariamente as operações efetuadas numa unidade de saúde está armazenada num Sistema de Gestão de Base de Dados Relacional (*Relational Database Management Systems - RDBMS*). Com o intuito de gerir a dispersão de informação, as unidades de saúde devem ser possuidoras de um sistema de EC, de forma a transformar dados brutos em conhecimento útil para a organização. Num processo de tomada de decisão é considerado de extrema importância tratar a qualidade dos dados. A qualidade interfere diretamente no sucesso do processo de EC [8].

O objetivo do processo de EC é a consolidação da informação a partir de diversas BDs num único *Data Warehouse* (DW), para depois, se proceder à análise e visualização da informação de forma a realizar e melhorar a tomada de decisão. Primeiramente, a informação deve encontrar-se organizada e estruturada para posteriormente ser submetida ao processo de EC. Este encadeamento de etapas dá-se pelo nome de processo ETL (*Extract, Transform and Load*). O objetivo de uma ferramenta ETL é a extração dos dados a partir de diversas fontes; proceder à sua transformação, limpeza e otimização de forma a remover inconsistências e definir regras para relacionar estes mesmos dados; e por fim carregar os dados modificados para um DW [10]. Numa fase posterior, e com o desenvolvimento dos sistemas de armazenamento de grandes quantidades de dados clínicos electrónicos, computadores potentes e algoritmos estatísticos, cresceu o interesse pelo domínio do *Data Mining* (DM). A aplicação bem sucedida de DM tem tido grande visibilidade em diversas áreas, e mais recentemente na área médica e da saúde pública [11, 12]. Os investigadores e especialistas adquirem conhecimento através da procura de padrões, formulação de teorias e de testes de hipóteses com observação [13].

Assim sendo, os especialistas médicos e clínicos devem adotar as ferramentas informáticas e sistemas inteligentes de modo a suportar a tomada de decisão. Toda esta perícia consiste em BI e, tal como explicado em [6], um processo de BI é composto por 3 passos fundamentais: a acumulação de dados *brutos* através do processo de *data warehousing*; o processamento de ETL em dados *brutos* num ambiente manobrável tal como um *Data Mart*; e a análise e os relatórios da informação para criar conhecimento imprescindível aos gestores. Em suma, torna-se importante monitorizar o funcionamento das instituições de saúde de forma a que estas procurem ser melhores no atendimento, nas instalações, nas condições de trabalho, na capacidade de resposta, nos custos, na satisfação dos utentes e na satisfação dos profissionais, devendo sempre colocar o paciente no centro das operações e articular todas as atividades de modo a garantir a máxima satisfação do utente.

Evidenciam-se nos hospitais problemas de elevados tempos de espera tanto para uma consulta de especialidade como para um tratamento cirúrgico, sendo que uma grande parte das reclamações na saúde em Portugal são devidas aos elevados tempos de espera e consequentemente à impossibilidade de marcação de consultas, contribuindo também decisivamente para tornar o número de episódios de urgência por habitante, um dos maiores da Europa Ocidental. Assim, um dos intuitos deste projeto de dissertação consiste na monitorização das listas de espera para consulta e para cirurgia de um subconjunto de dados extraídos de uma BD de um hospital do norte de Portugal. Por outro lado, outro dos grandes propósitos é o estudo da importância do processo de EC nas instituições de saúde, utilizando a tecnologia de BI e da ferramenta *open source* de BI, Pentaho. É pretendida uma análise e avaliação desta ferramenta tanto

ao nível de desenvolvimento das soluções como ao nível da sua usabilidade, de forma a concluir-se até que ponto é viável e eficiente a implementação de uma ferramenta de BI, *open source*, no seio de uma instituição hospitalar. Por conseguinte, é pretendida a simulação e monitorização de dados clínicos de forma a encontrar tendências e indicadores suscetíveis de apoiar a tomada de decisão e a implementação de medidas preventivas nos serviços de saúde.

Os objetivos iniciais do projeto prendiam-se com:

- Análise dos dados clínicos do hospital objeto de estudo;
- Avaliação do desempenho da ferramenta de BI Pentaho;
- Exploração da potencialidade e aplicabilidade do Pentaho numa instituição de saúde;
- Criação de padrões, indicadores através da técnica de *Data Mining* e de outras técnicas de análise de dados;
- Comparação do Pentaho com outras ferramentas existentes e atualmente implementadas em meio hospitalar.

No entanto, devido a questões de tempo e devido à complexidade de cada assunto em particular, os objetivos principais do projeto foram sendo ajustados ao longo do mesmo sendo que os principais objetivos prenderam-se com:

- Avaliação do desempenho da ferramenta de BI Pentaho;
- Exploração da potencialidade e aplicabilidade do Pentaho numa instituição de saúde;
- Análise a um nível superficial dos dados clínicos do hospital objeto de estudo;
- Aplicação de algumas técnicas de *Data Mining* também de forma superficial, para análise e criação de padrões e indicadores.

Este estudo está direcionado principalmente para os técnicos de informática e para o desenvolvimento técnico de soluções de BI. A EC é automatizada por sistemas de agentes e a monitorização é automática, permanente e constantemente atualizada, sendo que os profissionais de saúde apenas terão acesso à visualização dos dados para interpretação e procura de tendências/padrões. Por outro lado, será feito um estudo dos diferentes produtos oferecidos pelo Pentaho com uma abordagem da complexidade das ferramentas, ao ponto de ser ou não possível o usufruto por parte de um utilizador final tecnicamente menos especializado.

Esta dissertação encontra-se organizada da seguinte forma: o capítulo 2 onde são abordados os fundamentos teóricos subjacentes ao projeto; o capítulo 3 onde são introduzidos conceitos de *open source* e usabilidade, e onde são exploradas as diversas ferramentas *open source* de BI existentes atualmente, com especial destaque para o Pentaho (objeto de estudo); o capítulo 4 onde são apresentados os casos de estudo, é feita uma avaliação contínua da ferramenta ao longo dos casos e por fim uma apreciação global da ferramenta, dividida por funcionalidades; e finalmente, o capítulo 5 onde se apresentam as conclusões globais retiradas do projeto de dissertação apresentado e onde se enuncia possível trabalho futuro.

Capítulo 2

Enquadramento

Neste capítulo serão explorados e refletidos os fundamentos teóricos subjacentes ao problema. Desta forma, serão abordadas questões como o armazenamento e o tratamento de dados, a gestão da informação e do conhecimento no geral e, em especial, nos sistemas de saúde, o processo de EC, a BI e o funcionamento das listas de espera nas instituições de saúde.

2.1 Importância das TI na Saúde

Os processos de cuidados de saúde requerem cooperação e coordenação interdisciplinar. Eles dependem fortemente tanto da informação como do conhecimento. Neste contexto, a gestão da informação desempenha um papel muito importante. As Tecnologias de Informação (TI) encontram-se completamente integradas no mundo atual e auxiliam a ação humana em diversas atividades. São um dos principais motores da inovação. Existem inúmeros estudos que conseguem comprovar efeitos positivos nas instituições de saúde quando são usados sistemas de TI. A insuficiência de comunicação e a falta de informação encontram-se entre os principais fatores que contribuem para eventos adversos na saúde. O suporte das TI nas instituições de saúde têm o potencial de reduzir significativamente a taxa de eventos adversos e inesperados fornecendo de forma seletiva informações precisas e cuidadas correspondentes aos cuidados prestados. Porém, ainda hoje existe uma discrepância entre o potencial das TI e a sua real utilização na saúde.[14, 5]

A implementação de soluções de TI numa instituição é feita a dois níveis distintos: processos organizacionais (p. ex. entrada de ordem médica) e processos de tratamento médico (p. ex. diagnóstico e procedimentos terapêuticos). Enquanto que os padrões e procedimentos organizacionais auxiliam na coordenação entre os profissionais de saúde e a unidade organizacional, os processos de tratamento estão mais ligados ao paciente. Assim, diferentes desafios são tidos em conta dependendo destes dois diferentes níveis. Os processos de tratamento dependem fundamentalmente do conhecimento médico e das decisões específicas para o caso. Estas decisões são tomadas através da interpretação dos dados específicos de cada paciente de acordo com o conhecimento médico. Assim sendo, o processo de decisão é extremamente complexo uma vez que o conhecimento médico depende de diretrizes médicas e níveis de evidências, bem como da experiência individual de cada profissional. Exatamente por isto alguns autores consideram, então, que o processo de tomada de decisão médica não pode ser automatizado.

Por outro lado, considera-se cada vez mais de extrema importância a implementação de soluções de TI para suportar os processos organizacionais. Atualmente, grande parte do trabalho dos profissionais de saúde é afetado e sobrecarregado por tarefas organizacionais. Têm que se planear e preparar os procedimentos médicos, agendar consultas de diversas especialidades e serviços, transporte dos pacientes, marcação de visitas médicas de outros departamentos e a escrita, transmissão e avaliação de relatórios. Como consequência, ocorrem muitos erros e eventos adversos que se podem traduzir num aumento de custos para a instituição e na morte de pessoas. O suporte das TI não pode significar um plano restrito e pré-definido etapa-por-etapa, mas sim uma contribuição para fornecer a melhor evidência ao médico de um modo compreensível e rápido. O conhecimento médico explícito é sim importante e necessário, mas não suficiente para suportar a tomada de decisão. O processo de tomada de decisão é complexo e por isso as aplicações computacionais devem ser aceites, confiáveis e suficientemente integradas na prática clínica.[14]

2.2 Bases de Dados e de Conhecimento

A partir das BDs que contêm as operações diárias de uma empresa pode-se extrair conhecimento útil e de valor para apoiar a tomada de decisão. Na generalidade, as BDs são capazes de armazenar e manipular grandes quantidades de dados a uma velocidade considerável. Existem diversos motores de BDs tanto *open source* como proprietários. Os motores livres mais conhecidos são PostgreSQL, MySQL, SAP Database, entre muitos outros; enquanto que um dos motores proprietários com maior sucesso pertence à Oracle, também utilizado no âmbito deste projeto de dissertação para armazenamento, manipulação e consulta de dados.

Cada vez mais tem sido dada uma maior importância à gestão da informação e do conhecimento nas instituições com o intuito de implementar estratégias de sucesso, satisfazer os clientes/utentes, melhorar continuamente os processos, inovar os produtos e medir o desempenho instituicional. As bases de conhecimento clínico devem ser compreensivas, atualizadas permanentemente e de fácil usabilidade de forma a satisfazer as necessidades dos cuidados de saúde adequadamente. As aplicações informáticas médicas avançadas requerem uma ampla variedade de ativos de conhecimento, incluindo conjuntos de ordens, regras de interação entre medicamentos, diretrizes de prática médica e protocolos clínicos. Estes agentes do conhecimento suportam, por isso, as práticas de trabalho diárias dos profissionais de saúde, suportando simultaneamente a adoção de melhores práticas de trabalho e de estratégias baseadas em evidências. Assim, os sistemas clínicos com um sistema de apoio à decisão incorporado permitem reduzir a incidência de erros médicos e melhorar a qualidade dos serviços de saúde prestados, conduzindo a uma redução significativa dos custos. O sucesso de um sistema de apoio à decisão é em grande parte dependente da qualidade do conhecimento clínico a partir do qual foi construído. Tal como constatado eloquentemente por Matheson [15], "O grande desafio de abrangência informática que a sociedade enfrenta é a criação de sistemas de gestão de conhecimento que sejam capazes de adquirir, conservar, organizar, recuperar, visualizar e distribuir o que é conhecido hoje de uma forma que informa e educa, facilita a descoberta e criação de novos conhecimentos, e contribui para a saúde e bem-estar do planeta"[16, 5, 17].

O aumento de conhecimento clínico é notório, porém o processo de aceitação do

mesmo para suportar a prática clínica ainda é muito lento. Os profissionais de saúde apresentam alguma dificuldade em localizar a informação clínica e avaliar a sua relevância e credibilidade, tomando decisões baseadas neste conhecimento. Os técnicos de informática e os profissionais de saúde têm-se aliado no sentido de melhorar a qualidade dos processos de cuidados de saúde [16].

2.3 Extração de Conhecimento

Atualmente, os sistemas de informação empresariais demonstram uma elevada capacidade de armazenamento dos dados relativos a todas as atividades da organização. Porém, o mesmo já não se verifica na questão fulcral de análise e compreensão desses mesmos dados. Surge, assim, a necessidade de implementação de um sistema eficaz de EC no seio das organizações para auxiliar as funções de gestão e o processo de tomada de decisão. Desta forma, é urgente o desenvolvimento e otimização de métodos eficazes e eficientes para extrair a referida informação desconhecida das BDs [18, 13, 9].

A informação digital presente em grandes BDs é ubíqua, isto é, a informação encontra-se dispersa por diversos sistemas de armazenamento. Desta forma, considera-se que o que tem realmente valor é o conhecimento que se pode retirar a partir da concentração de informação e a sua utilização. Nas décadas de 80 e 90, os métodos tradicionais de conversão de dados em conhecimento consistiam numa interpretação e análise manual. Porém, a quantidade de dados armazenados tem aumentado exponencialmente, o que levou a que este tipo de análise se tornasse impraticável pois o processo era lento, caro e altamente subjetivo. A EC a partir de grandes BDs envolvia muitos passos e por isso se tornou, inicialmente, num desafio com algumas dificuldades inerentes como a manipulação dos dados e a recuperação da inferência de fundamentos matemáticos e estatísticos, de procura e de raciocínio. A EC apareceu, e continua a desenvolver-se a partir da interseção de diferentes campos de investigação como BDs, aprendizagem de máquina, reconhecimento de padrões, estatística, inteligência artificial, aquisição de conhecimento através de sistemas inteligentes, e visualização dos dados [18].

Por definição, o processo de EC é um "processo não-trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em dados". O termo processo significa várias etapas envolvendo pré-processamento dos dados, procura de padrões, avaliação do conhecimento e refinamento. Este processo é complexo e iterativo, isto é, dá-se em múltiplas iterações e envolve variadas tarefas. É não-trivial uma vez que envolve a procura de uma estrutura, de modelos, de padrões e de parâmetros. Por outro lado, os padrões devem ser considerados válidos porque o mesmo se deve aplicar a novos dados que surjam; novos para o sistema e para o utilizador; potencialmente úteis para a tarefa, problema e utilizador; e compreensíveis após o processamento [18, 9].

O processo de EC é, então, interativo e iterativo apresentando diversas etapas de processamento [18, 11, 9, 17, 7]:

1. **Compreensão do domínio do problema e seleção das variáveis** (identificação dos principais objetivos e metas de aplicação, bem como das técnicas mais adequadas ao problema).

2. **Seleção da BD e criação do conjunto de dados a ser processado** (seleção dos dados apropriados ao problema).

3. **Pré-processamento e limpeza dos dados** (inúmeras operações, tais como a remoção de ruído, agrupamento da informação de acordo com o problema, decisão de estratégias, mapeamento de valores em falta ou desconhecidos,... O pré-processamento envolve uma amostragem e verificação da qualidade dos dados de modo a garantir que se está na presença de informação limpa e bem descrita).

4. **Escolha das técnicas de análise de dados** (realização de DM, de análises *Online Analytical Processing* (OLAP - secção 2.3.3), ...).

5. **Interpretação e visualização do conhecimento** (compreensão dos padrões descobertos).

6. **Utilização do conhecimento extraído** (incorporação e integração do conhecimento num sistema de desempenho para auxílio da tomada de decisão). Esta etapa final é de extrema relevância para a utilização do conhecimento adquirido na prática clínica como um suporte de um sistema de apoio à decisão clínica.

Os sistemas de EC apresentam demasiados desafios que devem e têm sido ultrapassados de forma a tornar o processo mais eficiente. Estes são [18, 9, 11]:

- *BDs massivas e de elevada dimensionalidade.* BDs são da ordem de *terabytes* com milhões de registos e um elevado número de campos;

- *Interação com o utilizador e conhecimento prévio.* O processo deve auxiliar o utilizador na seleção apropriada das ferramentas e técnicas mais adequadas ao problema, dependendo do utilizador na decisão de caminhos a seguir em etapas intermédias;

- *Super-ajuste e avaliação da significância estatística.* Na procura dos melhores parâmetros para um modelo em específico muitas vezes acontece um uso limitado dos dados resultando num desempenho deficiente do modelo no teste do conjunto de dados;

- *Falta de dados.* Problema recorrente nas BDs empresariais, em que atributos importantes não se encontram registados na BD;

- *Incompreensibilidade de padrões.* Alguns modelos descobrem padrões incompreensíveis para a mente humana, tornando-se necessária uma representação gráfica, uma estruturação das regras, uma geração da linguagem natural ou técnicas de visualização de dados e conhecimento;

- *Gestão de dados e de conhecimento variável.* Os dados são não estacionários, isto é, estão em constante alteração e atualização, tornando os padrões descobertos anteriormente inválidos;

- *Integração.* As aplicações por si só não trazem vantagem ao utilizador, sendo necessária a sua integração com outros sistemas como ferramentas de visualização, interfaces de consulta em tempo real, tornando o processo de EC semi-automático senão mesmo completamente automático; por outro lado, não é possível a interligação entre diversos *datasets* e analisá-los através de um técnica de DM para identificar possíveis eventos inesperados;

- *Dados não padronizados, multimédia ou orientados a objetos.* As BDs não contêm apenas dados numéricos, muitos deles são do tipo textual, não numérico, geométrico, gráfico, texto livre, ...;

- *Conceção fragmentada.* Os modelos do processo não evidenciam as dependências importantes que existem num processo típico de EC; por dependências entende-se as relação existentes entre as vários etapas e tarefas do processo;

- *Falta de apoio nas tarefas da etapa de compreensão do negócio.* Esta é considerada uma das etapas mais importantes de todo o processo, pois tomam-se decisões que vão influenciar e definir todas as fases seguintes; falta de documentação e explicação de como definir e realizar a tarefas na implementação desta etapa.

A qualidade da informação interfere diretamente no processo de EC. A qualidade num modelo conceptual traduz a qualidade semântica do modelo tendo em conta o quão válido e completo este é no que diz respeito ao domínio do problema. A validação indica se a transmissão de informação através do modelo é correta e relevante para o problema, enquanto que a plenitude expressa se o modelo contém toda a informação correta e relevante sobre o domínio do problema [9, 8].

Aproximadamente 80% do tempo de análise de dados é gasto no processo de transformação. Por isso, a utilização de uma arquitetura de DW adequada pouparia cerca de 80% do tempo, o que torna imprescindível a limpeza dos dados para uma maior acessibilidade, interoperabilidade, utilidade, confiança e validação do DW [19].

Os sistemas de EC estão cada vez mais a ganhar um espaço nos sistemas de informação implementados nas instituições de saúde, com o principal objetivo de auxiliar a tomada de decisões, tanto clínica como administrativa, e na previsão de eventos. Em 2003, os autores de [11] foram uns dos pioneiros na aplicação de casos de estudo clínicos ao processo de EC e às técnicas de DM. É importante que as aplicações de EC quando aplicadas no contexto clínico apresentem determinadas caraterísticas, como aprendizagem *online*, processamento em tempo real, adaptabilidade, modelos de decisão e de DM, otimização de processos, integração de sistemas inteligentes e automatização de processos, precisão, garantia de privacidade e segurança, restrição de acesso externo e existência de políticas de utilização [12, 8, 20].

Importância da utilização de um sistema de EC na área da saúde [12]:

- *Sobrecarga de dados.* Os sistemas de armazenamento de dados clínicos contêm cada vez mais dados, dificultando a tarefa de descoberta de conhecimento através do olho humano e da simples navegação pelos dados.

- *Medicina baseada em evidências e prevenção de erros médicos.* Segundo dados relativos aos Estados Unidos, verificou-se que 87% das mortes podem ser prevenidas, podendo evitar-se erros por parte dos profissionais de saúde. Através da extração e análise dos dados, indicadores e tendências podem ser detetados e utilizados na melhoria da gestão hospitalar.

- *Mais valor na poupança e nos custos.* É possível a descoberta, por exemplo, de fraudes com cartões de crédito e anomalias nas reivindicações de seguro.

- *Prevenção e descoberta precoce de doenças.* Através da utilização de DM e de técnicas de visualização os especialistas clínicos podem encontrar padrões e indicadores de anomalias.

- *Gestão de pandemias de doenças e formulação das políticas na saúde pública.* Os especialistas utilizam, atualmente, as técnicas de EC com o intuito de prever precocemente e gerir situações de pandemias.

- *Apoio na tomada de decisão e diagnóstico não-invasivo.* Alguns procedimentos de diagnóstico e laboratório são invasivos, caros e custosos para os pacientes. Os autores de [21] utilizaram o algoritmo de agrupamento K-Means para analisar pacientes com cancro cervical e perceberam que o agrupamento de similaridades obtinha melhores resultados que uma opinião médica (considerada subjetiva e dependente do indivíduo). Foi encontrado um conjunto de atributos que poderiam ser utilizados pelos médicos para auxiliar a tomada de decisões.

2.3.1 Processo ETL

O processo ETL (*Extract* – Extração, *Transform* – Transformação, *Load* - Carregamento) tem como principal objetivo extrair dados provenientes de uma BD, encontrando-se representado o seu processamento global na Figura 2.1. De forma sucinta, os dados são, então, extraídos e processados a partir de diferentes fontes de dados; posteriormente, propagam-se para uma instância denominada *Data Staging Area* (DSA); sofrem uma transformação/modificação e são "limpos" para finalmente serem carregados para uma segunda BD, o DW. Considera-se o processo ETL a etapa mais crítica interveniente na construção de um DW, uma vez que grandes quantidades de dados são processados. Além disso, torna-se muitas vezes numa complexa combinação entre processamento e tecnologia conduzindo ao consumo de uma parte significativa do DW [22, 1, 23].

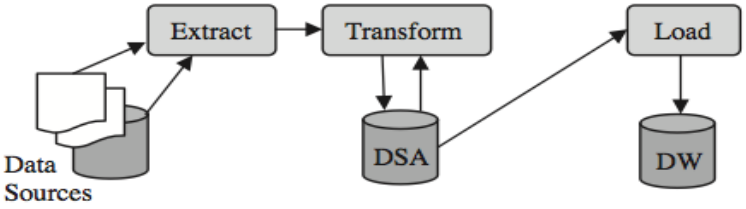


Figura 2.1: Esquema geral do processo ETL. Retirada de [1]

Tal como o próprio nome indica, o processo é efetuado através de 3 fases que consistem em: extrair dados de diferentes fontes de informação; proceder à sua transformação; e, finalmente, carregar a informação para um DW [22, 1, 24].

Extração: Os dados são provenientes de fontes de informação distintas, sendo por isso necessário que o processo recorra de forma efetiva à integração de sistemas, que tenham diferentes plataformas, linguagens, sistemas operativos, sistemas de gestão da BD e protocolos de comunicação. Esta fase consiste em 2 etapas: extração inicial e extração de dados alterados. A extração inicial representa o primeiro conjunto de dados extraídos e o seu carregamento no DW. Paralelamente, a extração de dados alterados corresponde à atualização contínua do DW.

Transformação: Nesta fase ocorre a limpeza, integração e otimização dos dados de entrada de forma a obter dados precisos, que são corretos, completos, consistentes e inequívocos. De salientar que os dados modificados não afetam as fontes originais mas apenas a data contida na extração para o repositório DW.

Carregamento: Neste passo final, os dados extraídos e transformados são escritos para uma estrutura multidimensional precisamente endereçados pelos utilizadores finais e pelos sistemas aplicacionais. Na fase de carregamento são processados mapeamentos sintáticos e semânticos entre esquemas, sendo considerada uma fase mais complexa de efetuar uma vez que está dependente da heterogeneidade da BD utilizada.

Após o carregamento, os dados permanecem disponíveis no DW para, depois, serem processados e darem origem a conhecimento útil para o processo de tomada de decisão. Por vezes, os dados são armazenados num passo intermédio, *Data Marts*, que consolida pequenas partes de informação relacionada entre si, como se definirá mais à frente. Posteriormente, o *Data Mart* precisa de estar ligado e cria o DW com informação relevante em função da especificidade de determinado assunto [22, 23].

Uma ferramenta ETL deve ser eficiente na comunicações entre as diversas BDs e deve conter mecanismos capazes de ler diferentes formatos. Por outro lado, a ferramenta deve ser desenhada e projetada de forma a ser facilmente modificada. Uma vez que as fontes de informação estão constantemente em mudança e por isso o DW atualiza periodicamente, o processo ETL vai-se alterando e desenvolvendo também. Outros aspetos importantes e a ter em consideração na escolha de uma ferramenta ETL são: o suporte à plataforma (o sistema deve ser capaz de ser executado independentemente da plataforma em uso); o tipo de fonte independente (deve ser capaz de ler o ficheiro diretamente da fonte de dados); o apoio funcional (deve apoiar em diversas tarefas tais como a extração de dados a partir de inúmeras fontes, a limpeza dos dados, a transformação, agregação, reorganização e carregamento); a facilidade de uso pelo utilizador; o paralelismo (deve ser capaz de correr código em paralelo de modo a facilitar a execução dos processos; é também pertinente que a ferramenta distribua tarefas por múltiplos servidores em caso de sobrecarga); o apoio ao nível do *debugging* (deve apoiar a limpeza da lógica de transformação e possibilitar ao utilizador visualizar os dados antes e depois da transformação); a programação (deve ser possível o agendamento de tarefas ETL para melhor aproveitamento do tempo); a implementação (deve ser capaz de implementar os objetos ETL em ambiente de teste ou de produção); e a reutilização (deve reutilizar a lógica de transformação de forma a que o utilizador não tenha permanentemente que reescrever a mesma lógica de transformação). Também, e não menos importantes, devem ser tidos em conta os seguintes aspetos na escolha de uma ferramenta ETL: Integração (integração do processo ETL em sistemas e processos já existentes, não exigindo assim grande esforço); Interfaces (as fontes de dados devem ser suportadas pelos sistemas); Editor gráfico (de modo a facilitar o trabalho do utilizador, é importante a modelação gráfica do processo ETL); Funcionalidade (ter em conta as funcionalidades das ferramentas); Suporte (existência do suporte do vendedor ou então de foruns *online*); Documentação (extremamente importante a existência de documentação com qualidade sobre as ferramentas); e *Up-to-dateness* (as ferramentas devem conter uma base desenvolvedora ativa, garantindo o desenvolvimento futuro a tempo) [22, 1, 25].

2.3.2 *Data Warehouse*

O sistema de um DW consiste essencialmente em agregar informação proveniente de uma ou mais fontes de dados de forma a se tratar, organizar e consolidar esta mesma informação numa única estrutura de dados. Um DW é, por isso, um repositório de dados através do qual ocorre a pesquisa de valiosa informação de negócios em grandes BDs, ou seja, constrói-se através de um processo de extração de dados a partir de diferentes aplicações, internas ou externas, realizando de seguida a estruturação da informação para ser armazenada no repositório central (DW). Neste contexto, a capacidade analítica do DW está dependente do tipo de dados disponíveis nas fontes [22, 10, 26].

De realçar que os termos *Data Warehouse* e *Data Warehousing* são distintos. Se por um lado um DW é considerado um repositório "inteligente" através do qual pode derivar o processo BI, por outro o processo de *Data Warehousing* corresponde ao desenvolvimento, gestão, métodos operacionais e práticas que determinam o modo como os dados são coletados, integrados, interpretados, geridos e utilizados pelos gestores que têm o poder de tomar decisões [27].

No processo de *Data Warehousing*, muitas vezes a informação é organizada em pequenos repositórios denominados *Data Marts* (Figura 2.3). Um *Data Mart* é um subconjunto de um DW que está projetado para um propósito específico, similar à forma como um DW pode ser personalizado para uma determinada organização. Os *Data Marts* servem de armazenamento de dados cumulativos a partir de outras BDs. Em termos funcionais, um *Data Mart* é utilizado para realçar relações complexas entre as diferentes fontes de dados. São agregadas grandes quantidades de informação que são, frequentemente, confidenciais [6].

Um DW deve ser adaptativo, ou seja, no sentido de lidar com alterações rápidas e frequentes das atividades, estratégias e ambientes de negócios, o DW deve então ser capaz de se modificar também. Por outro lado, deve ser capaz de facilitar e simplificar o manuseamento e a gestão da BD reduzindo a quantidade de pessoal responsável pela administração e manutenção da mesma. Um DW deve ter a capacidade de suportar simultaneamente transações pequenas (realizadas pelos utilizadores) e transações grandes (executadas pelos agentes de software) durante o processo de carregamento de dados. As principais vantagens da implementação e utilização de um DW são de: fornecer uma única fonte de dados para o negócio; oferecer informação precisa, relevante e oportuna para uma efetiva tomada de decisões; projetar uma solução de DW que possa ser escalável e extensível através da organização; e identificar e resolver o problema da qualidade e limpeza de dados. Os dados são extraídos de fontes variadas, diferentes, heterogéneas e distribuídas, sendo posteriormente a informação resultante usada em consultas/*queries* e relatórios. Esta arquitetura de armazenamento de dados baseia-se num modelo de dados multidimensional, o qual permite a análise dos dados a partir de diversas perspetivas e providencia aos gestores um elevado poder de tomada de decisão. A visualização da informação num esquema multidimensional baseia-se na dicotomia medida/dimensão e é caraterizada pela representação da informação segundo um espaço n-dimensional, isto é, com tantos eixos quantas as dimensões com interesse para análise. Um DW pode ter diferentes modelações multidimensionais dependendo do esquema de representação que é implementado: esquema em estrela, em floco de neve ou em constelação. Tradicionalmente, a multidimensionalidade é classificada segundo dimensões, níveis e atributos (explicado detalhadamente na secção 2.3.3), dando

Base de Dados Operacional	**DataWarehouse**
Tempo Crítico	Informação histórica
Acesso a leitura/escrita	Acesso a leitura
Acesso a poucos registos de cada vez	Acesso a grande quantidade de registos de cada vez
Atualização da informação em tempo real	Atualização periódica da informação
Estruturado para OLTP	Estruturado para OLAP

Tabela 2.1: Base de Dados Operacional vs Data Warehouse [5].

origem ao esquema multidimensional. Este paradigma permite que se compreenda e visualize a informação a partir dos diferentes pontos de vista de análise de um determinado assunto. Deste modo, um DW garante o fornecimento de informação consistente, integrada, organizada e histórica, preparada para posterior análise quando submetida a um sistema de BI e utilizada para processos de tomada de decisão no seio de uma organização. Através do armazenamento de informação histórica, é possível que o DW disponibilize informação acerca da evolução organizacional ao longo de um determinado período. Considera-se importante realçar que esta informação não é modificada como acontece nos sistemas transacionais, mas sim acrescentada/aumentada [28, 10, 5].

O DW disponibiliza, assim, uma visão global e detalhada da organização e sendo pretendida a sua exploração através de ferramentas específicas, surgiram vários mecanismos de navegação e análise de desempenho. Uma das ferramentas mais conhecidas e com maior relevância para exploração de uma DW é a análise OLAP (*Online Analytical Processing*) que facilita a navegação e análise de informação através dos dados comerciais baseando-se no paradigma multidimensional já referido. O DW suporta o processamento analítico *online* (OLAP), os requisitos funcionais e de desempenho que diferem consideravelmente daqueles do processamento transacional *online* (OLTP - *Online Transactional Processing*), aplicação habitualmente suportada por BDs operacionais. Um DW tende a abranger ordens de magnitude superior às de uma BD operacional: enquanto uma BD tende a ter centenas de *megabytes-gigabytes* de tamanho, um DW geralmente alberga tamanhos de *gigabytes-terabytes*. Para além disso, um DW é orientado ao problema, integrado, objetivo, não volátil, dependente do tempo e não normalizado. As principais diferenças entre uma BD operacional e um DW estão apresentadas na Tabela 2.1 [10, 9, 24, 5].

No projeto e implementação de um DW para uma organização devem ser tidos em conta variados fatores como: custos, tempo, utilizadores, pessoal, hardware e serviços. Os custos encontram-se relacionados com o montante que a organização pretende gastar em *hardware* e que tipo de *software*, ferramentas e serviços do fornecedor serão necessários. O fator tempo importa no sentido de se conhecer a quantidade de tempo que demora o projeto do DW e quanto tempo a organização realmente tem. Em relação aos utilizadores é importante saber qual a utilização e os objetivos concretos dos utilizadores finais em relação ao DW a implementar. O fator pessoal refere-se àqueles que desenvolvem e realizam a manutenção do DW. Finalmente, o *hardware* traduz-se nas ferramentas necessárias para a construção do DW, e os serviços considerados extra que possam vir a ser precisos ao longo do processo. Por outro lado, de forma a construir-se um modelo conceptual de um DW, é necessário analisar os requisitos de consulta,

as alterações da estrutura de dados, o tempo de resposta, a gestão dos recursos, ferramentas de interface, e verificação e possuir um sistema capaz de realizar cópias de segurança e recuperação de dados. O processo de desenvolvimento de um DW consome tempo e sem o auxílio das ferramentas adequadas pode tornar-se bastante prolongado. A construção de um DW não necessita da adição de nova informação nas fontes, mas de um rearranjo desta. Desta forma, o *Data Warehousing* não passa de uma prática de visão estratégica sobre os dados de uma organização [10].

As instituições de saúde são consideradas ambientes com um elevado grau de automação e, por isso, podem ser grandes beneficiadores da implementação de um DW clínico. A elevada disponibilidade e confiança na tecnologia atual torna o DW relevante nestes campos de aplicação. Um DW, combinado com relatórios e ferramentas de consulta, pode fazer surtir resultados bastantes promissores, apesar de todas as dificuldades de implementação inerentes bem como da necessidade de pessoal altamente especializado no assunto. Há cerca de 2 décadas, começaram a existir vários estudos e atualmente diversas aplicações também baseadas no conceito de DW implementado e utilizado na área médica [26, 29, 30].

2.3.3 Análise e Visualização

Existem diversas ferramentas de exploração e análise de BI que permitem a obtenção, visualização e apresentação da informação armazenada no DW, de forma a constituir conhecimento para apoiar no processo de tomada de decisão. Nesta secção serão abordadas: a análise OLAP (já introduzida na secção 2.3.2, definida no contexto da multidimensionalidade de um DW); consultas (*queries*) e relatórios; *Data Mining*; e *dashboard* (painel gráfico). Estas ferramentas, com principal enfâse para os *dashboards*, constituem grande potencial para visualizar o estado dos indicadores chave de rendimento (KPI: *Key Performance Indicator*) dentro da organização e em tempo real.

Análise OLAP

Os dados contidos em BDs OLTP são considerados de extrema importância do ponto de vista operacional, embora a sua disposição não permita o auxílio na tomada de decisões. Neste contexto, surge a tecnologia de análise OLAP com o propósito de obter informação proveniente das BDs para assistir no processo de tomada de decisão. São transformados dados transacionais em dados interpretáveis de acordo com a lógica do esquema de DW e do esquema OLAP. Significa, assim, a análise de grandes quantidades de dados em tempo real. O termo *online* implica que apesar de estarem envolvidas enormes quantidades de dados, o sistema deve responder às consultas efetuadas de modo rápido para permitir uma exploração interativa e dinâmica dos dados. A análise OLAP permite aos utilizadores o acesso a informação organizada, podendo formar subconjuntos de dados numa estrutura multidimensional que possam responder a questões específicas. O OLAP emprega a técnica denominada de Análise Multidimensional (análise do hipercubo - Figura 2.2) de dados, que permite uma visão mais rápida e interativa destes. Nos modelos de dados multidimensionais existe um conjunto de medidas numéricas que constituem os objetos de análise. Cada medida numérica é associada a um conjunto de dimensões, que dá o contexto para a medida. Cada dimensão é descrita por um conjunto de atributos, sendo que estes podem estar

relacionados por uma hierarquia de relações. O conjunto de dimensões, hierarquias e medidas designa-se de cubo [31, 28, 32].

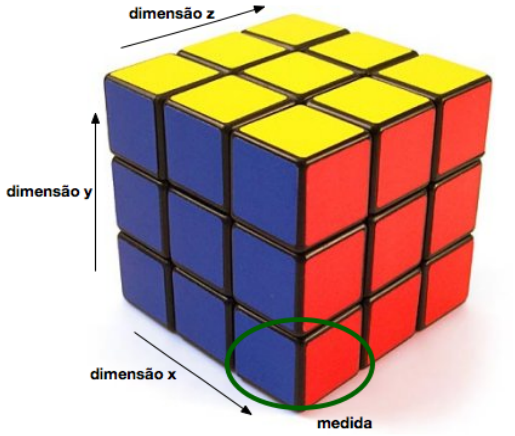


Figura 2.2: Representação do cubo multidimensional para análises OLAP.

Uma ferramenta OLAP facilita a navegação e análise da informação através dos dados possibilitando a tomada de decisão, na medida em que melhor conhecimento traduz-se em melhores decisões. Quanto maior for o número de medidas de análise, mais e mais variadas perspetivas existirão para compreender e analisar a informação disponível. A grande vantagem de uma ferramenta OLAP é a enorme flexibilidade de realizar relatórios. O tempo de resposta de uma consulta é mais importante do que o rendimento da transação uma vez que a consulta é considerada a principal utilização de uma BD OLAP. Numa aplicação OLAP, os dados não necessitam de estar normalizados como numa OLTP e, geralmente, a BD não tem uma grande número de tabelas. Porém, as tabelas são muito mais extensas. Isto deve-se ao fato de numa BD OLAP a principal operação ser o *insert* (inserção de dados), enquanto que o *delete* (eliminação de dados) e o *update* (atualização de dados) são utilizados apenas quando ocorrem erros de carregamento de dados [10, 33].

Dependendo da BD em que os dados estão armazenados, os sistemas OLAP podem ser classificados como [31, 5]:

- ROLAP: armazena dados em BDs relacionais. A grande parte dos sistemas ROLAP usam um esquema estrela para representar o modelo de dados multidimensional. A BD consiste numa única tabela de fatos e numa tabela para cada uma das dimensões. Cada linha da tabela de fatos consiste num apontador para cada uma das dimensões que fornece as coordenadas multidimensionais, armazenando as respetivas medidas numéricas; e cada tabela de dimensões consiste em colunas que correspondem aos atributos das dimensões.

- MOLAP: suporta a visualização multidimensional de dados através de um sistema de armazenamento que utiliza um *array* multidimensional. Apresenta como principais vantagens o tempo de resposta a consultas e as propriedades de indexação excelentes e tem como aspeto negativo a utilização para armazenamento.

- HOLAP: sistema híbrido que combina os dois sistemas anteriores, dividindo os dados num sistema MOLAP e num sistema de armazenamento relacional (ROLAP).

Beneficia da grande escalabilidade do ROLAP e da velocidade de processamento do MOLAP.

As operações permitidas para análise OLAP incluem [34, 5]:

- *Roll-up* (subir): aumento do nível de agregação, ou seja, permite agregar os dados visualizados no cubo utilizando uma dada hierarquia, sendo que cada repetição de análise ocorrerá a um nível mais elevado de agregação.
- *Drill-down* (descer): operação oposta do *roll-up*; diminui o nível de agregação e aumenta o detalhe ao longo de uma ou mais hierarquias de dimensões; o seu objetivo é fornecer uma visão mais pormenorizada dos dados que estão a ser analisados.
- *Slice and dice* (cortar e reduzir): seleção e projeção, restringindo a informação a visualizar.
- *Pivot* (rodar): re-orienta a visualização multidimensional dos dados, permitindo visualizar e interatuar com a informação do cubo OLAP rodando o eixo de visualização dos dados.

Consultas (*queries*) e relatórios

Na maioria das aplicações empresariais a presença de um componente de relatórios torna-se extremamente importante e necessário para organizar a informação e para a distribuir entre os colaboradores para que estes tenham acesso a toda a informação empresarial relevante. A consulta e relatórios são ferramentas para reportar e listar em detalhe ou sobre informação agregada proveniente do DW ou *Data Mart*. Geralmente, em vez de ter que se aprender uma linguagem de consulta tal como o SQL, os utilizadores podem usar menus e botões para especificar os elementos, as condições e outros atributos através de operações simples facilitadas por um ambiente gráfico intuitivo. Tipicamente, as consultas e os relatórios fornecem informação tal como média, desvio-padrão e outras funções básicas de análise. A opção pela qual o utilizador opta depende essencialmente das necessidades específicas apresentadas pela organização e da informação que se pretende extrair, tendo em conta o acesso a detalhes da BD no menor tempo e custo possíveis [35].

Data Mining

O termo DM é muitas vezes considerado pelos autores como o próprio processo de EC, porém no âmbito desta dissertação, DM será referente à utilização de técnicas estatísticas numa etapa do processo de EC. Na prática, inerente ao DM está a EC uma vez que mesmo que se realize a técnica de DM sobre um conjunto de dados é pertinente que estes sejam submetidos a todas as etapas de EC para melhorar a descoberta de padrões [11, 13].

Define-se, então, DM como a aplicação de técnicas estatísticas a um conjunto de dados, como por exemplo, modelos preditivos. São retirados de uma BD os dados referentes a um determinado problema e são realizadas análises estatísticas com o intuito de se "prosperar" estes dados e encontrar informação adicional desconhecida [11].

Segundo Turban [36], o DM é "o processo que utiliza técnicas estatísticas, matemáticas, de inteligência artificial e de aprendizagem-máquina para extrair e identificar informação útil e subsequente conhecimento a partir de extensas BDs". Assim, o DM é um campo que conjuga inteligência artificial, gestão de BD, visualização de dados, aprendizagem-máquina, algoritmos matemáticos e estatística. Esta tecnologia oferece diversas metodologias para a tomada de decisão, resolução de problemas, análise, planeamento, diagnóstico, deteção, integração, prevenção, aprendizagem e inovação.

A grande parte das técnicas estatísticas utilizadas no processo de DM são os métodos convencionais de análise de dados. Porém, as técnicas utilizadas no processo de EC diferem ligeiramente no sentido em que não formulam uma hipótese nula ou prévia, nem efetuam cálculos complexos. Por esta razão, considera-se que caso não seja produzido qualquer sinal no final da análise não é então possível afirmar se o sinal não existe ou se há insuficiência nos dados em análise. Basicamente, uma vez definidos os objetivos de análise e o conjunto de dados, as técnicas estatísticas têm como prinicipal propósito a descoberta de padrões, associações ou similaridades entre os grupos de dados sujeitos para análise, de modo a gerar um sinal ou a detetar informação nova [11].

Os algoritmos de DM são considerados como composições de técnicas e princípios básicos e consistem nos 3 componentes seguintes [18]:

- **O Modelo**: existem dois fatores relevantes, a função do modelo (classificação, agrupamento,...) e a forma de representação do modelo (função linear de variáveis múltiplas, função de densidade da probabilidade de Gaussian). O modelo contém parâmetros que são determinados a partir dos dados.

- **O Critério de Preferência**: preferência de um modelo ou conjunto de parâmetros em deterimento de outros, dependendo dos dados.

- **O Algoritmo de Procura**: consiste na especificação de um algoritmo para procurar modelos e parâmetros particulares, em função dos dados, do modelo e do critério de preferência.

Os modelos de representação incluem regras e árvores de decisão, modelos não-lineares (p. ex. redes neuronais), métodos baseados em exemplos (p. ex. vizinho mais próximo e raciocínio baseado em casos), modelos de dependência gráfica probabilística (p. ex. redes Bayesianas) e modelos de atributos relacionais. A representação dos modelos determina a flexibilidade do modelo em representar os dados e a interpretabilidade do modelo em termos humanos. Tipicamente, os modelos mais complexos são capazes de encaixar melhor os dados porém são de mais difícil compreensão. Por exemplo, as árvores de decisão podem ser bastante úteis na procura da estrutura em espaços de elevadas dimensões e em problemas com dados categóricos e contínuos combinados [18].

O algoritmo de procura, também comummente chamado de sistema de aprendizagem, é o método utilizado para o processo de EC. Nem sempre é uma tarefa simples determinar qual o método que melhor se adapta ao problema e ao tipo de dados uma vez que nenhum método é inteiramente eficaz em todas as áreas, porém existem determinados parâmetros através dos quais se deve basear a escolha: tipo de aprendizagem, paradigmas de aprendizagem, linguagens de descrição,... A qualidade do modelo utilizado para realizar a técnica de DM influencia fortemente a eficácia e eficiência da

implementação do processo deEC, bem como do resultado final obtido. A aproximação consiste em selecionar um conjunto de dados (*dataset*) a partir do DW, efetuar sofisticadas análises de dados no conjunto selecionado para identificar caraterísticas estatísticas e de forma a construir modelos preditivos. Finalmente, estes modelos de previsão são desenvolvidos na BD operacional [9].

Uma seleção, pré-processamento e transformação cuidada dos dados é importante para o sucesso da análise e dos resultados obtidos. Devem ser tidas em conta questões como: quais as variáveis a selecionar; quais as relações e a informação contextual a utilizar; quais as medidas que se devem usar; saber se os dados selecionados são suficientes e adequados para representar a complexidade e natureza do problema [13].

Uma ferramenta de DM consiste numa previsão automática de tendências e comportamentos, permitindo uma descoberta automática e previsão de padrões desconhecidos. Existem inúmeros algoritmos utilizados em DM, entre eles: redes neuronais artificiais, árvores de decisão, algoritmos genéticos, indução de regras,... As técnicas mais importantes de DM e também as utilizadas nesta dissertação são a classificação, a segmentação, a associação e a visualização.

Modelagem preditiva / Classificação e Previsão: Técnica utilizada no desenvolvimento de um modelo para relacionar uma variável dependente com um conjunto de variáveis independentes, similar a uma análise de regressão múltipla. Existem dois tipos de modelagem preditiva: classificação, para variáveis dependentes categóricas; e previsão de valor, para variáveis dependentes contínuas. A **classificação** é o agrupamento de dados em classes (categorias) de acordo com as suas propriedades (valores de atributos) e tem como principais objetivos atribuir uma classe a um objeto, prever a classe de novas observações com base na EC de acontecimentos passados. Torna-se, assim, apropriada quando o objetivo é prever a relação entre um grupo de novos registos baseada nas suas características (variáveis dependentes). Para tal, necessita de um *dataset* de treino para treinar (configurar) o modelo de classificação, e uma *dataset* de teste para avaliar o desempenho do modelo treinado. A classificação é também designada de classificação supervisionada (ao contrário da classificação não-supervisionada - agrupamento). Para que seja possível proceder à classificação, as classes devem ser discretas, devem existir em número finito, não ter ordem e devem ser sempre identificadas por um nome. Os métodos de classificação incluem, por exemplo, árvores de decisão, redes neuronais artificiais, estimação da vizinhança máxima (algoritmo largamente utilizado na área de decisão clínica), função de discriminação linear, métodos do vizinho mais próximo e raciocínio baseado em casos. No contexto médico, as técnicas de classificação podem ser usadas como ferramentas de segunda opinião para apoiar as decisões clínicas. A **previsão de valor** utiliza tanto a classificação como a regressão para prever o resultado final, como por exemplo no caso dos dados relativos a um paciente é possível prever as suas características sócio-económicas ou demográficas [11, 13, 37, 7].

Segmentação ou *clustering*: A segmentação utiliza um algoritmo que analisa os dados e avalia a similaridade entre os registos formando grupos (*clusters*) de dados similares entre si e com comportamentos semelhantes. Pares de registos são comparados com valores de campos individuais dentro deles, sendo que agrupamentos dentro de grupos conduz à ordenação rápida e eficaz dos dados em grandes BDs. A segmentação poderia ser, por exemplo, usada num grupo de pacientes que apresentem sintomas ou diagnósticos similares de forma a determinar se existe alguma associação de medica-

mentos. O agrupamento é uma técnica interessante quando o objetivo consiste em reduzir uma grande amostra de dados num conjunto de grupos heterogéneos específicos e mais pequenos, sem perder informação significativa. Os métodos de agrupamento podem ser classificados em 2 grupos: agrupamento de particionamento e agrupamento hierárquico. Os métodos de agrupamento de particionamento, tal como o K-Means, dividem o conjunto de dados num número de grupos sem sobreposição. Um item de dados é atribuído ao grupo "mais próximo", baseado na medida de proximidade e de similaridade. O agrupamento hierárquico, por outro lado, organiza os dados num hierarquia com uma sequência partições ou grupos [11, 13, 20].

Associação: Esta técnica refere-se a métodos que identificam padrões frequentes, associações, correlações ou estruturas ocasionais no conjunto de dados. As regras de associação são usadas para encontrar elementos que ocorrem conjuntamente em *datasets*. Baseiam-se num tipo de regra 'se x então y', descobrindo padrões de comportamento ou identificando sequências de tempo similar de eventos. A confiança corresponde à percentagem com que uma ocorrência antecedente está relacionada com uma ocorrência consequente (medida de certeza) enquanto que o suporte indica a frequência com que uma regra de associação surge no conjunto de dados (medida de utilidade). É aconselhável dar maior importância às regras que apresentam maior valor de suporte [11, 13].

Visualização: Consiste no desenvolvimento de teorias e métodos para facilitar a formação de conhecimento através da exploração visual e da análise de dados e da implementação de ferramentas visuais para a subsequente recuperação de informação. Para processar grandes quantidades de dados e visualizar padrões gerais, as aproximações visuais são normalmente combinadas com métodos computacionais (tal como o agrupamento, a classificação, a associação) para sumariar a informação e auxiliar os utilizadores na descoberta de indicadores e padrões desconhecidos [13].

Concluindo, o processo de DM realiza transformações nos dados, tarefas de classificação, regressão, regras de associação e algoritmos de agrupamento (*clustering*), que podem ser utilizadas para uma melhor compreensão do negócio e melhorar o desempenho futuro da organização através de análises preditivas.

A aplicação da técnica na área médica é um desafio devido às indissincrasias da profissão médica porém estudos relacionados já comprovaram que tal é possível e viável. Na investigação clínica, DM começa com uma hipótese sendo os resultados ajustados de forma a ir de encontro com a hipótese. Enquanto que a técnica tradicional de DM se preocupa essencialmente com a descoberta de padrões e tendências nos conjuntos de dados, na área médica é também considerada de extrema importância atentar na minoria que não se enquadra nos padrões normais. Para além disso, na medicina é muito importante que se encontrem esforços no sentido de providenciar explicações para os padrões e tendências descobertos pois uma pequena diferença pode significar a morte do paciente. Os principais desafios a serem considerados na aplicação do DM à saúde são: incompletude e ruído dos eventos clínicos, a quantidade de variáveis de processo que necessitam de ser claramente distinguidas, e os casos médicos excecionais. Em conclusão, estes processos são dinâmicos, complexos e multidisciplinares. Os sistemas de informação que suportam os processos de prestação de cuidados de saúde e registam toda a atividade clínica contêm dados altamente valiosos para serem submetidos a análises e, daí, retirarem conclusões e conhecimento útil para a instituição [12, 20].

Dashboard (Painel Gráfico)

Um *dashboard* é uma ferramenta que alinha os objetivos das diferentes áreas com a estratégia da organização e a monitorização do seu progresso. Esta ferramenta pode ter inúmeros usos, podendo variar segundo avaliações pessoais, exercícios de treino e planos de negócios. Existem dois tipos de *dashboards*: analítico e integral. Os *dashboards* analíticos permitem a obtenção, a partir do DM, de relatórios e principais indicadores de desempenho (KPI). São operacionais ou táticos e analisam as áreas de negócio não relacionadas entre si. Na prática, são uma ferramenta de consulta que visam a obtenção e apresentação de indicadores para a gestão. Os *dashboards* integrais (também designados de *Balanced Scorecard*) são desenvolvidos num nível estratégico por toda a organização, e os diferentes níveis de gestão e liderança possuem visões estratégicas com um conjunto de objetivos e indicadores que cobrem toda a organização [9].

2.4 *Business Intelligence*

A BI tem sido considerada nos últimos anos uma área de investigação e desenvolvimento com grande impacto nas empresas modernas, conduzindo estas ao aumento e melhoria do desenvolvimento de produtos, políticas de gestão e estratégia de negócios. Hoje em dia, as organizações começam a perceber a importância da gestão da informação e as vantagens competitivas que implicam o seu uso. A BI surgiu na década de 1970 (não sendo completamente consensual a data do seu aparecimento) e tinha por caraterística a exaustiva programação, o que implicava elevados custos no seu desenvolvimento. Enquanto que inicialmente a BI era utilizada nas organizações apenas com a finalidade de suporte estratégico à tomada de decisão, atualmente é largamente usada para suporte das mais variadas atividades de negócio, como melhorias dos processos operacional e tático, cadeia de fornecedores, produção e apoio ao consumidor. A BI é vista como a chave estimuladora do crescente aumento de valor e desempenho empresarial, e define-se sucintamente como o processo de transformação dos dados em informação e da transformação desta em conhecimento. Este processo é monitorizado por equipas de gestão e consultores de TI, e tem como principal objetivo aproveitar os dados organizacionais de forma a suportar análises e tomadas de decisão inteligente, responsável e de forma empírica. A BI é reconhecida como tendo grande relevância para o proveito das empresas, sendo que está comprovado que aquelas que não operam sistemas de BI podem perder competitividade. Por outro lado, cresce a necessidade de se implementarem sistemas que suportem as tarefas de BI em tempo real, conduzindo a que a tomada de decisão se baseie nos dados operacionais. O principal propósito da BI em tempo real é reduzir a latência existente entre o momento em que os dados operacionais são adquiridos e quando é que a análise dos mesmos se procede [25, 6, 34, 31, 38].

Os sistemas de BI referem-se a uma importante classe de sistemas de análise de dados e relatórios de forma a possibilitar aos gestores das diferentes camadas organizacionais o acesso a informação relevante, proveitosa e em tempo útil para uma melhor tomada de decisão. Eles combinam dados com ferramentas de análise, de forma a constituir conhecimento útil e relevante para a tomada de decisões. A entidade IDC[1]

[1] http://www.idc.com

estimou em 2011 que o gasto total em sistemas BI por parte das organizações atingisse os 7 biliões de dólares, cerca de 5 mil milhões de euros. A escala de investimento em sistemas de BI é refletida na crescente importância estratégica e realça a necessidade de iniciar e desenvolver um maior número de projetos de investigação [38, 5].

Com o crescimento dos negócios e das organizações em escala, a grandeza de dados incluídos nas aplicações de BI aumentam similarmente, aumentando a complexidade dos desafios de desempenho da análise dos dados. Em resposta a esta complexidade, a BI baseia-se no processo de EC (secção 2.3) e é composta por três etapas principais (Figura 2.3): a acumulação de dados brutos através de *Data Warehousing*, o processo ETL e a análise e relatório da informação para criação de conhecimento para suporte à decisão. O componente ETL representa aproximadamente 80% de todo o consumo dos projetos de BI. Uma das atividades mais significativas no âmbito da BI constitui a arquitetura e construção dos armazéns de dados, o DW [25, 34].

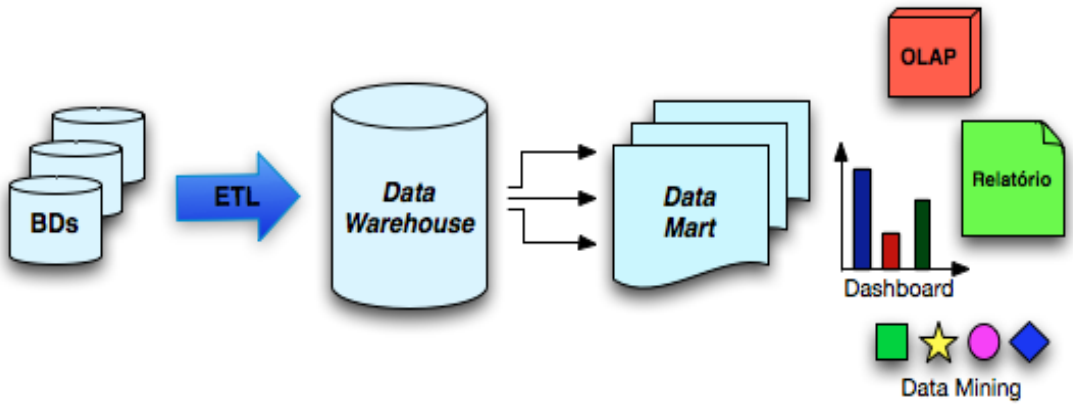


Figura 2.3: Esquema representativo do processo de BI.

Atualmente, a BI é utilizada em diversos campos de aplicação como no fabrico de expedição de pedidos e apoio ao cliente, no retalho para criação de um perfil de utilizador, nos serviços financeiros para análise de reclamações e deteção de fraudes, nos transportes para gestão de frotas, nas telecomunicações para a identificação das razões da rotatividade de clientes e nos cuidados de saúde para análise de resultados. Os sistemas de BI são conhecidos pela sua elevada aplicabilidade, necessitando apenas de determinadas configurações para se ajustar aos diferentes tipo de empresas, instituições ou negócios. Na área particular da saúde, os sistemas de informação das instituições prestadoras de cuidados de saúde devem ser projetados e desenhados com o intuito de suportar ambos os processos clínicos e administrativos, integrando e coordenando a prática médica. Porém, parte destes sistemas devem ser repensados uma vez que ainda se verifica atualmente uma grande escassez de interoperabilidade, aspeto de elevada importância no que concerne a processos de cuidados de saúde [31, 38, 20].

O conceito de Gestão do Relacionamento com o Cliente (CRM – *Customer Relationship Management*) está inteiramente associado à BI, consistindo em diretrizes, procedimentos, processos e estratégias através dos quais as organizações conseguem estabelecer a interligação das interações dos consumidores. Assim, é importante manter o consumidor no centro do negócio como uma das tendências chave na indústria. Em analogia, esta política deve ser também implementada na saúde considerando sempre

o paciente como o ponto central [19].

O núcleo da BI consiste na análise OLAP e nos Sistemas de Informação Empresarial (SIE). Estas componentes suportam diretamente a tomada de decisão. Na teoria, uma aplicação de BI pode ser construída com estes 2 componentes apenas. Porém, a BI orientada para a análise abrange vários conceitos e aplicações que permitem a análise de dados baseada no modelo e no método. Com isto, para além dos 2 componentes nucleares, incluem-se o desenvolvimento de relatórios *ad hoc*, componentes de DM bem como outras funcionalidades avançadas, por exemplo a análise CRM. O processo integral de BI para além dos componentes já referidos, necessita de outros dois imprescindíveis para qualquer solução potente de BI: um DW para armazenar os dados para análise e ferramentas de ETL [25].

Em conclusão, os sistemas de BI são definidos como ferramentas altamente especializadas para análise de dados, consulta de dados, relatórios, análises OLAP e *dashboards* que suportam a tomada de decisão organizacional potencializando o desempenho dos processos empresariais. As suas principais tarefas são a elaboração de previsões baseadas em dados históricos e fatos passados, a criação de cenários que evidenciem o impacto da alteração de determinadas variáveis e a análise detalhada da instituição, obtendo um conhecimento mais profundo da mesma. Os sistemas BI são complementados por infraestruturas de TI, incluindo DW, DM e ferramentas ETL. As ferramentas de BI de código aberto (*open source*) têm ganho, recentemente, uma maior ênfase e importância, começando a competir no mercado com soluções comerciais, tal como se poderá ler na secção 3.2 [38, 25, 5].

2.5 Listas de Espera em Sistemas Hospitalares

A existência de listas de espera é um dos aspetos determinadores do acesso à prestação de cuidados de saúde. O fato de existirem listas de espera pode-se concluir como o resultado de uma incapacidade do sistema de saúde em satisfazer as necessidades elementares de saúde dos cidadãos [2, 39]. Considera-se que a inexistência de listas de espera é impossível, sendo no entanto possível reduzir a sua duração média para valores mínimos.

As listas de espera geram preocupações quer ao nível da eficiência que ao nível de equidade, existindo conflitos frequentes entre as duas dimensões quando se coloca, por exemplo, a questão de pagamento entre pacientes para troca de lugares na fila. Este tema é assim bastante controverso, levantando também questões éticas [2, 40].

O funcionamento das listas de espera em Portugal baseia-se no esquema da Figura 2.4. O paciente é submetido a uma consulta de cuidados primários, geralmente ocorrida nos centros de saúde. Perante a necessidade de prosseguir para uma consulta de especialidade no hospital é emitido pelo médico um documento designado de P1 (atualmente registado eletronicamente) que permite que o paciente seja alocado na lista de espera para uma consulta de especialidade. Posteriormente, caso a decisão de tratamento tomada seja a necessidade de intervenção cirúrgica, este prossegue para a lista de espera para cirurgia [2, 41].

O tempo de espera total no percurso de um doente é outro aspeto que levanta algumas questões entre a comunidade. Teoricamente, este tempo é dado pela soma das parcelas de todos os tempos de espera desde que um paciente visita pela primeira vez um médico até que o tratamento é finalizado. Porém, para efeitos de valores

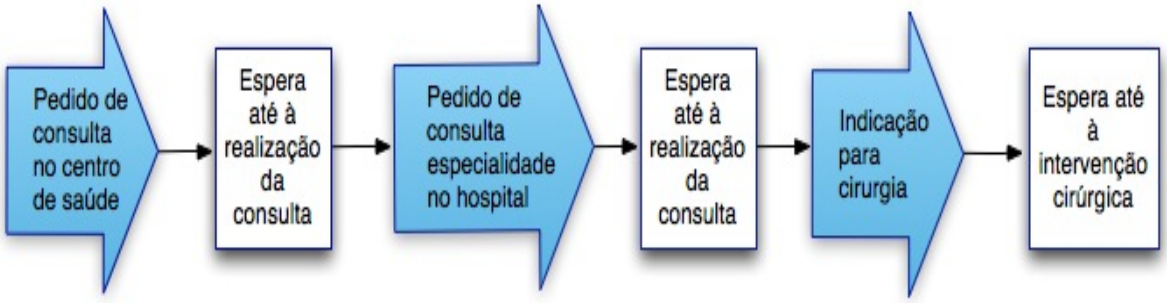


Figura 2.4: Esquema representativo do funcionamento das listas de espera em Portugal. Adaptado de [2]

hospitalares o tempo de espera foca-se na espera desde que existe a necessidade de uma consulta de especialidade até à realização da mesma, e desde que se decide prosseguir para uma intervenção cirúrgica até ao agendamento da mesma. Para uma organização prestadora de cuidados de saúde é inevitável a existência de listas de espera para aceder aos cuidados de saúde [2].

Observa-se que para o conjunto de países pertencentes à OCDE, as listas de espera para cirurgia tendem a ser mais acentuadas nos países que combinam seguro de saúde [39]. Em Portugal, tem-se registado um aumento das listas de espera nas últimas décadas, apresentando-se como possível explicação o avanço das tecnologias na área cirúrgica e de anestesia, aumentando assim o alcance, a segurança e eficiência dos procedimentos cirúrgicos e por conseguinte a procura por intervenções cirúrgicas. Segundo o Observatório Português dos Sistemas de Saúde o "aumento da procura de cuidados de saúde, na maior parte dos países europeus, é explicada pelas alterações demográficas verificadas nos últimos 20 anos, a qual tem sido acompanhada por um correlativo aumento da expectativa de melhoria da qualidade de vida. A conjugação destes fatores colocou os sistemas de saúde numa situação paradoxal" [2, 42].

Existem alguns fatores que podem conduzir à existência de listas de espera para cirurgia, estando entre eles: o aumento da procura de intervenções cirúrgicas (explicado pelo envelhecimento da população, expectativas dos utentes, opções clínicas associadas à introdução de novas tecnologias,...); a definição da capacidade de oferta (número de salas operatórias existentes e disponíveis); e a eficiência da instituição de saúde (explicada por fatores como a organização funcional dos serviços de saúde, a incerteza/variabilidade do desempenho clínico, o empenho dos recursos humanos, a distribuição irregular dos recursos...) [2, 42].

O tempo considerado ótimo de espera por uma cirurgia não é zero, no sentido em que pode-se tornar eficiente para o estabelecimento de saúde em termos de custos ter pequenas filas de espera para cirurgias do tipo programado. Na saúde, as consequências adversas existentes para os ligeiros atrasos são pouco significativas e, deste modo, poupa-se na capacidade dos hospitais aquando da formação de filas de espera. Obviamente, a noção de tempo aceitável de espera por um tratamento clínico depende diretamente da condição do doente e da patologia identificada. Há, por isso, uma lista de espera que suaviza as flutuações de chegadas de novos doentes alocados de acordo com a prioridade no sistema permitindo uma melhor programação da atividade. Ge-

ralmente, as listas de espera não se regem por um comportamento de *first in – first out*, tendo sempre em consideração o aparecimento de casos com um grau de urgência mais elevado. Torna-se, então, claro que o custo da capacidade não utilizada pela instituição (chamada de desperdício) é maior do que o custo associado a algum tempo de espera. Concluindo, é de realçar novamente que não é uma situação ótima ter, em média, tempos de espera nulos, situação em que não existiriam listas de espera [2, 39].

Nos últimos anos, tem-se vindo a proceder à implementação de programas de combate às listas de espera para cirurgia. Os programas aplicados em Portugal dividem-se em dois grandes grupos: os programas que têm como principal objetivo estimular a realização de atividade adicional, disponibilizando verbas para cirurgias adicionais; e os programas que pretendem gerir pormenorizadamente os doentes inscritos para cirurgia, de forma a conhecer-se em cada momento a situação destes (no que se refere ao número de inscritos e aos tempos de espera) [2].

Neste contexto, pretende-se apenas analisar a última política (SIGIC – Sistema Integrado de Gestão de Inscritos para Cirurgia) implementada no seio das instituições de saúde portuguesas em 2004. Essencialmente, o SIGIC é um sistema de informação que tem como propósito principal a redução dos tempos das listas de espera através de uma melhor gestão das mesmas. Tem como principais objetivos melhorar o serviço, aumentar a eficiência, gerar equidade no acesso e criar conhecimento e transparência. Neste âmbito, passou a existir um tempo máximo de espera (1 ano) aceitável para o paciente ser tratado, sendo que se o respetivo tempo limite fosse ultrapassado o paciente teria um cheque de cirurgia com o direito de ser operado noutro estabelecimento de saúde, mesmo que privado. A filosofia passou então a ser a resolução da questão das listas de espera, através de um melhor desempenho das instituições hospitalares da zona geográfica dos pacientes. Com a implementação do SIGIC, verificou-se uma melhor programação da atividade e conhecimento das situações correntes através do sistema de informação contribuindo assim para uma gestão mais eficaz dos recursos e uma maior capacidade de resposta do sistema de saúde (Figura 2.5). Contudo, o sistema de informação para a gestão das listas de espera por si só não resolve o problema fulcral, sendo que se torna essencial a existência complementar de um desenho do sistema de saúde de forma a garantir maior produtividade e produção [2, 43].

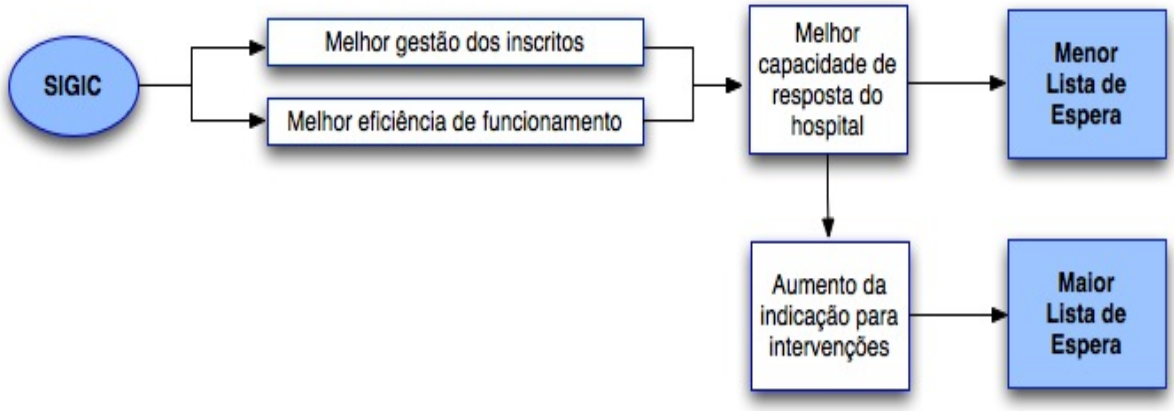


Figura 2.5: Efeitos do SIGIC sobre as listas de espera. Adaptado de [2]

A criação do SIGIC permitiu o acesso à informação sobre o número de pessoas

alistadas para cirurgia e o tempo de espera até à intervenção. Este tipo de informação deslocou a discussão do número de pessoas em lista de espera para o tempo mediano de espera (correspondente ao tempo que o paciente da lista de espera terá ainda que aguardar até ser submetido à intervenção). Assim, a apresentação dos indicadores referentes às listas de espera baseiam-se no tempo mediano de espera, sendo que a tendência mais recente é a redução dos tempos médios de espera. Em termos de resultados obtidos, valores relativamente recentes adiantam uma redução tanto do número de inscritos para cirurgia como do tempo mediano de espera. Do ponto de vista da eficiência da realização das cirurgias, quanto maior for a eficiência maior será o número de pacientes tratados num mesmo espaço de tempo, o que permite a redução do tempo de espera para uma intervenção cirúrgica. Perante uma maior capacidade de resposta do hospital, podem-se contemplar três dimensões importantes: eficiência produtiva, eficiência de custos e a qualidade dos serviços prestados.

De uma modo geral, hospitais mais eficientes com melhor capacidade operacional, no sentido de um menor custo unitário por cirurgia realizada, têm menores listas de espera e menores tempos de espera para intervenções cirúrgicas [2, 41].

Atualmente, já é possível aceder-se à informação relativa à situação na lista de espera para cirurgia através do programa e-SIGIC. Pela Internet, é possível conhecer a posição que se ocupa na lista, bem como o tempo dentro do qual se prevê que a intervenção cirúrgica seja realizada. Desta forma, consegue-se assegurar a transparência e o rigor do processo, melhorar o acesso à informação e centrar o sistema no paciente[2].

[2] http://crohnsnews.wordpress.com/2009/12/23/e-sigic-portugueses-passam-a-\
aceder-a-sua-situacao-na-lista-de-inscritos-para-cirurgia/

Capítulo 3

Ferramentas de Business Intelligence

Neste capítulo será feito um estudo e uma exploração das ferramentas *open source* de BI, começando por apresentar a teoria de código aberto, apresentando-se em seguida um estado atual das ferramentas de BI, fazendo uma descrição e exploração do Pentaho e dos seus módulos constituintes e finalizando com a apresentação do termo usabilidade, fulcral na avaliação de desempenho do Pentaho nas soluções obtidas e no âmbito deste projeto.

3.1 *Software Open Source*

Com a urgência atual de implementar novas aplicações informáticas na área médica e da saúde surgiu uma nova abordagem às aproximações de cógido aberto (*open source*). O *software open source* é um modelo de desenvolvimento no qual o código fonte está inteiramente disponível aos utilizadores para visualização, leitura, modificação e redistribuição sem as restrições de direito de propriedade do produto. Assim, o desenvolvimento de aplicações *open source* permite um progresso viável das aplicações para os cuidados de saúde. As aplicações *open source* prometem acelerar a difusão das soluções informáticas nos cuidados de saúde, diminuir custos de desenvolvimento, estimular a inovação por parte de indivíduos que se possam encontrar geograficamente distribuídos e aumentar a usabilidade das aplicações. Porém existem alguns fatores que influenciam tal evolução como as capacidades funcionais, patrocínio, licenças *open source* entre outras características [44, 45].

O *software open source* difere do proprietário fundamentalmente nos processos de desenvolvimento e nas licenças do produto. As aplicações *open source* são desenvolvidas de forma colaborativa e distribuída alavancando a *internet* para a coordenação. Todas as aplicações de código aberto encontram-se licenciadas por uma licença *open source*, que proporciona ao utilizador o direito de utilização do *software*, acesso e modificação do seu código fonte e a redistribuição desse mesmo *software*. Este tipo de licenças são numerosas e complexas. A escolha da licença para *software* livre na saúde e informática médica é importante porque determina os direitos dos utilizadores e influencia o estímulo dos desenvolvedores para participar num projeto, a qualidade do produto e a vontade dos utilizadores na adoção de determinada aplicação. Todavia, um dos grandes paradoxos do *open source* é a explicação da motivação dos indivíduos que contribuem para projetos deste tipo sem qualquer remuneração financeira [44, 45].

As aplicações *open source* promovem a inovação no sentido em que facilitam aos

utilizadores a modificação, customização e reutilização do código fonte de forma a obterem novas soluções. A confiança de um produto *open source* é bastante elevada uma vez que o acesso ao código fonte por parte de todos permite que se identifiquem problemas e se proponham soluções. O *software* aberto promete também outros benefícios, especificamente quando aplicado na área da saúde, como a barreira de venda do produto ou a independência do proprietário da tecnologia, aumentando também as possibilidades de escolha, a flexibilidade e a interoperabilidade entre soluções abertas. Em termos de custos, as organizações podem economizar nas taxas de licenciamentos e reduzir gastos na compra de material informático específico do vendedor proprietário de determinado produto. Contudo, as organizações necessitam de desenvolver e formar pessoal especializado na adoção e gestão de soluções *open source*. Este tipo de situações tem que suportar custos de pessoal altamente especializado, implementação, manutenção e suporte podendo conduzir a uma verba de custos superiores à adoção de soluções proprietárias [44, 45].

No mercado mundial, estima-se que em 2012 a adoção de plataformas *open source* seja 3 vezes superior às proprietárias, sendo que a linguagem de programação mais utilizada, atualmente, no desenvolvimento das aplicações é Java. As soluções abertas têm vindo a aumentar a sua oferta no mercado e a ganhar um espaço importante também nas instituições de saúde, obrigando a um patrocínio contínuo das partes interessadas (utilizadores, desenvolvedores e gestores) conduzindo a uma crescente adoção deste tipo de aplicações em informática médica e na saúde [44].

3.2 Estado da Arte

A BI providencia aos seguidores da técnica uma mudança no modo como as pessoas trabalham, conseguindo competir no mercado mais eficaz e eficientemente. Assim, com o avanço tecnológico tem-se registado na última década um crescente aparecimento de novas soluções de BI, com especial enfoque para as de código aberto. São ferramentas capazes de serem integradas em qualquer organização e ambiente de negócios que com o passar do tempo têm sido cada vez mais enriquecidas com integração de outros produtos e com o desenvolvimento de novas funcionalidades.

Nos dias que correm, cada vez mais surgem no mercado novas soluções *open source* de BI com enorme potencial e capazes de competir com as soluções proprietárias. As principais ferramentas de BI a realçar são o Talend Open Studio[1], o SpagoBI[2], o JasperSoft[3] e o BIRT[4]. Talend, fundada em 2005, foi a primeira solução comercial *open source* de integração de dados. É uma ferramenta de código aberto versátil e com elevado potencial constituída por uma versão da comunidade e uma versão comercial. Contém vários componentes: Integração da Dados (Talend Integration Suite), Gestão da Qualidade dos Dados (Talend Open Profiler), Controlo e Visualização Gráfica dos Processos (Talend Open Studio's Job Designer), entre outras. O Spago é uma ferramenta de BI totalmente de código aberto (existindo uma única versão), constituída por diversos módulos como uma plataforma de BI (SpagoBI Server), ambiente de desenvolvimento de soluções (SpagoBI Studio), análise de metadados (SpagoBI Meta),

[1] http://www.talend.com
[2] http://www.spagoworld.org/xwiki/bin/view/SpagoBI/
[3] http://www.jaspersoft.com
[4] http://www.eclipse.org/birt/phoenix/

integração de ferramentas externas (SpagoBI SDK) e conjunto de modelos analíticos (SpagoBI Applications). A ferramenta de BI intitulada de Jasper, desenvolvida em 2001 nas linguagens Java e Perl, incorpora soluções de ETL, OLAP e relatórios todas integradas numa plataforma comum. Em analogia com o Pentaho, o Jasper inclui ainda ferramentas independentes de ETL (JasperETL), de análise (JasperAnalysis) e relatórios (JasperStudio). Tal como o Pentaho, o Jasper utiliza o servidor Mondrian para a realização de análises OLAP. Esta ferramenta é também uma das mais utilizadas a nível mundial e, por isso, talvez a maior concorrente do Pentaho possuindo uma interface bastante atrativa e extensa documentação. Desenvolvida em 2005 com a parceria da Eclipse e da IBM, a ferramenta BIRT (*Business Intelligence Report Tool*), tal como o próprio nome indica foca as suas soluções no desenvolvimento de relatórios sem qualquer grau de complexidade, adequados às necessidades do negócio e com capacidades de análises OLAP. A ferramenta é constituída por dois componentes principais: um desenhador de relatórios visual baseado em Eclipse e um componente para gerar relatórios com rotina que pode ser integrado em qualquer ambiente Java. O projeto BIRT possui também um motor de construção de gráficos integrado no desenvolvimento de relatórios. Por último, as ferramentas proprietárias de BI com maior sucesso e presença no mercado pertencem à Oracle (Oracle BI) e à Microsoft (*Business Intelligence Development Studio* - BIDS).

Atualmente, as organizações utilizam e adotam as ferramentas de BI como imprescindíveis no seu fluxo de negócios. Cada vez mais se tem verificado um aumento da implementação de sistemas de BI em computação *cloud* (Cloud BI), possibilitando assim a redução de gastos ao nível de *hardware*. Nesta dissertação a ferramenta BI explorada foi o Pentaho (apresentada e descrita na secção 3.3) uma vez que é neste momento a ferramenta líder no mercado *open source* de BI e é aplicada em diversas áreas como transportes, farmácias, comercial, recursos humanos, tecnologia, indústria, saúde, gestão empresarial, serviços financeiros, entre muitas outras. Em relação à área da saúde este é um projeto pioneiro no sentido em que existe muito pouca ou praticamente nenhuma documentação de estudos ou projetos da implementação do Pentaho na saúde. Foram encontrados escassos casos, em particular o testemunho de Joshua Greenbaum [46] que trabalhou com o Pentaho na área da saúde e conseguiu obter resultados positivos. Outros casos de sucesso foram encontrados no estudo de construção de um DW [6] e na realização de análises OLAP [32].

3.3 Pentaho

O *software* Pentaho[5] foi desenvolvido pela Pentaho Corporation em 2004, totalmente em Java, tendo sido a primeira plataforma BI tão abrangente a ser projetada e lançada como a alternativa *open source* do mercado. O Pentaho é um serviço *Web* que oferece a arquitetura e a infraestrutura necessárias para a construção de soluções de BI. Tal como se encontra representado na Figura 3.1, a Pentaho BI Suite disponibiliza áreas de aplicação sobre dados e integração de aplicações desde a camada inferior de integração de aplicações à camada superior de apresentação passando pelas camadas intermédias das funcionalidades e administração de processos. A maioria dos utilizadores finais interatua com a camada superior de apresentação dos resultados e soluções.

[5] http://www.pentaho.com/

Esta solução de BI tornou o Pentaho na suíte BI de código aberto líder e a mais amplamente desenvolvida a nível mundial, com mais de 8 milhões de *downloads* já efetuados e projetos em variadas organizações mundiais de mais de 185 países.

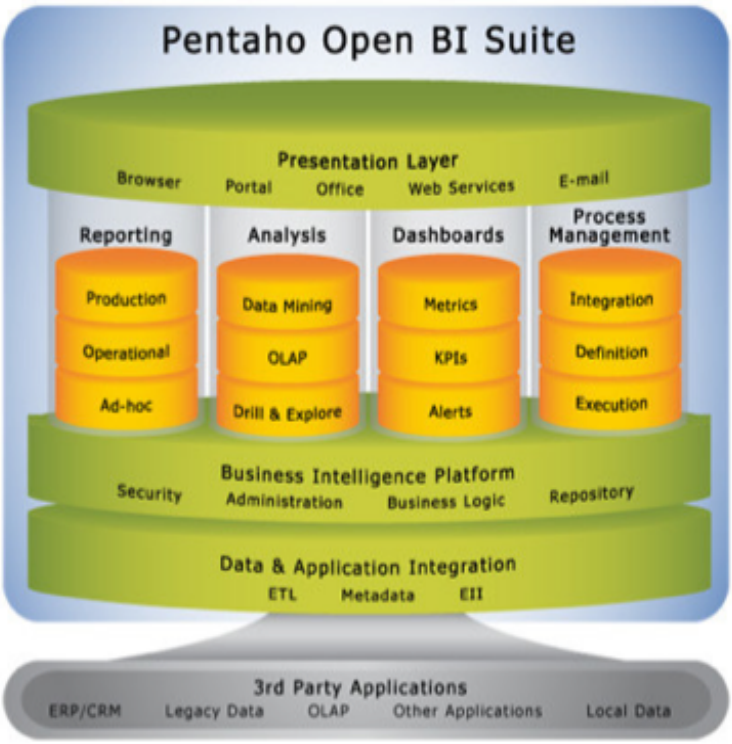


Figura 3.1: Pentaho Open BI Suite. Retirada de [3]

O projeto Pentaho BI Suite engloba um conjunto de produtos: plataforma de BI (servidor), relatórios, análises OLAP, integração de dados (ETL), *dashboards* e *data mining* [3, 34].

3.3.1 Plataforma de BI

A arquitetura da plataforma Pentaho BI baseia-se num motor de execução de sequências de ações, denominadas de *Xactions*, definidas em XML através da ferramenta gráfica Pentaho Design Studio (extensão do IDE Eclipse). A partir de um *input* lançado pelo utilizador é executada uma sequência de ações bem definidas.

É um servidor *Web* que pode ser acedido via uma interface *Web* e permite aos utilizadores interagir com o servidor. A plataforma inclui um motor de soluções que integra a apresentação de relatórios, análises, *dashboards* e componentes de DM de forma a constituir uma plataforma de BI completa e sofisticada. Fornece vários serviços para os utilizadores finais, tais como, autenticação, registo, auditoria, fluxo de trabalho, serviços *Web*, motores de regras, alerta de agendamento de subscrições, notificação e integração de ferramentas para o utilizador final, gestão de processos e integração de segurança centralizada. O servidor de BI é um servidor *Web* compatível com J2EE e utiliza uma instância Tomcat pré-configurada. Tomcat é um pacote *open source* Java Servlet desenvolvido pela The Apache Software Foundation. Para aceder à plataforma existe um suporte completo de utilizadores definidos e a função de cada um. Existe,

também, a PAC (*Pentaho Administration Console*) onde estão definidas as contas de utilizadores e as respetivas funções e onde é possível se criarem e autenticarem novas contas.

O Pentaho disponibiliza duas versões de plataformas para o Pentaho Suite: a edição da comunidade (Pentaho CE) baseada numa licença GPL e a edição empresarial (Pentaho EE) baseado num modelo de subscrição. A principal diferença entre as duas baseia-se maioritariamente no tipo de suporte oferecido, sendo que a EE contém mais componentes e produtos que a CE. Neste projeto foram utilizadas as duas edições para o desenvolvimento de relatórios, *dashboards* e análises. A EE apresenta uma funcionalidade do *software* melhorado, suporte técnico profissional, conhecimentos sobre o produto, *software* certificado e manutenção do produto. Ao longo do desenvolvimento dos casos de estudo vai-se fazendo uma comparação entre as duas edições, tanto ao nível técnico como da sua usabilidade. A plataforma CE pode facilmente ser integrada e direcionada para as PMEs devido ao código livre e utilização grátis, enquanto que a EE pode ser integrada numa empresa de maior dimensão uma vez que abrange várias áreas de BI.

3.3.2 Pentaho Reporting

Com esta solução Pentaho, torna-se possível desenvolver facilmente um relatório, permitindo às organizações o acesso, formatação e distribuição da informação entre o corpo de colaboradores, consumidores e parceiros.

A Pentaho Suite possui possibilidades diferentes para o desenvolvimento de relatórios: *Web Ad-Hoc Query and Reporting* (WAQR), *Pentaho Reporting* (pertencente ao servidor CE) e *Pentaho Interactive Report* (pertencente ao servidor EE), e a ferramenta *Pentaho Report Designer* (PRD). A WAQR não foi utilizado no âmbito deste projeto. Permite o desenvolvimento de relatórios, utilizando apenas dados pré-definidos por modelos de metadados via *Web*. Apesar de ser um produto rápido é bastante limitado na utilização dos dados e também não suporta gráficos.

Por sua vez, os produtos dos servidores e o PRD apresentam funcionalidades similares, sendo que o PRD contém obviamente uma maior número de opções de desenvolvimento, tendo mais potencial. Estes produtos permitem a costumização dos relatórios de acordo com as preferências do utilizador e suportam gráficos. O PRD não se baseia na metodologia WYSIWYG (*What You See Is What You Get*) e, por isso, o utilizador tem que se adaptar a um ambiente de trabalho que não corresponde ao resultado final do relatório. Por este motivo, o desenvolvimento de relatórios no PRD exige alguma aprendizagem e compreensão prévia por parte do utilizador avançado. O PRD possui também a funcionalidade de relatórios *Wizard* proveniente de um assistente que mediante um processo simples, passo a passo, permite editar as caraterísticas mais comuns para a construção de relatórios. O desenvolvimento é assim mais rápido e fácil, bastando para tal escolher a fonte de dados, a estrutura do relatório e os campos e cálculos associados sendo lançado pelo assistente o resultado de um relatório usado para visualização. Esta funcionalidade destina-se principalmente a utilizadores iniciais ou que não pretendem relatórios demasiado elaborados.

3.3.3 Pentaho Analysis

A Pentaho Suite possui um conjunto de opções para procedimentos de análises OLAP, entre elas Pentaho Analysis (edição da comunidade), Pentaho Analyser (edição empresarial) e a ferramenta Mondrian (servidor OLAP). O Mondrian é então um servidor para análises OLAP desenvolvido em Java e suportado pela biblioteca JPivot[6] para navegação e análise dos dados. Executa *queries* MDX (*Multidimensional Expressions*), permite a leitura de dados a partir de BDs relacionais (RDBMS), e permite a visualização dos resultados num formato multidimensional através de uma Java API. Os módulos Pentaho Analysis e Analizer são baseados no projeto JPivot,uma interface *Web-based* extremamente útil para a realização de análises multidimensionais de dados. Neste projeto não foi explorada a ferramenta Mondrian uma vez que as diferenças entre estas soluções de análise são mínimas, por questões de tempo e porque não se encontrava nos objetivos principais da dissertação. Constata-se que as principais diferenças entre a ferramenta Mondrian e o módulo de análise do servidor remetem-se simplesmente a questões de segurança (o Mondrian tem suporte de regras de segurança que limita os utilizadores no acesso a dados) e aos tempos de resposta (o Mondrian possui capacidades de *cache* e *buffer* que otimizam o seu desempenho).

As soluções Pentaho enunciadas proporcionam excelentes análises OLAP auxiliando os utilizadores no processo de tomada de decisão e providenciando o conhecimento geral do negócio. Estas soluções facilitam a tarefa dos utilizadores na exploração e estudo da informação empresarial, explorando e cruzando os dados. Oferecem também uma integração completa com todos os outros produtos disponíveis na Pentaho Suite. Em todos os módulos, a consulta de dados realiza-se através da linguagem MDX, que se traduz em consultas SQL sobre a BD relacional. A MDX é uma linguagem que expressa as seleções, os cálculos e as definições de alguns metadados (informação relativa a outros dados) numa BD OLAP.

O Pentaho Mondrian pode ser útil na agilização dos tempos de resposta das consultas MDX efetuadas referentes a um cubo OLAP que, por sua vez, é descrito mediante um arquivo de metadados que mapeia o esquema armazenado na BD relacional codificado em XML, podendo ser construído manualmente pela ferramenta Mondrian Schema Workbench. Para além do Mondrian, a Pentaho Suite é composta pelo Schema Workbench e o Aggregation Designer. O Schema Workbench consiste numa ferramenta gráfica para criação do cubo OLAP, gerando o esquema multidimensional concordante com o cubo projetado. O Aggregation Designer é utilizado para desenhar tabelas agregadas de forma a aumentar o desempenho do cubo.

3.3.4 Pentaho Data Integration (PDI)

É uma ferramenta poderosa para a execução do processo ETL, utilizando uma aproximação inovadora orientada a metadados, com mais de 100 objetos de mapeamento. Para além disso, é uma aplicação amigável do utilizador no sentido em que possui um ambiente intuitivo, gráfico, esquema arrastar-soltar (*drag-and-drop*), tendo também uma arquitetura comprovada, escalável e baseada em padrões. A sua interface gráfica denomina-se de Spoon e permite a extração, transformação e carregamento de informação de origens e destinos diferentes. O PDI é constituído pelos diferentes blocos de

[6] http://jpivot.sourceforge.net

construção:

Step: Um *step* possui uma entrada de dados e uma saída de dados sendo que estes são canalizados através do *step*. Cada *step* tem uma função que pode ser a de alterar, filtrar ou até verificar dados para identificar padrões. Existem inúmeros *steps* no PDI.

Transformação: Uma transformação é um conjunto de *steps*. Quando é construído o esquema de uma transformação com todos os *steps* ligados entre si e com uma entrada e saída de dados apropriadas o utilizador pode executar a transformação. Esta execução é acompanhada por um ficheiro de *log* que permite auxiliar o utilizador em possíveis erros que possam ocorrer. Uma transformação tem que ter sempre um *step* para a entrada de dados, um *step* para a saída de dados e entre estes um ou mais *steps* para a transformação dos dados.

Job: Um *job* consiste em uma ou mais transformações que vão ser executadas pela ordem desejada. É importante realçar que os dados não vasam entre transformações, estas apenas correm de acordo com a ordem estabelecida. Os *jobs* têm uma opção de agendamento que permite que um determinado *job* seja executado sem que o utilizador tenha que iniciar o processo.

As principais caraterísticas do PDI são apresentadas de seguida:

- Inclui um conjunto de componentes para realizar ETL. Um dos seus objetivos é que o processo de ETL se torne fácil de gerar, de manter e de implementar;
- Exploração do repositório de registos (tabelas, vistas) e metadados;
- Assistente para a criação de ligações a BDs;
- Suporta muitos tipos de transformações básicas: mapeamento de campos e valores, filtros de linhas, ordenação, sequências, divisão de campos, agrupamento, adição de constantes, normalização/desnormalização de linhas, uniões (*join*) de linhas, fusão de linhas e algumas operações matemáticas;
- Execução de procedimentos armazenados e SQL;
- As transformações podem ser chamadas por *jobs*, e os *jobs* por outros *jobs*, existindo mecanismos de passagem de informação entre transformações e *jobs*;
- Execução de *shellscripts* e comprovação da existência de ficheiros e tabelas;
- Definição do intervalo de execução num calendário de *jobs*;
- Ambiente gráfico de desenvolvimento;
- Código: aplicação 100% java com transformações. Desenho orientado a metadados;
- Conetividade: suporta Oracle, DB2, SQL Server e Sybase. Compatível também com MySQL, PostGres, Hypersonic, FireBird SQL e Ingres;
- O desenho da interface pode ser considerado um pouco pobre não havendo uma interface unificada em todos os componentes;
- Baseado em 2 tipos de objetos: Transformações e *Jobs*;

- Inclui 4 ferramentas complementares: Spoon (para desenhar transformações ETL usando o ambiente gráfico), PAN (para executar as transformações desenhadas no Spoon), CHEF (para criar *jobs*) e Kitchen (para executar *jobs*).

3.3.5 Community Dashboard Edition (CDE)

Dispõe de um ambiente gráfico, fornecendo aos utilizadores a informação crítica necessária para compreender e aprimorar o desempenho organizacional. É possível integrar completamente o módulo CDE com o PRD e o Pentaho Analysis. Para gerar gráficos o módulo apoia-se numa bilbioteca JFreeChart que contém gráficos mais comuns (2D, 3D, barras, linhas, séries temporais,...), possibilita o acesso a diferentes fontes de dados e exportação para PNG, JPEG e PDF.

3.3.6 Pentaho Data Mining (Weka)

Para a realização de DM, o Pentaho integrou no seu produto a plataforma *open source* independente desenvolvida em Java na Universidade de Waikato na Nova Zelândia, denominada de Weka. Esta ferramenta importa ficheiros de dados do tipo ARFF.

Contém uma interface gráfica para pré-processamento dos dados, classificação, regressão, segmentação, regras de associação e visualização dos dados de acordo com a seleção de determinados atributos. Possibilita o melhoramento do desempenho futuro do negócio através de análises preditivas, dispondo do conhecimento de padrões e relações ocultas nos dados e indicadores de desempenho futuro.

3.3.7 Pentaho Design Studio

Este módulo baseia-se no ambiente de programação Eclipse. A sua integração tem como principal intuito a criação e manutenção das sequências de ação (*xaction*). Estas sequências são a base das soluções Pentaho onde se encontram todos os componentes.

3.4 Usabilidade

O termo usabilidade está diretamente relacionado com o modo e a facilidade com que o utilizador lida com a plataforma sem formação prévia. Esta deve sempre ser considerada no desenvolvimento de um sistema, para além dos aspetos funcionais. A usabilidade é definida como a forma com que o utilizador e o sistema conseguem "comunicar" claramente. É também considerada o grau de compatibilidade do sistema com as capacidades cognitivas do utilizador para se estabelecer a comunicação, entendimento e resolução dos problemas. Por outro lado, a usabilidade é uma medida de qualidade da experiência do utilizador com a interação com uma aplicação [47, 7].

Segundo a norma ISO 9421-11, usabilidade é "a eficácia, eficiência e satisfação com que os utilizadores alcançam os seus objetivos em ambientes particulares de utilização". Assim, podem-se considerar como principais atributos da usabilidade a eficácia, eficiência e satisfação do utilizador (Figura 3.2). A eficácia representa a precisão e exatidão com que os utilizadores atingem os seus objetivos finais. Está diretamente relacionada com a funcionalidade da plataforma. A eficiência representa os recursos utilizados para

o alcance dos objetivos. Os utilizadores consideram um sistema eficiente quando este lhes permite atingirem os objetivos de uma forma rápida, simples e sem muito esforço a nível cognitivo. Finalmente, a satisfação é definida como o conforto e a aceitação por parte dos utilizadores. A plataforma deve dispor de um conjunto variado de opções e garantir confiança, segurança e privacidade dos serviços [48, 4, 49].

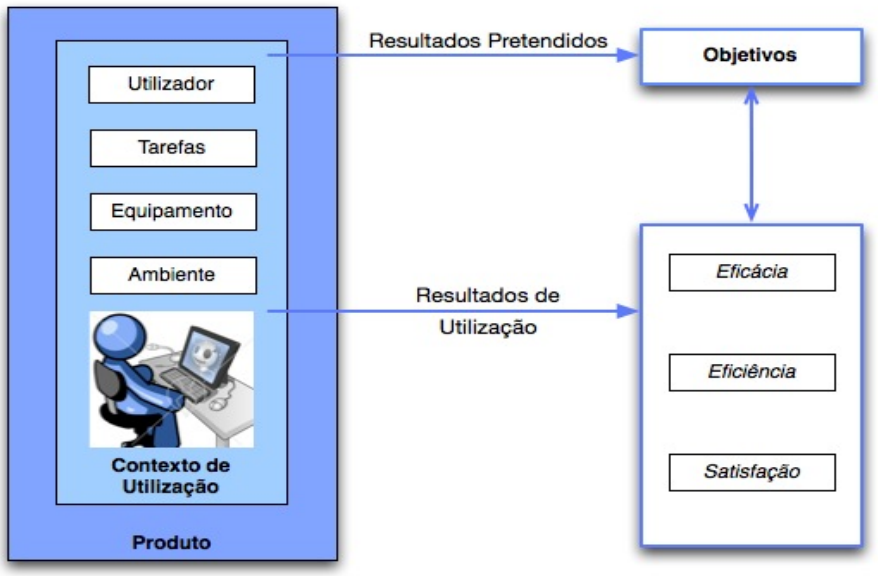


Figura 3.2: Ilustração da usabilidade contextual. Adaptado de [4]

Uma parte das plataformas *Web* já disponibilizam, atualmente, um campo de avaliação sistemática da satisfação do utilizador quando interage com o sistema [47].

No contexto da saúde, os erros médicos têm ganho cada vez mais espaço na área de investigação de informática médica. Considera-se que a usabilidade de uma plataforma tem uma relação direta com o aumento do número de erros que se têm vindo a registar [50].

Os princípios da usabilidade são [47, 51, 4, 49]:

- **Suporte:** capacidade de resposta, documentação sobre utilização, possibilidade de *feedback* FAQ, explicação dos procedimentos, suporte técnico, customização e operabilidade do sistema;
- **Consistência:** qualidade da linguagem, qualidade de representação, estrutura lógica, taxa reduzida de erros;
- **Relevância do contexto:** relevância precisa dos conteúdos, precisão da informação, tempo de realização de tarefas, disponibilidade de informação, imparcialidade da informação, terminologia, reconhecimento do contexto e ajuste automático do sistema ao mesmo;
- **Credibilidade:** segurança, privacidade, confiabilidade, garantia;
- **Legibilidade:** facilidade de leitura de dados, leitura estruturada de conteúdos, abrangência, introdução correta de dados;

- **Aprendizagem:** facilidade de aprendizagem e de utilização, aparência e apresentação, resultados cognitivos;
- **Simplicidade:** *layout* simples, simplicidade dos menus, tempo de resposta, tempo de *download*, velocidade de visualização dos dados;
- **Navegação:** facilidade em encontrar informação, procura múltipla e avançada, reversibilidade das ações;
- **Interatividade:** apresentação e aparência da plataforma, qualidade gráfica, tempo de resposta e aparência visual simples com *design* minimalista.

A usabilidade afeta as preferências do utilizador e a probabilidade dos mesmo regressarem à plataforma para a utilizarem novamente. Por isso, é importante que exista uma interação eficiente e bem sucedida entre o utilizador e o sistema [47].

Ao longo da apresentação e explicação dos casos de estudo realizados no contexto deste projeto vai sendo feita uma avaliação das ferramentas utilizadas e das soluções obtidas ao nível da usabilidade, apresentando-se no final um ponto conclusivo e geral a cerca da usabilidade do Pentaho.

Capítulo 4

Resultados e Discussão

Neste capítulo apresentam-se os resultados dos diversos casos de estudo realizados com o intuito de se explorarem as variadas ferramentas e módulos do Pentaho Suite. Foram testadas diversas soluções, apresentando-se os passos mais importantes que foram seguidos e referindo-se também outras funcionalidades das ferramentas, erros e falhas detetadas, e uma análise e avaliação em termos de usabilidade. Algumas soluções apresentam as informações em inglês uma vez que o seu desenvolvimento resultou numa publicação científica e, por isso, foram construídas nesse âmbito.

4.1 Dados Experimentais

Os dados utilizados nos casos experimentais são um subconjunto de alguns dados clínicos extraídos de uma BD de um hospital no norte de Portugal.

A manipulação da BD foi efetuada com o suporte da ferramenta Oracle SQL Developer, ambiente de desenvolvimento integrado aberto.

4.2 Listas de Espera do Centro Hospitalar

Nos sistemas de saúde de países mais desenvolvidos é frequente a espera para se ser submetido à prestação de cuidados de saúde, como uma consulta de especialidade ou uma intervenção cirúrgica. A existência de listas de espera nas instituições de saúde indicam uma incapacidade do hospital em responder a todos os pedidos, podendo também tratar-se de medidas administrativas para uma melhor gestão da organização.

Inicialmente, realizou-se uma exploração e análise dos dados das listas de espera para consulta e intervenção cirúrgica. As listas de espera são dinâmicas no sentido em que os registos são permanentemente atualizados podendo conter diversos estados e situações. No caso das listas de espera para cirurgia o registo em lista pode admitir 8 situações diferentes: agendado; cancelado; inscrito; operado; pré-inscrito; readmitido; transferido de; e transferido para; e 3 estados diferentes em lista: ativo, pendente ou fechado.

Nos casos de estudo realizados de forma a perfazer os objetivos deste trabalho de dissertação foram estabelecidos alguns parâmetros de forma a definir os registos que se encontram ativos em lista de espera num determinado momento uma vez que foram realizadas análises em que se marcou um ponto temporal e se consideraram apenas os

casos ativos em lista de espera. Assim, os casos de análise ativos em lista de espera para cirurgia seriam os que apresentassem uma data de marcação da cirurgia, que não apresentassem data de cancelamento ou operação (momento em que se finaliza o processo) e que também não apresentassem uma situação de "transferidos para". A tabela de lista de espera para cirurgia contém um total de 147878 registos, entre os anos de 2007 e 2012. No caso das listas de espera para consulta o registo em lista pode admitir 4 estados diferentes: agendada, pendente, marcada ou realizada. Assim, foram considerados registos ativos na lista num determinado instante de tempo aqueles que apresentassem uma data de receção de pedido para consulta de especialidade e que não existisse nesse mesmo registo um código de motivo de recusa (correspondente ao cancelamento da consulta e consequente saída da lista) ou um número de taxa (número atribuído no momento de marcação da consulta e por isso sai também da lista). A lista de espera para consultas é constituída por um total de 303820 registos, entre 2007 e 2012.

4.3 Monitorização das Listas de Espera para Cirurgia

4.3.1 Registo de Casos na Lista de Espera para Cirurgia

Numa primeira aproximação ao caso de estudo da lista de espera para cirurgias considerou-se importante organizar os dados de acordo com a especialidade médica cirúrgica e o respetivo total de casos de espera. Para tal, foram desenvolvidas soluções para analisar e visualizar a informação de forma simples e eficaz. Foi construída uma tabela e foram desenvolvidos *dashboards*, apresentando-se descritos de seguida. O desenvolvimento das soluções foi feito tanto no Pentaho CE como no Pentaho EE de forma a poder comparar as duas plataformas a vários níveis, como por exemplo os resultados obtidos e a usabilidade das soluções.

Numa fase inicial, procedeu-se à criação das ligações estabelecidas à BD. Neste sentido, a plataforma do Pentaho possui uma funcionalidade que possibilita a gestão das conexões, permitindo a criação de novas ligações, tal como se verifica na Figura 4.1. A fonte de dados pode ser um ficheiro do tipo CSV, através de *queries* SQL (utilizadas neste projeto) ou ainda de tabelas da BD. Posteriormente, pode-se definir e configurar o esquema do modelo de dados criado.

Os dados foram selecionados a partir da BD através de uma *query* SQL (transcrita abaixo), sendo que para a construção das soluções apenas se consideraram as 10 especialidades com maior número de registos em espera uma vez que simplificava a sua construção. Foram então utilizadas as funções SQL *between* (para selecionar o intervalo de anos que se pretende analisar) e o *rownum* (para selecionar apenas as 10 primeiras linhas da lista ordernada).

```
   Select * from(
select des_grupo, count(des_grupo) as Total
from bimov_listaesperablo
where to_char(dta_marcacao,'yyyy') between '2007' and '2012'
group by des_grupo
order by Total desc)
where rownum<=10
```

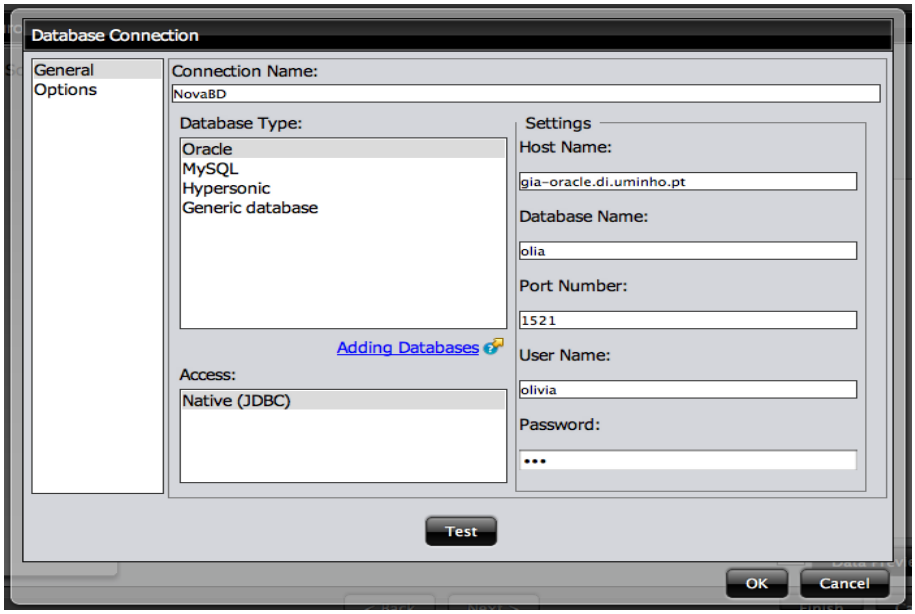


Figura 4.1: Gestão das conexões com a BD na plataforma.

Com o suporte do Pentaho EE foi construída uma tabela e o respetivo *dashboard* representado por um gráfico de barras verticais, utilizando o módulo de desenvolvimento de *dashboards*. A construção destes meios de visualização é bastante intuitiva e simples na EE uma vez que as dimensões são arrastadas ou selecionadas de forma interativa, sem existir uma definição de todos os pormenores de construção como acontece na CE. A construção das soluções baseia-se na filosofia de *drag-and-drop*, tornando a plataforma mais amigável ao utilizador. Na EE existe a possibilidade de selecionar o número de elementos do topo da lista que sejam apresentados na solução, enquanto que na CE tem que se definir na *query* SQL através da função *rownum*, tal como se apresenta transcrito na *query* acima. A tabela gerada através da plataforma EE pode ser visualizada na Figura 4.2.

DES_GRUPO	TOTAL (Sum)
OFTALMOLOGIA	27636
ORTOPEDIA	18156
CIRURGIA VASCULAR	13257
CIRURGIA AMBULATORIO	12578
UROLOGIA	8145
GINECOLOGIA MJD	8108
O.R.L	6647
ESTOMAT/CIR.MAX_FACIAL	6439
NEUROCIRURGIA	6141
DERMATOLOGIA	4800

Figura 4.2: Tabela das especialidades com os totais de espera para cirurgia (construído no Pentaho EE).

Daqui podem-se concluir as especialidades que têm maior registo de casos de espera ao longo dos 5 anos e, porventura, agir com medidas preventivas de forma a fazer uma gestão mais adequada destes casos.

Em relação ao gráfico de barras (Figura 4.3), a sua construção foi consideravelmente simples uma vez que após o esquema de dados estruturado, procedeu-se à seleção das dimensões e medidas a utilizar bem como o seu agrupamento e ordenação. De seguida, escolheu-se o tipo de gráfico, o tema de cores a utilizar, e a escolha das dimensões para as séries das colunas, a categoria das colunas, os valores das colunas e as escalas. Tais definições na CE são realizadas inteiramente de modo manual como se poderá verificar de seguida, o que torna o processo mais complicado e técnico mas também mais dependente do utilizador. Por outro lado, na EE pode-se legendar o *dashboard* e os eixos manualmente, bem como a rotação da legenda do eixo das abcissas. Quanto à legenda dos eixos este mecanismo facilita muito a construção do gráfico no sentido que na CE a legendagem está automaticamente dependente da *query* SQL, enquanto que a rotação dos valores do eixo das abcissas é vertical e caso se pretenda com outro ângulo como na diagonal esta definição tem que ser escrita por *Javascript*. Assim sendo, percebe-se claramente que a Pentaho EE está mais vocacionada para o mundo empresarial e para a utilização de pessoas informática e tecnicamente menos especializadas como gestores e administradores. Por sua vez, a Pentaho CE está mais vocacionada para o mundo dos técnicos de informática que desenvolvem as aplicações e, que por isso, têm conhecimentos mais elevados ao nível da programação.

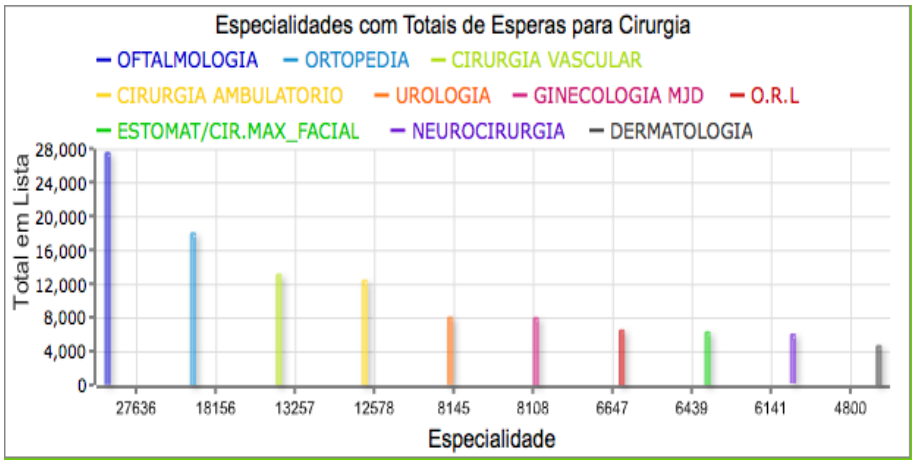


Figura 4.3: Gráfico de barras verticais das especialidades com os totais de espera para cirurgia (construído no Pentaho EE).

No Pentaho CE, a construção de *dashboards* divide-se em 3 etapas fundamentais. A primeira etapa consiste na escolha do *layout* do *dashboard*, onde podem ser definidos o *template* (pode ser aplicado um *template* pré-definido ou pode ser construído de raiz), os títulos principal e de cada secção, as cores de fundo, as dimensões de cada uma das divisões, o texto de rodapé bem como a inserção de imagens logo. A grande parte das definições de texto, cores e inserção de imagens são efetuadas através de HTML. A definição do *layout* baseia-se numa hierarquia de linhas e colunas (cada uma é uma *tag* html da camada do CDF (*dashboards framework*)). Nesta etapa pode também ser adicionado um recurso (CSS ou JS), definindo um esquema principal com as definições dos títulos, dimensões e organização. Numa solução desenvolvida foi testado um esquema principal através de CSS, sendo que a sua definição foi considerada relativamente simples.

Na etapa seguinte são definidos os componentes que se pretendem manipular e de-

senvolver. Apresenta-se na Figura 4.5 o *dashboard* desenvolvido no Pentaho CE, o qual é composto por um gráfico de barras verticais e um gráfico circular. No componente de um gráfico existem um número de opções que podem ser definidas para a sua construção. Cada componente gráfico tem que lhe ser atribuído um nome, um valor para a altura e largura, a fonte de dados a que está associado, o objeto HTML do *layout* onde se pretende que o gráfico seja inserido. Existe também a possibilidade de dar um título ao componente, definir se a legenda será ou não mostrada, se apresentará a opção de permitir de ser dinâmico no sentido em que um clique no gráfico mostra outros resultados ou se conterá parâmetros. Estas são as propriedades principais (Figura 4.4) de um componente gráfico, existindo ainda a opção de propriedades avançadas onde são especificadas propriedades como a opção de animação na visualização de um gráfico, opção de um gráfico empilhado, a mostragem ou não dos valores no próprios gráfico e a definição do seu formato, configurações mais específicas do título, legenda e eixos como o seu tamanho e alinhamento, entre muitas outras configurações.

A lista de componentes disponibilizada pelo Pentaho é impressionante, destacando-se os componentes CCC que são baseados no tipo Protovis. Podem-se criar gráficos de linhas, barras, circular, entre muitos outros, e diversos diagramas baseados em Protovis, OpenFlashChart e Rephael. São possíveis também outros componentes como elementos HTML tais como texto, tabelas, botões.

Figura 4.4: Propriedades principais de um componente gráfico do tipo barras.

A última etapa consiste na criação da fonte de dados baseada no CDA (Community Dashboard Access), onde se podem utilizar variados tipos de queries: SQL, MDX, MQL, XPATH, PDI, xaction, JS, ...

Para o desenvolvimento de *dashboards*, a plataforma Pentaho CE apresenta ainda inúmeras outras opções como: a criação de parâmetros (simples, costumizáveis e de data), componentes genéricas (integração com relatórios PRPT e análises Pivot, comentários, tabelas, menu de navegação, texto, botões, consultas, xactions...), construção de variados tipos de gráficos (componente CCC com diversos tipos de gráficos como linhas, de pontos, de área, componente Portovis, TimePlot, OpenFlashChart, gráfi-

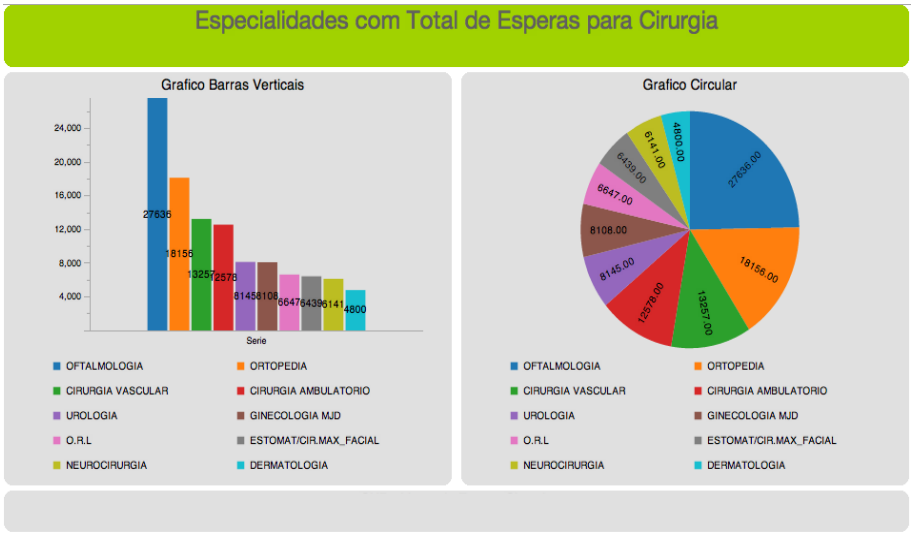


Figura 4.5: Dashboard das especialidades com os totais de espera para cirurgia (construído no Pentaho CE).

cos Dial, entre outros), componentes de seleção (entrada de texto, verificação, escolha de mês, completação automática, multi-seleção, entre outros,...), entre outras opções. Neste projeto de dissertação, os componentes utilizados foram os considerados adequados aos casos de estudo, porém outros tantas foram experimentadas para testar a sua funcionalidade, e uma parte foi apenas estudada a sua funcionalidade ao nível teórico.

Numa apreciação em termos de usabilidade das soluções obtidas e das edições utilizadas foram retiradas algumas conclusões. Em geral, os resultados obtidos pelo Pentaho EE têm um aspeto mais profissional do que o Pentaho CE. Porém, ambas as solução são consideradas atrativas e apelativas para visualização e análise de informação. No Pentaho CE é possível personalizar o conjunto de cores apresentado nos gráficos, sendo por isso a interatividade muito maior com o utilizador. Tanto a tabela como os *dashboards* produzidos apresentam simplicidade na leitura e perceção dos dados. Por outro lado, a dificuldade e o tempo de realização das tarefas é maior no Pentaho CE uma vez que é praticamente tudo definido e personalizado pelo utilizador. Em ambas as edições, a geração dos gráficos exige algum tempo de espera para visualização dos dados, porém esse tempo não é considerado prejudicial ao processo de tomada de decisão (são apenas alguns segundos). A realização de tarefas no Pentaho EE, ao contrário do CE, apresenta reversibilidade das ações. Ou seja, para além de ser possível voltar atrás um passo, quando o trabalho não está gravado e se pretende fechá-lo aparece uma mensagem para lembrar a gravação enquanto que no CE fecha automaticamente.

Em relação aos *dashboards* em geral, existe muito pouca documentação teórica de suporte à utilização desta funcionalidade. A aprendizagem tem que ser totalmente autónoma e por tentativa por parte do utilizador, existindo um número de fóruns de suporte que estão bem fundamentados no que diz respeito a erros de execução, *bugs* da ferramenta, impossibilidade de realizar determinadas tarefas. Basicamente, tem-se acesso ao que não dá ou não é possível, sendo que tudo o que é execuível tem que ser feito autonomamente e a aprendizagem feita de forma independente pois não existe muita documentação teórica.

Finalmente, foi construído um relatório onde estão apresentadas todas as especia-

lidades médicas e os respetivos totais de casos de espera. Este relatório encontra-se no anexo A.1 e foi desenvolvido através do *Interactive Report* da Pentaho EE. Este módulo é bastante simples para construção de relatório quando comparado com o *software* PRD, no sentido em que estando definido o esquema de dados para análise, a sua construção baseia-se no *drag-and-drop* das dimensões para os locais pretendidos do relatório, bem como as funções de paginação e data que estão definidas automaticamente.Mais à frente no contexto de outros casos de estudo serão apresentados outro tipo de relatórios desenvolvidos no PRD.

4.3.2 Registos Ativos em Lista de Espera para Cirurgia

Para a análise dos casos ativos em lista de espera para bloco construiu-se um DW com o suporte do PDI, apenas com os dados relevantes para o caso. O instante temporal marcado na lista de espera como de análise de registos ativos em lista naquele momento foi o dia 17 de Abril de 2012, num total de 7534 registos ativos para esse mesmo instante. Para realizar a análise foram desenvolvidos *dashboards* através do Pentaho CE e do Pentaho EE, em que se apresentou o tipo de gráfico de pontos. Neste caso de estudo para cirurgias apresenta-se um *dashboard* desenvolvido no Pentaho EE (Figura 4.6), enquanto que no caso análogo para consultas se apresenta o *dashboard* desenvolvido no Pentaho CE, permitindo assim fazer uma avaliação e comparação final das principais diferenças entre os dois, apresentada na secção 4.4.2.

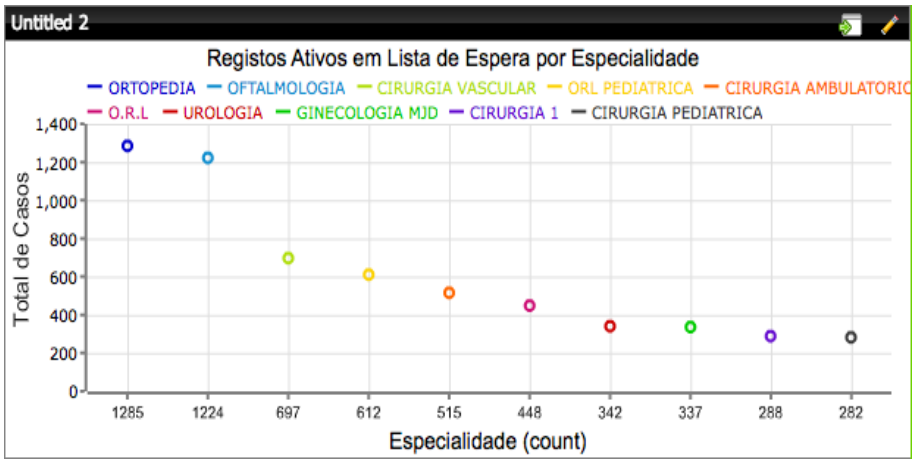


Figura 4.6: Gráfico de pontos desenvolvido no Pentaho EE, onde estão representadas as 10 especialidades com o maior número de registos ativos naquele momento.

Este tipo de gráfico no Pentaho EE é desenvolvido utilizando o tipo de gráfico área. Porém de acordo com o tipo de valores que são introduzidos ele forma um gráfico de pontos, pois não se verifica uma continuidade entre os valores. Da análise do gráfico conclui-se que no momento de análise as especialidades médicas mais representadas na lista de espera para cirurgia eram Ortopedia e Oftalmologia.

4.3.3 Cirurgias em Lista Espera (períodos mensais)

Neste caso em particular, foram considerados para estudo os tempos de espera para submissão a uma intervenção. O principal objetivo prendeu-se com a simulação

e monitorização da operação das listas de espera de forma a encontrar tendências e indicadores.

Primeiramente, foi realizada a análise do número de registos de cirurgias realizadas que se encontravam em lista de espera, correspondente à percentagem mais significativa de saídas da lista de espera (estado fechado). Para tal, usou-se o PDI como auxílio à construção do DW utilizando-se a função de junção (*merge*) de diferentes períodos ao longo de um mês (1-10, 11-20 e 21-31) para a transformação dos dados (Figura 4.7). Assim, dividiu-se um mês em 3 períodos diferentes: inicial, médio e final. Estes dados tiveram dois *outputs* distintos, tendo sido introduzidos novamente na BD numa nova tabela e também transcritos para um ficheiro de Excel.

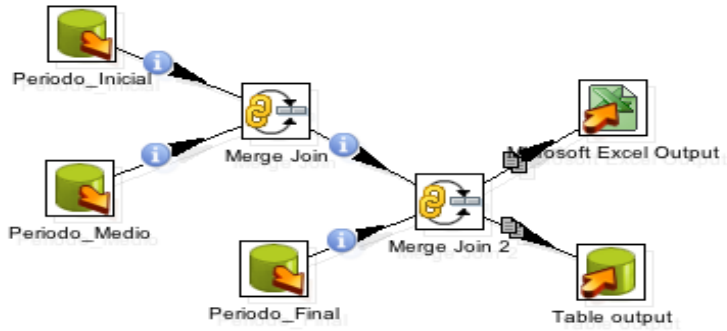


Figura 4.7: Projeto de construção do DW através do PDI.

A extração dos dados realizou-se através do *step Table Input* onde se definiu a BD e o esquema donde se pretendia extrair os dados. A seleção dos dados foi feita por uma *query* SQL, sendo que o PDI permite a pré-visualização da consulta requisitada por forma ao utilizador se certificar que está a extrair os dados corretos. Seguidamente, de forma a introduzir os dados todos juntos e organizados num DW, procedeu-se à função de junção. No PDI existe um conjunto de *steps* desenvolvidos especificamente para diferentes tipos de junção como o *Merge Join* (junta as entradas de acordo com o tipo de junção), *JoinRows* (possibilita a combinação das linhas de entrada através do produto cartesiano), *Merge Rows* (junta conjuntos de linhas de acordo com uma determinada chave), *Sorted Merge* (junção ordenada) e *XML Join* (junção de um conjunto de *tags* XML). Esta informação, bem como todos os *steps* existentes no PDI e respetiva função encontram-se na página *Web* que suporta a documentação do Pentaho[1]. Neste caso de estudo foi utilizado o *step Merge Join* em que se definiu o tipo de junção como INNER, isto é, recolha de dados dos diferentes períodos do mês em que coincidissem o mês e o ano. Existem ainda inúmeras possibilidades de entrada e saída de dados distintas, como por exemplo entradas de ficheiros CSV ou XML, mensagens de email e do tipo HL7, ficheiros de texto e Excel, análises OLAP,...; e saídas de ficheiros de diversos tipos, para tabelas de BD, para relatórios com integração do PRD, inserção ou atualização de dados,...

Em termos de usabilidade a ferramenta PDI apresenta vantagens e desvantagens para o utilizador. Considera-se que é uma ferramenta com alguma dificuldade de utilização, tendo o utilizador que ter uma ideia inicial dos dados que vai extrair inicialmente e daquilo que pretende obter como resultado final. Não é uma ferramenta simples e

[1] http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+Steps

exige conhecimentos teóricos prévios por parte do utilizador. Em termos de documentação o conjunto de *steps* que constituem o PDI encontram-se bem apresentados e descritos no tutorial da página Wiki do Pentaho já referido acima. Dependendo do número de dados que se estão a tratar pode demorar mais ou menos tempo de execução das transformações, porém considerou-se que a velocidade de processamento é aceitável para o problema. A interface gráfica de desenvolvimento é atrativa e tem qualidade, bem como todos os *steps* têm um ícone que os distingue entre eles tornando mais fácil a sua utilização. Em relação às soluções resultantes pode-se considerar que o *output* para ficheiros de texto ou Excel não é muito apelativa pois não há qualquer formatação, o que dificulta também a leitura dos dados. Por fim, os atalhos de teclado do PDI para um utilizador Mac não se encontram completamente adaptados ao sistema operativo. Uma ação de guardar ou refazer ou qualquer outra efetua-se através da tecla ctrl em vez da tecla comando.

De seguida, procedeu-se a técnicas de visualização através do módulo CDE do Pentaho CE para se analisarem os resultados obtidos. Na Figura 4.8, encontra-se representado o *dashboard* com o número registos que estiveram em lista de espera para cirurgia no passado ano 2011, e na Figura 4.9 está representado o número de registos em lista de espera para cirurgia ao longo de vários anos (2007-2011) num determinado mês, neste caso o mês 07 (correspondente a Julho). A escolha por estes ano e mês específicos em detrimento dos outros resultados foi completamente aleatória uma vez que todos os gráficos apresentavam um comportamento similar (restantes gráficos no anexo A.3). A linha azul significa o período inicial do mês, a linha laranja representa o período médio e a linha verde corresponde ao período final do mês. No desenvolvimento do *dashboard*, inseriu-se um parâmetro com um componente de seleção que através de consulta de dados na BD dispõe de todos os meses possíveis para análise, evitando assim a construção de *dashboards* independentes para cada um dos meses. Para a criação de um parâmetro deve-se definir uma *query* que consulte todos os dados que vão definir o parâmetro, depois cria-se um componente simples de parâmetro onde lhe é atribuído um nome e um valor por defeito. Posteriormente, define-se na *query* relativa ao componente gráfico o atributo que vai ser variável substituindo por ${nome_parâmetro}, define-se o tipo de componente para selecionar o parâmetro (de seleção, auto-completação, de verificação, de seleção múltipla, entre outros). Por fim, associa-se ao componente gráfico o parâmetro definido e o respetivo *listener*.

Em conclusão ao processo de EC realizado, foram sumariadas algumas tendências verificadas na Figura 4.8, e consideradas importantes. Foi observado um decréscimo do número de cirurgias alistadas ao longo do ano 2011. Por outro lado, foi facilmente verificado um decréscimo acentuado do número de registos na lista de espera durante o primeiro período de Janeiro, durante todo o mês de Agosto e no período final de Dezembro, tendo como possível explicação o fato destes períodos serem meses comuns para férias frequentes do pessoal.

Na Figura 4.9 foi observado um decréscimo acentuado do número de cirurgias alistadas entre os anos de 2010 e 2011. Este fato pode provir da implementação de novas medidas de gestão no hospital, as quais limitam o número máximo de dias nas listas de espera para cirurgia (12 meses). O módulo CDE permite saber o valor exato de cada um dos pontos do gráfico. Para tal, basta colocar o cursor do rato em cima de cada um dos pontos, tal como se pode verificar na Figura 4.9, onde se conclui que o número de cirurgias em espera para o período médio do mês de Julho para o ano de 2008 é de

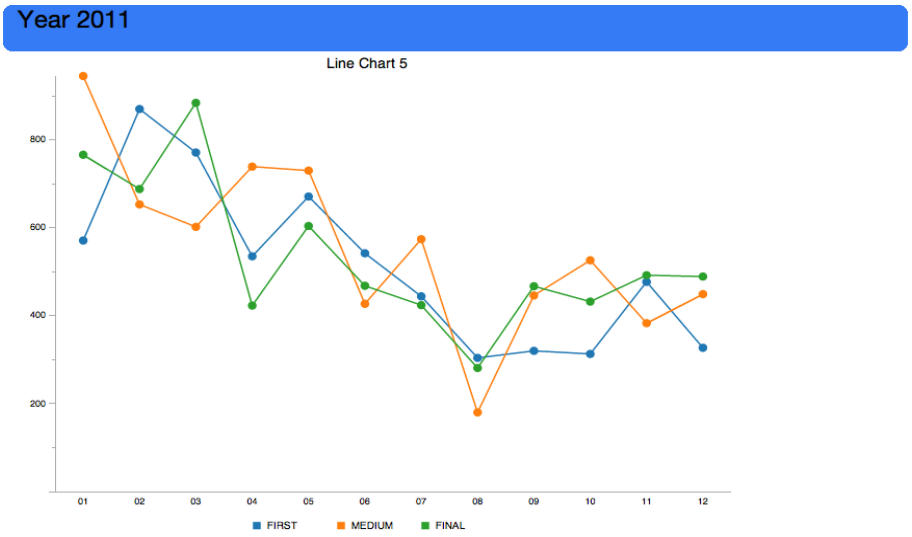


Figura 4.8: Número de cirurgias em lista de espera ao longo do ano 2011.

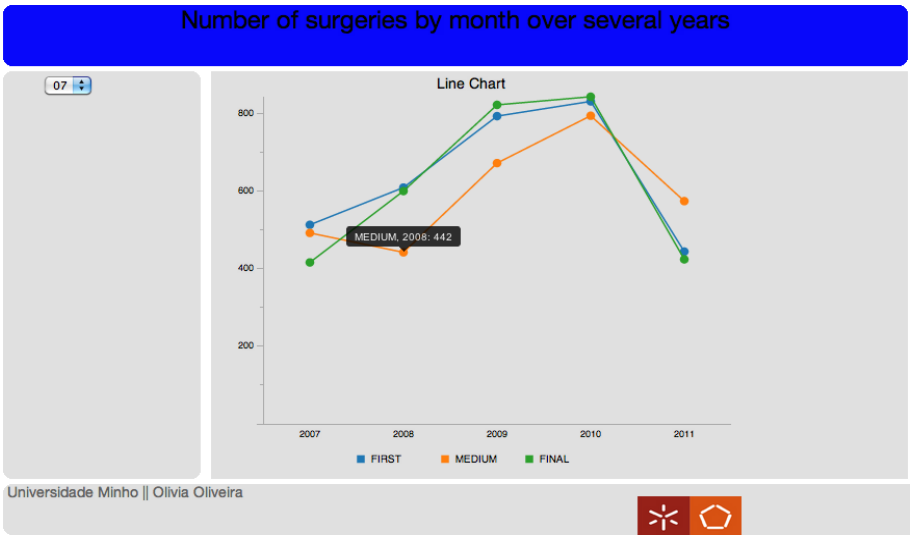


Figura 4.9: Número de cirurgias em lista de espera durante vários anos no mês de Julho.

442 cirurgias.

As soluções apresentadas neste caso de estudo com a utilização do módulo CDE do Pentaho CE são apelativas para o utilizador final, os gráficos utilizam cores atrativas e revelam facilidade na leitura dos dados. Em geral, o tempo de processamento e geração dos *dashboards* é aceitável, bem como o tempo em que se muda de parâmetro. Por outro lado, existe interatividade com o utilizador na apresentação dos resultados, pois este colocando o cursor do rato em cima dos pontos pode obter informação mais detalhada sobre aquele ponto. Em relação ao módulo CDE, é considerado de alguma dificuldade de utilização e está direcionado para utilizadores minimamente avançados que possuam algumas bases de desenvolvimento e até de alguma programação. Para além disso, existe pouca documentação de suporte à utilização da ferramenta como tutoriais. A grande parte dos componentes e funcionalidades encontra-se apresentada apenas em fóruns, sendo que algumas funcionalidades não existem mesmo referidas. Por outro

lado, o módulo CDE possui a grande desvantagem de não apresentar reversibilidade das ações. Em relação ao ambiente gráfico da interface pode-se afirmar que existe qualidade gráfica tanto na simplicidade e organização do *layout* e das diferentes secções, como nas soluções apresentadas.

4.3.4 Pentaho Google Maps (GeoMap)

A edição empresarial do Pentaho desenvolvida especialmente para apoiar o mundo empresarial na tomada de decisões e, por isso, com ferramentas muito mais específicas, possui um módulo de integração de análise e visualização de dados com o Google Maps, denominado de GeoMap, pertencente ao Pentaho Analyzer para realização de análises OLAP. Desta forma, é permitida a resolução de problemas com componentes e dados geográficos, uma vez que atualmente uma elevada percentagem de dados apresenta um componente espacial. A combinação da visualização gráfica dos resultados fornecida pelo Pentaho com a visualização da informação geográfica das respetivas medidas auxilia as empresas na identificação de padrões relevantes como por exemplo a associação do seu pacote de clientes/vendas/fornecedores por regiões, permitindo encontrar padrões de comportamento essenciais à tomada de decisão. Explorado no âmbito deste projeto, no contexto hospitalar, este módulo revelou-se igualmente útil no sentido em que foi possível organizar os pacientes em lista de espera para cirurgia e consulta de acordo com a sua origem. A informação geográfica foi organizada por distritos e, de modo pormenorizado, por concelhos do distrito do Porto. Ao contrário do que se possa pensar, a integração de informação geográfica em dados clínicos pode trazer importantes conclusões como por exemplo saber os locais em que as pessoas são mais sensíveis a chuvas intensas ou ondas de calor, previsão de doenças em determinadas áreas geográficas ou o número de pessoas carenciadas que vivem sozinhas em pontos geográficos.

Lista de Espera para Bloco por Distritos

Um dos resultados pretendidos foi a análise da naturalidade dos pacientes em lista de espera para cirurgia. Desta forma, utilizou-se a informação geográfica dos mesmos organizando-a por distritos com o suporte do módulo GeoMap e cruzando-a com os casos em lista de espera. Por outro lado, através do Pentaho Analyzer foi possível a análise completa da informação, desde a visualização geográfica à construção de tabelas e gráficos de forma a constituir conhecimento útil para auxiliar a tomada de decisão e a implementação de medidas (Figura 4.10).

O passo inicial consistiu na criação de um esquema do modelo da fonte de dados com todos os dados relevantes para a realização da análise. Estes foram selecionados através da conexão à BD com a tabela referente às listas de espera e usando *queries* SQL. Desta forma, criou-se então o esquema e o cubo para análise dos dados. Seguidamente, foram definidas as dimensões e as medidas, sendo que neste módulo existe então a possibilidade de se atribuírem as dimensões geográficas (país, cidade, código postal,...) às dimensões escolhidas. A seleção das medidas e dimensões realiza-se através de *drag-and-drop*, o que torna o processo de análise mais simples e amigável, sendo que é possível depois definir funções de cálculo, ordenação e agrupamento clicando nas próprias medidas. Os distritos corresponderam às principais cidades portuguesas e as respetivas coordenadas geográficas foram introduzidas num ficheiro .csv, que é permanentemente importado pelo servidor do Pentaho com a informação geográfica

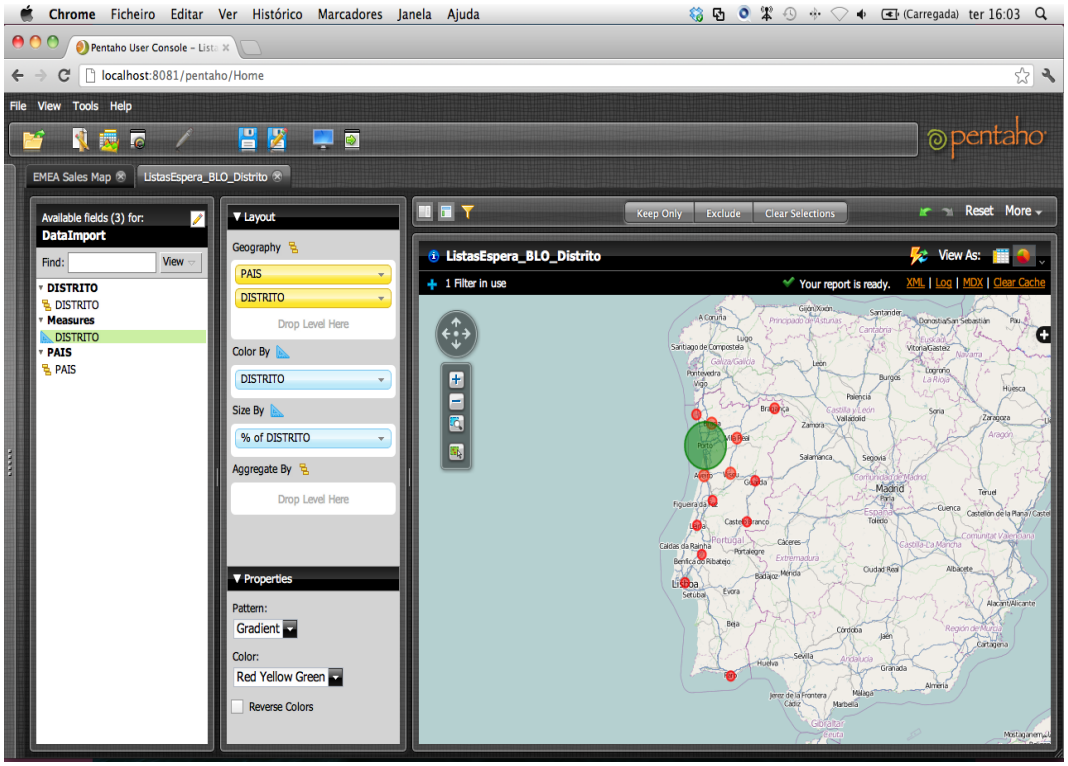


Figura 4.10: Módulo Pentaho Analyzer da edição empresarial; visualização da informação geográfica dos pacientes em lista de espera para bloco utilizando o módulo GeoMap.

relativa a diferentes países, estados, cidades. Dado isto, a informação encontrava-se estruturada e foi organizada numa tabela (Figura 4.11), através também da filosofia de *drag-and-drop* para constituir as diferentes colunas.

Em seguida, procedeu-se à integração da informação com o GeoMap, disponibilizando-a de acordo com os seus dados geográficos, neste caso pelos diferentes distritos de Portugal (Figura 4.10). Finalmente, procedeu-se à construção de diversos gráficos de forma a permitir uma estruturação e visualização gráfica da informação, sendo que neste caso de estudo apenas se apresentará o gráfico de barras horizontais (Figura4.12), apresentando-se os restantes gráficos no anexo A.4.1.

Da análise da informação gerada e disponibilizada, foi possível concluir que tal como já era previamente previsível o distrito do Porto era o mais representado na lista de espera (aproximadamente 81% dos casos). Por outro lado, os distritos adjacentes ao Porto (como Braga, Aveiro, Viseu, Vila Real) apresentavam números representativos de pacientes em lista de espera para cirurgia. Foi importante a tomada de conhecimento na medida em que se torna relevante para o hospital saber a naturalidade e origem da grande maioria dos pacientes. No âmbito da tomada de decisão e implementação de medidas poderia tornar-se pertinente para o centro hospitalar disponibilizar aos pacientes um transporte para os distritos com um número significativo de pacientes. Outro dos aspetos relevantes é o conhecimento por parte do hospital nos padrões de comportamento dos pacientes, dependendo da origem destes.

As soluções resultantes da utilização do módulo de análise do Pentaho EE (Analyzer) são atrativas e interativas. Baseiam-se na filosofia *drag-and-drop* tornando a análise

PAIS	DISTRITO	DISTRITO	% of DISTRITO
PORTUGAL	PORTO	5.252	81,28%
	AVEIRO	289	4,47%
	BRAGA	270	4,18%
	VISEU	257	3,98%
	VILA REAL	196	3,03%
	BRAGANCA	101	1,56%
	VIANA DO CASTELO	77	1,19%
	GUARDA	10	0,15%
	LISBOA	3	0,05%
	LEIRIA	2	0,03%
	SANTAREM	2	0,03%
	CASTELO BRANCO	1	0,02%
	COIMBRA	1	0,02%
	FARO	1	0,02%

Figura 4.11: Números de pacientes em lista de espera organizado por distritos.

mais dinâmica e amigável para o utilizador. Após definidas as dimensões e medidas, os gráficos são gerados praticamente de forma automática e com uma facilidade de utilização e interação muito maior. A qualidade gráfica tanto da interface como das soluções é bastante superior comparada com a do Pentaho Analysis (CE) como se poderá verificar na secção 4.3.5. Por outro lado, a integração do Google Maps originando o módulo GeoMap introduziu um maior grau de usabilidade à plataforma EE, no sentido em que a informação geográfica é agora apresentada de forma muito mais apelativa ao utilizador. O tempo de execução em geral no Pentaho Analyzer é bastante baixo, sendo que as soluções são apresentadas quase instantaneamente, com exceção do módulo GeoMap que demora algum tempo na geração dos mapas.

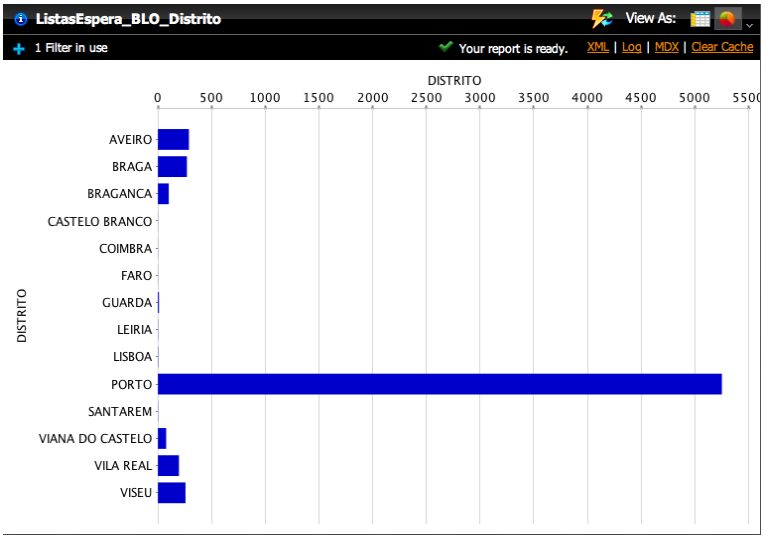


Figura 4.12: Gráfico de barras horizontais com o número de pacientes em lista de espera para cirurgia por distrito de origem.

Lista de Espera para Bloco por concelhos do distrito do Porto

Posteriormente, os dados geográficos dos pacientes foram organizados de acordo com o seu concelho de origem. Nesta caso, foi considerado apenas o distrito do Porto, uma vez que se considerou o mais importante para a análise no contexto do hospital em estudo. Novamente, através do módulo GeoMap a informação foi disponibilizada segundo as suas dimensões geográficas (Figura 4.13).

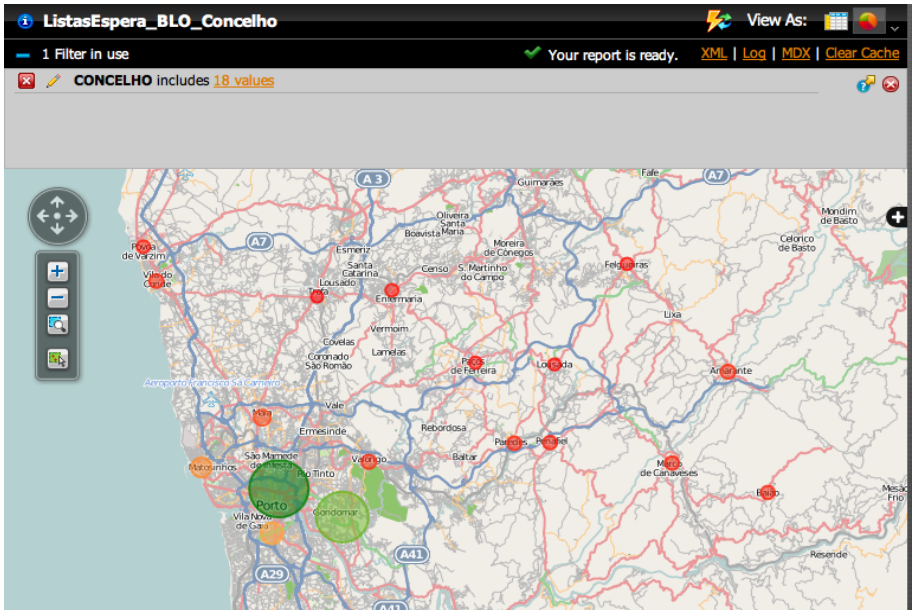


Figura 4.13: Visualização dos dados geográficos dos pacientes em lista de espera para bloco utilizando o módulo GeoMap.

A informação foi estruturada e tabelada por concelhos (Figura 4.14), sendo que anteriormente se teve que proceder à aplicação de um filtro no sentido de canalizar apenas os resultados referentes ao distrito do Porto.

PAIS	CONCELHO	CONCELHO	% of CONCELHO
PORTUGAL	AMARANTE	109	2,08%
	BAIAO	80	1,52%
	FELGUEIRAS	33	0,63%
	GONDOMAR	1.601	30,50%
	LOUSADA	20	0,38%
	MAIA	189	3,60%
	MARCO DE CANAVESES	81	1,54%
	MATOSINHOS	336	6,40%
	PACOS DE FERREIRA	31	0,59%
	PAREDES	91	1,73%
	PENAFIEL	84	1,60%
	PORTO	1.870	35,62%
	POVOA DE VARZIM	57	1,09%
	SANTO TIRSO	47	0,90%
	TROFA	6	0,11%
	VALONGO	110	2,10%
	VILA DO CONDE	94	1,79%
	VILA NOVA DE GAIA	411	7,83%

Figura 4.14: Números de pacientes em lista de espera para cirurgia organizado por concelhos.

Nas Figuras 4.15 e 4.16 encontram-se algumas das representações gráficas concretizadas com o intuito de se poder adquirir uma melhor compreensão da informação, sendo que as restantes representações se encontram no anexo A.4.2.

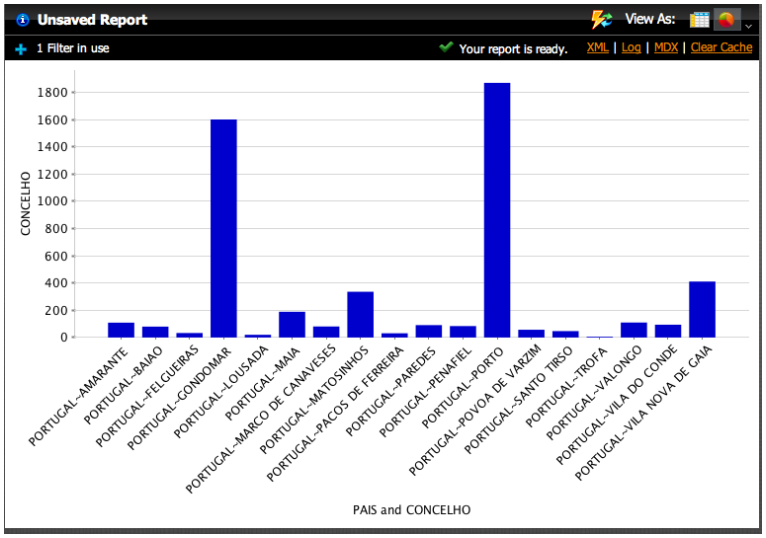


Figura 4.15: Gráfico de barras verticais com o número de pacientes em lista de espera para cirurgia por concelhos do distrito do Porto.

Em jeito de conclusão, a informação obtida corresponde às expectativas e previsões iniciais, isto é, os concelhos mais representados por pacientes em lista de espera são o concelho do Porto (cerca de 36% dos casos) e os concelhos vizinhos (tal como Gondomar, Vila Nova de Gaia, Matosinhos e Maia). Esta ferramenta permite, também, a reversibilidade das ações e a exportação das análises realizadas para PDF, Excel e CSV, bem como a criação de relatórios (no anexo A.4.2). Contudo, não é possível a

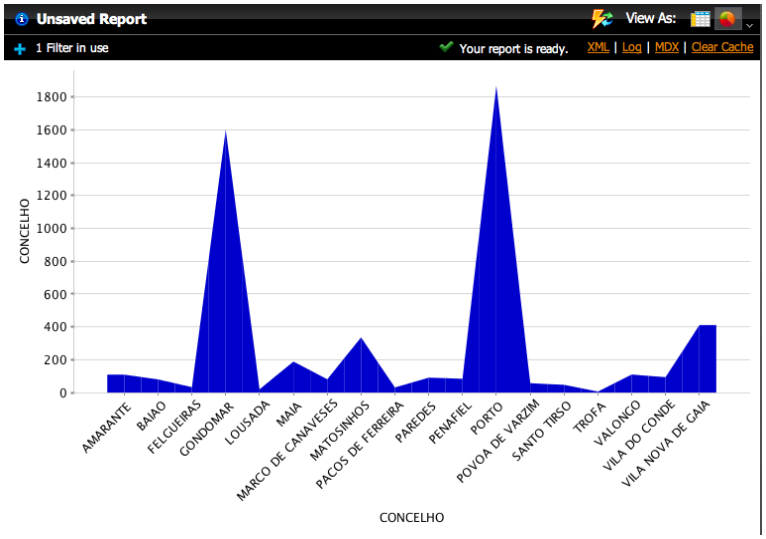


Figura 4.16: Gráfico de área com o número de pacientes em lista de espera para cirurgia por concelhos do distrito do Porto.

exportação dos mapas, apenas a tabela de análise, os gráficos construídos e as informações base relativas à análise. Finalmente, não existe praticamente documentação teórica que sirva de guia à utilização da ferramenta, porém considera-se a sua utilização razoavelmente simples e até de algum modo muito mais intuitiva do que o Pentaho Analysis.

4.3.5 Lista de Espera para Bloco por concelhos do distrito do Porto - Análise OLAP no Pentaho CE

Posteriormente, e baseando-se no mesmo caso de estudo foi realizada uma análise OLAP dos dados utilizados para explorar os concelhos do distrito do Porto de onde são originários os pacientes em lista de espera para cirurgia, utilizando-se o módulo Pentaho Analysis pertencente ao Pentaho CE. De início, foi utilizado o Navegador OLAP (Figura 4.17), através do qual é possível definir-se o cubo multidimensional a utilizar para a análise. Aqui define-se o *layout* global da consulta realizada por SQL, como quais as dimensões a utilizar como linhas, as que se utilizarão como colunas e quais os membros em que serão aplicados filtros.

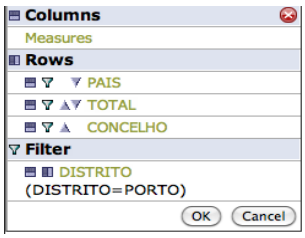


Figura 4.17: Navegador OLAP para a definição e estruturação do cubo multidimensional.

Consequentemente, é construída uma tabela de forma automática onde está organizada a informação definida no cubo, e através da qual se pode manipular a informação

dinamicamente (Figura 4.18).

			Measures
PAIS	TOTAL	CONCELHO	• TOTAL
⊞ All PAISs	⊞ All TOTALs	⊟ All CONCELHOs	5250
		AMARANTE	109
		BAIAO	80
		FELGUEIRAS	33
		GONDOMAR	1601
		LOUSADA	20
		MAIA	189
		MARCO DE CANAVESES	81
		MATOSINHOS	336
		PACOS DE FERREIRA	31
		PAREDES	91
		PENAFIEL	84
		PORTO	1870
		POVOA DE VARZIM	57
		SANTO TIRSO	47
		TROFA	6
		VALONGO	110
		VILA DO CONDE	94
		VILA NOVA DE GAIA	411

Slicer: [DISTRITO=PORTO]

Figura 4.18: Tabela gerada de forma automática a partir da definição do cubo OLAP.

Para a construção deste caso de estudo foi aplicado um filtro, aqui também designado de *slicer*, que permite filtrar os dados baseados num membro de uma dimensão oculta nas linhas ou colunas da visão global do esquema. Assim, foram filtrados apenas os concelhos pertencentes ao distrito do Porto (Distrito=Porto).

O Pentaho Analysis utiliza a linguagem de consulta MDX para definir consulta de dados multidimensionais. Esta linguagem MDX é uma linguagem muito mais detalhada e expressiva, e a sua utilização para o caso de estudo em questão apresenta-se na Figura 4.19. No caso de existir a necessidade de modificar a informação extraída para análise, o Pentaho Analysis possibilita a edição da *query* MDX através do MDX Query Editor.

MDX Query Editor

```
select NON EMPTY {[Measures].[TOTAL]} ON COLUMNS,
  NON EMPTY Crossjoin(Hierarchize({([PAIS].[All PAISs], [TOTAL].[All TOTALs])}), {[CONCELHO].
[All CONCELHOs], [CONCELHO].[AMARANTE], [CONCELHO].[BAIAO], [CONCELHO].[FELGUEIRAS],
[CONCELHO].[GONDOMAR], [CONCELHO].[LOUSADA], [CONCELHO].[MAIA], [CONCELHO].[MARCO
DE CANAVESES], [CONCELHO].[MATOSINHOS], [CONCELHO].[PACOS DE FERREIRA], [CONCELHO].
[PAREDES], [CONCELHO].[PENAFIEL], [CONCELHO].[PORTO], [CONCELHO].[POVOA DE VARZIM],
[CONCELHO].[SANTO TIRSO], [CONCELHO].[TROFA], [CONCELHO].[VALONGO], [CONCELHO].[VILA
DO CONDE], [CONCELHO].[VILA NOVA DE GAIA]}) ON ROWS
from [DataImport4]
where {[DISTRITO].[PORTO]}
```

Figura 4.19: Visualização da query MDX utilizada para o caso de estudo.

A partir da query definida para a consulta dos dados, o Pentaho Analysis também permite a construção de gráficos, tal como se pode observar na Figura 4.20.

Neste módulo, existe um botão que permite aceder às propriedades do gráfico (Figura 4.21) onde é possibilitado ao desenvolvedor alterar as definições gerais dos gráficos.

Para além dos diversos tipos de definições que se podem proceder, é possível também a escolha do tipo de gráfico a utilizar, como a formatação do gráfico, mostrar/ocultar a legenda,... Existem inúmeros tipos de gráfico possíveis: barras vertical ou horizontal, barras empilhadas vertical ou horizontal, linear vertical ou horizontal, de área vertical ou horizontal, de área empilhado vertical ou horizontal e circular, podendo ainda definir

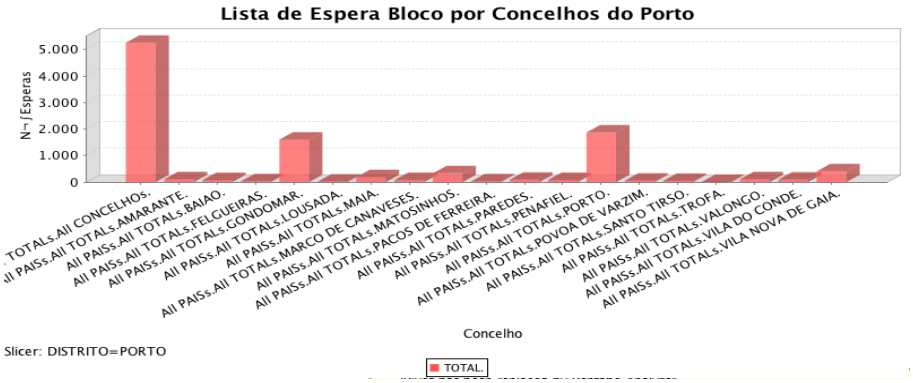


Figura 4.20: Visualização do gráfico criado pelo Pentaho Analysis.

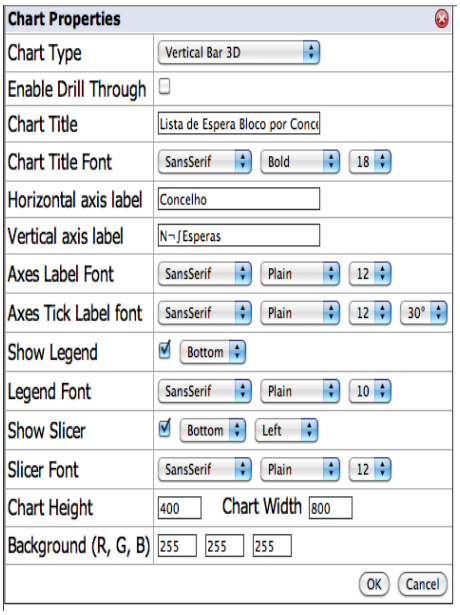


Figura 4.21: Janela de definição das propriedades dos gráficos.

em alguns destes tipos a opção de 2D ou 3D. Para este caso de estudo foi utilizado um gráfico de barras vertical 3D.

O Pentaho Analysis possui uma função com significativo potencial designada de *Swap Axes* (Figura 4.22), que fornece ao utilizador a possibilidade de articulação dos dados permutando as dimensões/membros da área das linhas com as dimensões/membros da área das colunas.

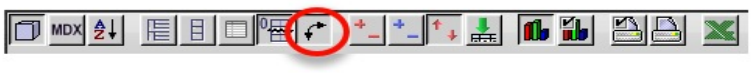


Figura 4.22: Barra de ferramentas do Pentaho Analysis com o botão de Swap Axes realçado.

Em consequência da utilização desta função as dimensões/membros são então permutados e, por isso, o gráfico também se altera tendo como resultado final a Figura 4.23.

Por outro lado, o Pentaho Analysis possui também um conjunto de botões (Figura

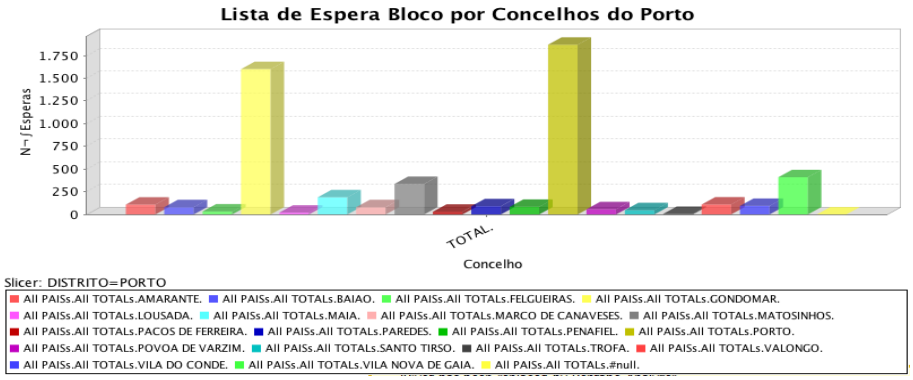


Figura 4.23: Visualização do gráfico resultante da utilização da função Swap Axes.

4.24) baseados na função designada de *Drilling* que permite aos utilizadores expandir as hierarquias dentro de uma determinada dimensão e desta forma descobrir detalhes adicionais.

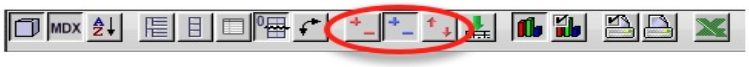


Figura 4.24: Barra de ferramentas do Pentaho Analysis com o conjunto de butões Drill realçados.

Dentro do mesmo âmbito existe também a função de *Drill Through* (Figura 4.25) que através dos indicadores da grelha de dados permite aos utilizadores clicar e ver os detalhes individuais (fatos) resultantes da agregação do valor da célula.

Figura 4.25: Barra de ferramentas do Pentaho Analysis com o botão de Drill Through realçado.

Em resultado da utilização da função de *Drill Through* foi possível a visualização detalhada da informação do concelho do Porto (Figura 4.26) e da Trofa (Figura 4.27), com o total de pacientes em lista de espera para bloco detalhado, bem como o distrito e país a que pertencem.

Desta análise as conclusões retiradas para o caso de estudo são as mesmas da análise realizada através do módulo GeoMap. Concluindo, o módulo Pentaho Analysis não é um módulo tão intuitivo quanto o Pentaho Analyzer. Porém, para a realização de análises OLAP e exploração dos respetivos módulos do Pentaho é necessário um conjunto de conceitos teóricos bastante consolidados. A realização de uma análise OLAP pode ser vantajosa no sentido em que o uso de gráficos e cubos OLAP produz um maior dinamismo e interatividade com a manipulação dos dados.

Analisando a realização do mesmo caso de estudo em módulos diferentes, considerou-se que a utilização do módulo Pentaho Analyzer (integração do GeoMap) é mais simples, mas verifica-se que o Pentaho Analysis fornece uma visualização mais detalhada da informação e permite uma manipulação mais dinâmica dos dados e mais dependente do utilizador. Considera-se que para um utilizador mais experiente pode-se

Drill Through Table for TOTAL_0

CONCELHO	DISTRITO	PAIS	TOTAL	TOTAL_0
PORTO	PORTO	PORTUGAL	1.870,00	1.870,00

Figura 4.26: Visualização detalhada dos dados relativos ao concelho do Porto através da utilização da função de *Drill Through*.

Drill Through Table for TOTAL_0

CONCELHO	DISTRITO	PAIS	TOTAL	TOTAL_0
TROFA	PORTO	PORTUGAL	6,00	6,00

Figura 4.27: Visualização detalhada dos dados relativos ao concelho da Trofa através da utilização da função de *Drill Through*.

tornar mais vantajosa a realização de uma análise OLAP com o suporte do Pentaho CE, enquanto que um utilizador básico e iniciante consegue mais facilmente efetuar uma análise OLAP no Pentaho EE. Através destas ferramentas, é possível permutar os eixos dos dados, ordenar os dados de forma ascendente e descendente, suprimir as colunas e linhas que estão vazias, visualizar gráficos de diversos tipos, utilizar a opção de mostrar dados em detalhe e exportar a análise realizada para Excel e PDF para a criação de relatórios.

As soluções apresentadas pelo Pentaho Analyzer são extremamente mais atrativas do que o Pentaho Analysis e, tal como já referido acima, este módulo não é muito intuitivo e requer algum conhecimento prévio e bem fundamentado por parte do utilizador. Porém, uma das grandes vantagens do Pentaho Analysis é que o utilizador possui uma maior participação e responsabilidade das soluções finais, estando estas mais dependentes das ações do utilizador uma vez que a ferramenta não é tão automática. A ferramenta possibilita a exportação da análise para PDF ou Excel mas não se conseguiu realizar essa ação nem encontrar o erro através do qual a ação está impossibilitada. O tempo de realização da análise é relativamente baixo, não possui reversibilidade das ações, nem integração com o Google Maps. Finalmente, verificou-se que a documentação de Analysis é escassa, não existindo muita documentação com descrições técnicas, apenas com a descrição da interface[2].

4.3.6 Listas de Espera para Bloco - Análise DM

Neste caso de estudo, apenas foram considerados os algoritmos de *clustering* uma vez que o objetivo é a descoberta de grupos de dados similares entre si, e também como já foi verificado em teoria a técnica de *clustering* é largamente aplicada em dados clínicos. Para análise, foi utilizada a tabela da lista de espera para cirurgia, tendo sido considerados os atributos da descrição do grupo (des_grupo), do código da patologia (cod_patologia), da prioridade do caso (prioridade), do tipo de cirurgia (tipo_cirurgia) e o do distrito do paciente. Todos os outros atributos foram removidos no Weka, na fase de pré-processamento, através de um filtro não supervisionado.

O algoritmo EM atribui a distribuição da probabilidade para cada instância no qual indica o valor de probabilidade para cada uma pertencer aos *clusters* encontrados.

[2] http://www.osbi.fr/wp-content/Pentaho-Analysis-Viewer-User-Guide.pdf

Assim, cada instância pertence a cada *cluster* probabilisticamente. Este algoritmo possibilita a decisão do número de *clusters* criados por validação cruzada, podendo também ser definido *apriori* o número de *clusters* que vão ser gerados. Da aplicação do algoritmo EM resultaram 3 *clusters* sendo que o mais utilizado agrupou 404 instâncias (73%). As caraterísticas principais deste grupo são: grupo clínico de Ortopedia, códigos de patologia 71515 (Osteoartrose Localizada Primária na Região Pélvica e Coxa) e 7350 (Hallux Valgo Adquirido, nome científico para Joanete), prioridade de cirurgia 1 (nível mais baixo), tipo de cirurgia 1 (primária) e distrito do Porto (resultados completos no anexo A.5).

O algoritmo SimpleKMeans agrupa dados utilizando o algoritmo K-Means. Os dados foram agrupados de acordo com 2 *clusters*, sendo que o principal com maior número de instâncias agrupadas (94%) apresenta as seguintes caraterísticas: grupo de Ortopedia, código de patologia 71515 (Osteoartrose Localizada Primária na Região Pélvica e Coxa), prioridade 1, tipo de cirurgia primária e distrito do Porto. Os resultados da computação deste algoritmo corroboram com os obtidos através do algoritmo EM.

O algoritmo Farthest First é uma variante do K-Means que coloca cada centróide no ponto mais distante dos centróides já existentes, e escolhe aleatoriamente um ponto como o primeiro centro. Na maioria dos casos, o FarthestFirst acelera o processo de segmentação, uma vez que é necessário um menor ajuste e processamento dos dados. Finalmente, utilizou-se o algoritmo FarthestFirst para a computação de somente 4 atributos (grupo, prioridade, tipo de cirurgia e distrito), onde foi definido um *cluster* central (centróide) com 504 instâncias (91%) e com as caraterísticas: grupo de Ortopedia, prioridade 1, cirurgia primária e distrito do Porto. O *cluster* mais distante contém 48 instâncias (9%) e apresenta as caraterísticas: grupo de Oftalmologia, prioridade 1, cirurgia secundária e distrito de Viseu.

Com a análise de DM foi possível retirarem-se algumas conclusões como por exemplo: o grupo clínico com mais casos cirúrgicos no hospital em estudo é Ortopedia, a patologia mais frequente é a de código 71515, o distrito mais representado é o Porto, e um grupo pequeno que agrupa cirurgias do tipo secundárias sugere que estes casos não são tão frequentes como o tipo 1 o que se torna benéfico para o hospital uma vez que cirurgias subsequentes representam custos adicionais significativos.

A facilidade de uso da ferramenta Weka está muito dependente da experiência que o utilizador possui com a técnica de DM. Porém, para quem já possui conhecimentos bem consolidados o Weka é intuitivo e bem organizado. Possui uma interface gráfica agradável, os menus e secções encontram-se bem organizados, divididos e estruturados por técnicas. Também os algoritmos estão bem organizados e estruturados, com fundamentação e explicação teórica do funcionamento de cada um. Existe documentação teórica bastante completa acerca do Weka, das técnicas e dos algoritmos. As soluções apresentadas não são muito atrativas, quase sempre exibidas em texto livre o que dificulta a leitura e análise dos dados; apenas a secção de visualização dos atributos é que já apresenta alguma atratividade na apresentação da informação. O tempo de execução depende inteiramente do algoritmo e *dataset* utilizados.

4.3.7 Tipos de Prioridades em Espera para Cirurgia

A colocação de um paciente em lista de espera para o bloco operatório está dependente de um fator atribuído pelo médico aquando da decisão de tratamento cirúrgico

do paciente. Esse fator é a prioridade e pode tomar quatro valores distintos, 1, 2, 3 ou 4, por ordem crescente de urgência. O problema colocado consistia em perceber qual o nível de prioridade atribuído mais frequentemente pelos profissionais, sendo que quanto maior for esse nível maiores são os custos associados para o hospital no sentido de ter que alocar os pacientes nos lugares mais cimeiros da lista de espera.

Para a obtenção, manipulação e visualização dos resultados relativos ao caso de estudo foram utilizados os módulos CDE (do Pentaho CE) e o PRD. Foram desenvolvidos *dashboards* apelativos e tabelas representativas da informação gerada. A Figura 4.28 apresenta um *dashboard* com a visão global dos resultados. São dispostos, tanto num gráfico circular como numa tabela as percentagens relativas à atribuição de cada nível de prioridade para cirurgia. Os valores *null* não foram ignorados porque representam dados em que faltam valores de registo.

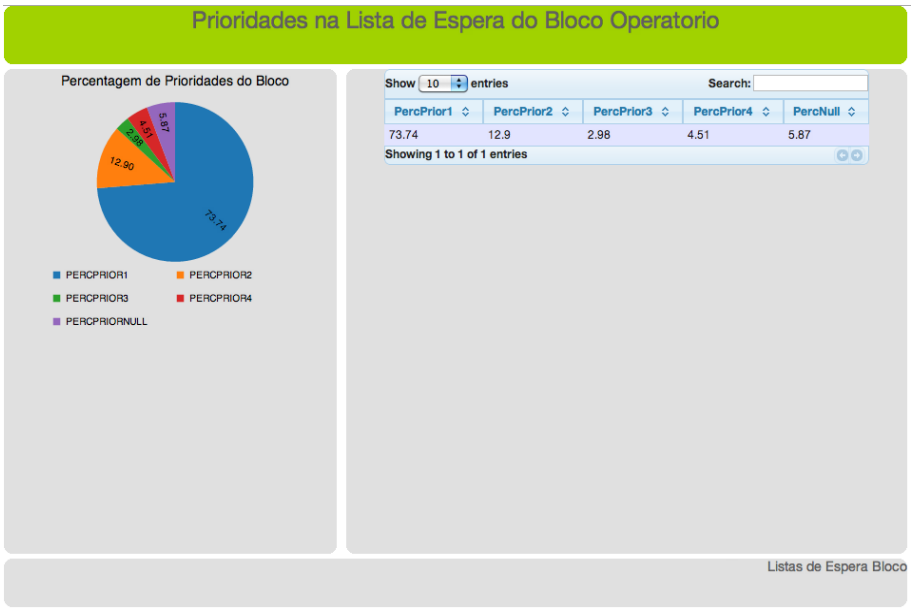


Figura 4.28: Valores percentuais da atribuição de cada nível de prioridade para cirurgia.

Dos resultados gerados, é possível concluir-se que o nível de prioridade mais atribuído aos casos cirúrgicos é o nível 1 (73.74%), considerada a situação menos urgente. Seguidamente, encontra-se o nível 2 (12.9%), sendo que o nível 4, a situação mais urgente é considerada mais vezes (4.51%) do que o nível 3 (2.98%) pelos médicos. A atribuição do nível 4 a um caso é o que aufere mais custos para a instituição uma vez que os pacientes têm que ser alocados na lista de espera nos lugares de topo, com a máxima urgência possível.

Por outro lado, considerou-se importante desenvolver uma tabela que apresentasse as diferentes especialidades por ordem de atribuição de determinado nível de prioridade aos respetivos casos clínicos. Os valores apresentados são percentuais. Assim, e sempre com o gráfico circular representativo dos valores percentuais gerais da atribuição de prioridades, foi desenvolvido um *dashboard* com uma tabela que dispõe de todas as especialidades e da percentagem de casos em que foi atribuída determinada prioridade. Por outro lado, construiu-se um gráfico de barras que apresenta o topo da lista de especialidades, isto é, apresenta sempre as 5 especialidades com maior valor de casos para aquele nível de prioridade. No caso de existirem mais de 5 especialidades com o

mesmo valor, este gráfico apresenta as 5 primeiras por ordem alfabética. No entanto, com o manuseamento da tabela é possível ter-se acesso a todas as especialidades e respetivos valores. Como o pretendido é a apresentação dos resultados em função do nível de prioridade atribuído, adicionou-se um parâmetro (*simple parameter*) de seleção da prioridade (*select component*), sendo que os valores da tabela e do gráfico são apresentados dependendo do nível de prioridade que se seleciona, tal como se pode verificar nas Figuras 4.29 e 4.30.

```
Select distinct des_grupo,
round(((A.parcial)/(select count(des_grupo) from bimov_listaesperablo
where prioridade=${prior}))*100,3) as X from bimov_listaesperablo,
(select count(bi.des_grupo) as parcial from bimov_listaesperablo bi
where bi.prioridade=${prior} group by bi.des_grupo) A
where bimov_listaesperablo.prioridade=${prior}
order by X desc
```

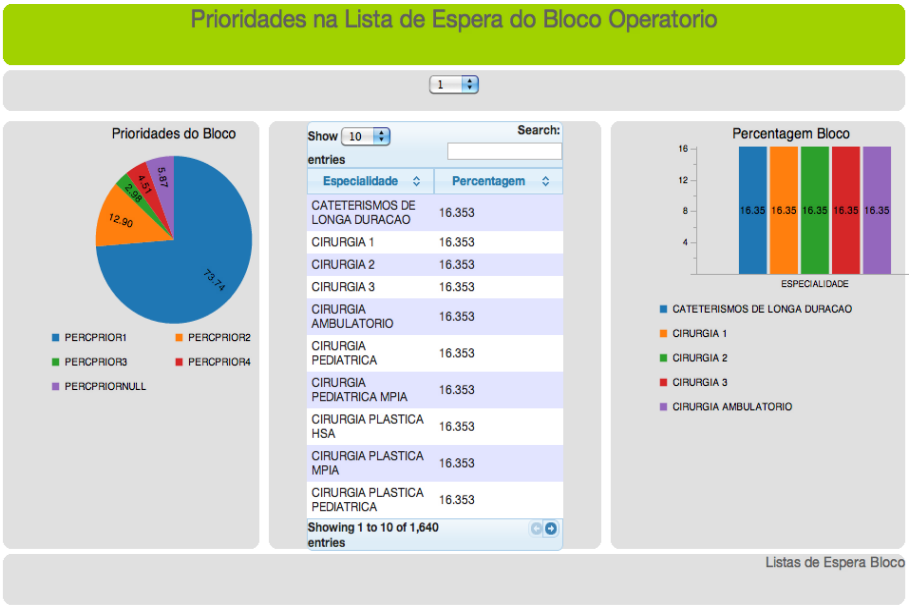


Figura 4.29: Apresentação dos resultados de atribuição do nível de prioridade 1 por especialidade médicas, com parâmetro de seleção simples.

Pela visualização da Figura 4.30, verifica-se a utilização de outros componentes: um componente de multi-seleção do parâmetro no topo do *dashboard* e um componente de texto na seção de baixo. A introdução de um componente de multi-seleção é semelhante ao de seleção simples, pois tem que se criar o parâmetro, definir o objeto HTML que lhe vai corresponder e a fonte de dados. O componente de texto permite o aparecimento dinâmico de um texto/frase de acordo com o parâmetro que se está a visualizar. Em parte é também semelhante nos passos de criação do parâmetro e seleção do objeto HTML e da fonte de dados, porém tem um ponto adicional onde é utilizada a linguagem JS para a geração do texto. Neste caso, na opção *'Expression'*, definiu-se a seguinte função:

function(){return 'Visualizando especialidades de' + nome_parâmetro}

Por outro lado, tentou-se também aplicar uma interatividade maior entre o *dashboard* e o utilizador aumentando assim a usabilidade da ferramenta, no sentido em que

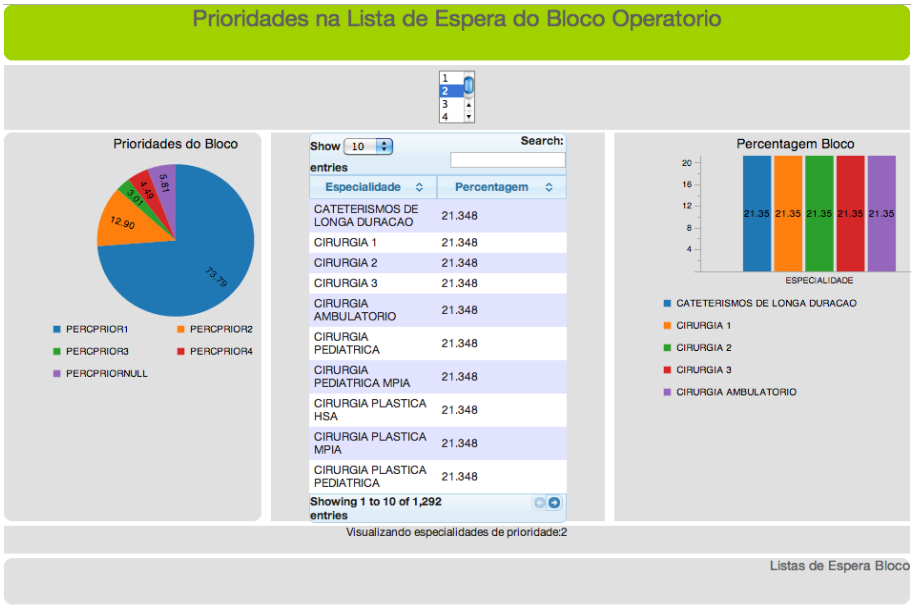


Figura 4.30: Apresentação dos resultados de atribuição do nível de prioridade 2 por especialidade médicas, com parâmetro de seleção múltipla e de texto.

se tornou o gráfico circular num gráfico interativo com capacidade de clique. Assim, o utilizador clica na fatia do gráfico circular correspondente a cada uma das prioridades e são gerados a tabela e o gráfico de barras em função desse parâmetro. Para tal, colocou-se a opção *'Cickable'* a *true* e definiu-se uma função *JavaScript* na opção *'Click Action'* para relacionar com os restantes *dashboards*, transcrita de seguida. Também se colocou uma opção para evitar o aparecimento da tabela e do gráfico de barras no processamento do dashboard, sendo que estes só aparecem após o clique no gráfico circular para a visualização da informação mais detalhada (*'Execute at Start'* = *False*).

function(s,c,v){Dashboards.fireChange('prior',s);}

A partir dos resultados obtidos, pode-se concluir que existem determinadas especialidades médicas que estão sempre no topo da lista com maior atribuição dos níveis de prioridade: Cirurgia 1, 2, 3 e Ambulatória. Posto isto, e devido à elevada incidência de casos nestas especialidade, as conclusões retiradas não se tornam muito relevantes. Porém, é possível observar-se que a especialidade Cateterismo de Longa Duração apresenta maior número de casos com níveis de prioridade mais baixa (1, 2 e 3) e que a especialidade Cirurgia Pediátrica apresenta uma elevada incidência de casos com prioridade máxima. Neste sentido, seria importante a implementação de medidas por exemplo de disponibilização de uma sala operatória mais direcionada para as intervenções cirúrgicas destas especialidades. O componente de tabela apresentado no *dashboard* não apresenta todos os resultados, uma vez que existe um número elevado de especialidades por prioridade, porém através de um botão é possível visualizarem-se as páginas seguintes.

Finalmente, com o conhecimento das especialidades pelo respetivo número de casos associado a cada nível de prioridade, considerou-se relevante calcular o número médio de dias de espera por cada nível de especialidade. Para tal, utilizou-se o PRD para apresentar os resultados numa tabela apelativa, Figura 4.31.

Prioridade	1	2	3	4
Média (Tempos Espera)	103.8	25.3	8.3	1.6

Figura 4.31: Média do número de dias de espera por nível de prioridade.

Tal como se pode verificar por observação da tabela, o nível de prioridade 4 apresenta a menor média de dias de espera para cirurgia, sendo então considerado o nível máximo de urgência. Em forma decrescente do nível de prioridade percebe-se que aumenta o número médio de dias de espera, sendo que o nível 3 apresenta uma média de espera de 8.3 dias, o nível 2 apresenta uma média de espera de 25.3 dias e o nível 1 - o menos urgente e o mais comum - apresenta uma média de tempo de espera de 103.8 dias.

4.3.8 Tipos de Cirurgias em Lista de Espera

No contexto das instituições de saúde é relevante a análise do tipo de cirurgias realizadas. Assim, foram analisados os tipos de cirurgia que se encontravam em lista de espera, podendo esta ser a primeira cirurgia do paciente ou então uma cirurgia subsequente. A realização de uma cirurgia subsequente representa uma percentagem significativa de custos para o hospital. Assim, foi desenvolvido um *dashboard* com o suporte do módulo CDE do Pentaho CE onde estão representados os totais de cirurgias em espera, divididas por primárias e secundárias. Foram construídos um gráfico circular e um gráfico de barras verticais, tal como se pode verificar na Figura 4.32.

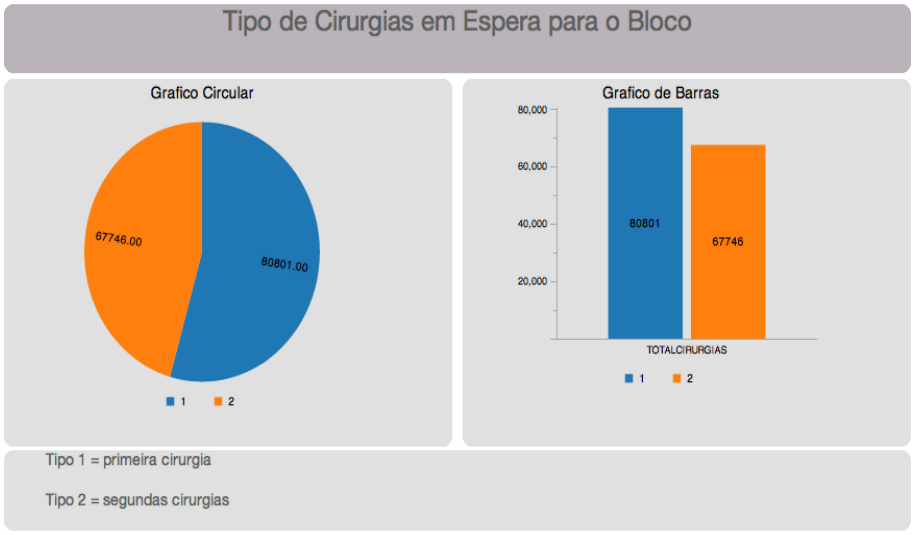


Figura 4.32: Representação gráfica (*dashboard*) do tipo de cirurgias em espera.

Para a análise foram considerados todos os casos presentes na lista de espera para cirurgia, exceto os casos em que a coluna tipo de cirurgia estava a *null*, perfazendo um total de 148547 casos de espera. Como conclusão, pode verificar-se que existe um maior número de cirurgias do tipo 1, primárias (80801). Porém, repara-se que o número de cirurgias subsequentes em espera contém um total significativo (67746) e, por isso, pode representar para o hospital um indicador importante de alerta que deve

ser analisado ao pormenor de forma a tentar diminuir este número no sentido em que se traduz em custos elevados para a instituição prestadora de cuidados de saúde.

4.4 Monitorização das Listas de Espera para Consulta

4.4.1 Registo de Casos na Lista de Espera para Consulta

A realização do estudo da lista de espera para consulta procedeu-se de modo similar e com o mesmo suporte de algumas das ferramentas utilizadas na lista de espera para cirurgia. Inicialmente, analisaram-se os totais de casos de espera por especialidade médica (Figura 4.33), apresentando-se assim uma visão global da informação referente à lista de espera para consulta. A consulta efetuada por SQL apresenta-se transcrita de seguida, bem como a tabela desenvolvida no módulo de *dashboards* do Pentaho EE com a organização da informação e a ordenação dos dados.

```
   Select * from(
select des_especialidade, count(des_especialidade) as Total
from bimov_listaesperacon
where to_char(dta_marcacao,'yyyy') between '2007' and '2012'
group by des_especialidade
order by Total desc
) where rownum<=10
```

DES_ESPECIALIDADE	TOTAL (Sum)
CE DERMATOLOGIA /HSA	27217
CE ORTOPEDIA /HSA	25281
CE CENTRO OFTALMOLOGICO /HSA	21585
CE O.R.L. /HSA	18791
CE CIRURGIA VASCULAR /HSA	15258
CE CIRURGIA AMBULATORIO /HSA	11341
CE NEUROLOGIA /HSA	10962
CE OFTALMOLOGIA ADICIONAL / HSA	10599
CE UROLOGIA /HSA	9710
CE ESTOMATOLOGIA /HSA	9589

Figura 4.33: Tabela das especialidades com os totais de espera para consulta (construído no Pentaho EE).

Foi também desenvolvido um gráfico de barras verticais, como se apresenta na Figura 4.34.

Por último, foi desenvolvido um *dashboard* composto por um gráfico de barras verticais e um gráfico circular através do módulo CDE do Pentaho CE (Figura 4.35).

A construção das soluções apresentadas seguiram a mesma metodologia das desenvolvidas para o caso de estudo das cirurgias. No anexo A.6 encontra-se o relatório

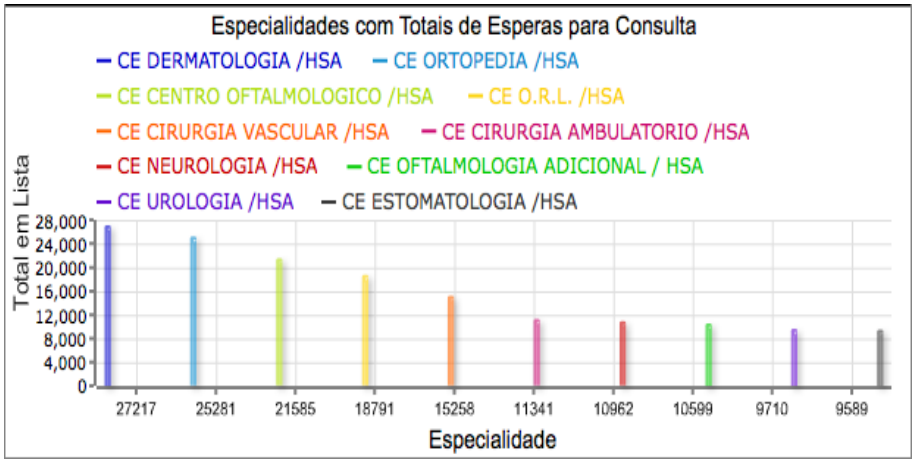


Figura 4.34: Gráfico de barras verticais das especialidades com os totais de espera para consulta (construído no Pentaho EE).

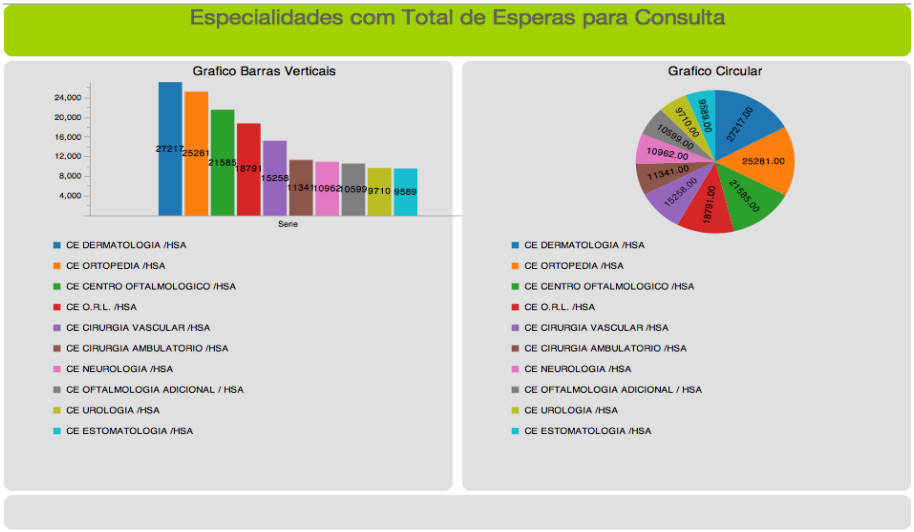


Figura 4.35: *Dashboard* das especialidades com os totais de espera para consulta (construído no Pentaho CE).

completo construído com o suporte *Interactive Report* do Pentaho EE, onde se apresentam todas as especialidades médicas e os respetivos totais de casos de espera para consulta.

4.4.2 Registos Ativos em Lista de Espera para Consulta

Para a análise dos casos ativos em lista de espera para consulta, em analogia com o caso de estudo do bloco, construiu-se um DW com o suporte do PDI apenas com os dados relevantes para o caso. O instante temporal marcado na lista de espera como de análise de registos ativos em lista naquele momento foi o dia 17 de Abril de 2012, num total de 13333 registos ativos para esse mesmo instante. Neste caso de estudo para as consultas apresenta-se o *dashboard* desenvolvido no Pentaho CE (Figura 4.36), onde se fará uma análise dos resultados e desenvolvimento do mesmo, finalizando com uma comparação entre os *dashboards* desenvolvidos no Pentaho CE e EE (4.3.2).

Figura 4.36: Gráfico de pontos desenvolvido no Pentaho CE, onde estão representadas as 10 especialidades com o maior número de registos ativos naquele momento.

Este tipo de gráfico no Pentaho CE é desenvolvido utilizando o componente CCC do tipo de gráfico pontos. Foi definido um gráfico horizontal para melhor visualização e compreensão, tendo-se registado uma maior complexidade na construção deste dashboard. Foi definido um *offset* do eixo de 0.1 para que todos os valores aparecessem no gráfico, e colocou-se uma grelha em relação ao eixo para melhor perceção dos limites, uma vez que se estão a visualizar pontos. Também se colocaram os valores precisos de cada ponto ao pé destes. Verifica-se uma falha no aparecimento dos dados do eixo das ordenadas onde as especialidades são cortadas, não se tendo conseguido resolver o problema. Também o último valor correspondente à especialidade com maior número de registos ativos corresponde a 5310, mas sem explicação aparente o *dashboard* não apresentou a casa das unidades. Da análise do gráfico conclui-se que no momento de análise as especialidades médicas mais representadas na lista de espera para cirurgia eram Oftalmologia e Cirurgia Ambulatória.

Numa avaliação global dos dois módulos verificou-se uma maior simplicidade e facilidade na construção do *dashboard* através do Pentaho EE. Por outro lado, pela análise dos mesmos constata-se uma maior atratividade das soluções geradas no Pentaho EE, sendo os gráficos muito mais apelativos, com cores mais atrativas e com maior dinamismo.

4.4.3 Consultas em Lista de Espera (períodos mensais)

Em relação ao estudo e análise do número de registos que obtiveram retorno de marcação de consulta de especialidade e, que por isso, correspondem a saídas da lista de espera para consultas por períodos mensais, os procedimentos foram os mesmos da análise do número de cirurgias utilizando-se as mesmas ferramentas de suporte e obtendo-se resultados do mesmo tipo e com as mesmas caraterísticas.

Na Figura 4.37, encontra-se representado o *dashboard* com o número de retornos para consulta em lista de espera no ano 2007, e na Figura 4.38 está representado o número de consultas em lista de espera ao longo de vários anos (2007-2011) num determinado mês, neste caso o mês 07 (correspondente a Julho). A escolha por estes ano e mês específicos em detrimento dos outros resultados foi aleatória uma vez que todos os gráficos apresentavam um comportamento similar (restantes gráficos no anexo A.8).

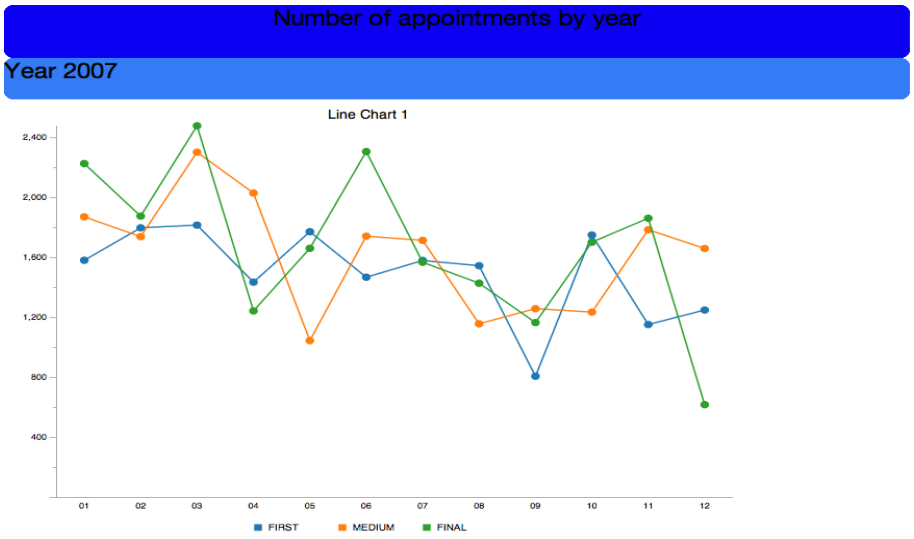


Figura 4.37: Número de consultas em lista de espera ao longo do ano 2007.

Da observação da Figura 4.37, podem-se retirar algumas conclusões de tendências detetadas pela visualização e análise da informação. Repara-se que no último período do mês de Dezembro e no primeiro período do mês de Janeiro existe um decréscimo do número de retornos em lista para consulta de especialidade, motivo justificado pelos períodos de férias. Por outro lado, ao contrário do que acontece com o caso de estudo das cirurgias não se nota um decréscimo do número de consultas ao longo do mês de Agosto. Também não se regista nenhuma tendência de aumento ou decréscimo do número de retornos em lista de espera para consulta ao longo do ano. Finalmente, importa realçar que os dados relativos ao ano 2012 encontram-se incompletos por ser o ano corrente.

Em relação ao número de retornos para marcação de consulta ao longo dos vários anos para um determinado mês, verifica-se em geral que o número de consultas tem aumentado ligeiramente ao longo dos anos. Através da visualização e análise dos gráficos é também possível detetarem-se erros de registo de dados. Na Figura 4.39 encontra-se um registo de uma consulta que recebeu um retorno no ano 2099, o que é obviamente impossível e por isso conclui-se que se trata de uma falha de introdução de dados. Contudo, este problema de erros de registo será mais aprofundado aquando da utilização

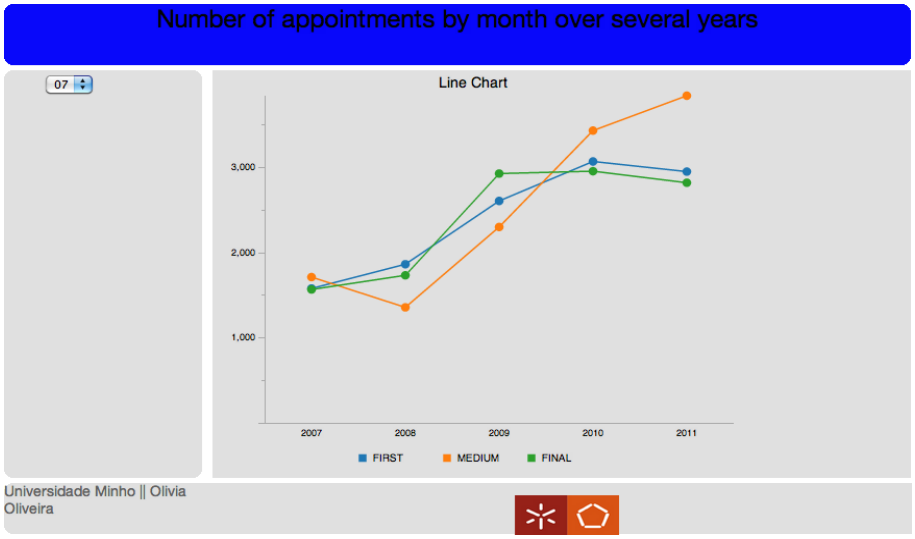


Figura 4.38: Número de consultas em lista de espera durante vários anos no mês de Julho.

da ferramenta de DM, WEKA.

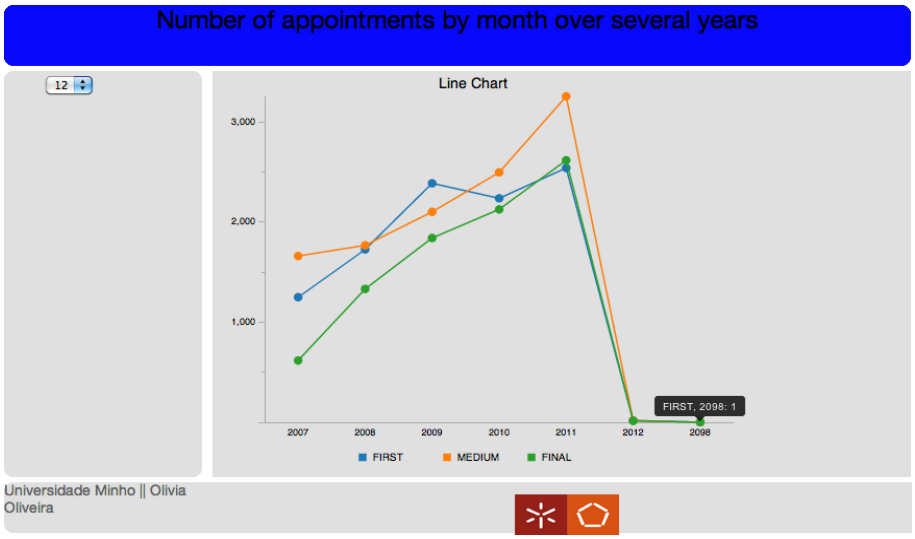


Figura 4.39: Número de consultas em lista de espera durante vários anos no mês de Fevereiro, com realce para uma falha de registo.

4.4.4 Lista de Espera para Consulta por Distritos

O mesmo tipo de problema foi colocado para as listas de espera para consulta médica. Em analogia ao caso de estudo para a lista de espera para bloco foi analisada a naturalidade dos pacientes para consulta. Utilizou-se também o módulo Pentaho Analyzer da EE com o suporte do módulo GeoMap, retirando-se da BD os dados geográficos dos pacientes.

Inicialmente, procedeu-se à criação do esquema multidimensional com os dados relevantes para realizar a análise. A manipulação da BD foi feita através da utilização

do *software* SQL Developer e a consulta e seleção dos dados através de *queries* SQL. De seguida, definiram-se as medidas e dimensões a analisar, sendo que no módulo GeoMap é possível atribuírem-se dimensões do tipo geográficas, tal como já foi referido anteriormente. Igualmente ao caso de estudo para a lista de espera para bloco também neste caso de estudo os distritos correspondem às principais cidades portuguesas. A informação extraída foi estruturada e organizada numa tabela (Figura 4.40).

PAIS	DISTRITO	TOTAL	% of TOTAL
PORTUGAL	AVEIRO	294	4,06%
	BRAGA	207	2,86%
	BRAGANCA	71	0,98%
	CASTELO BRANCO	2	0,03%
	COIMBRA	6	0,08%
	FARO	1	0,01%
	GUARDA	4	0,06%
	LEIRIA	2	0,03%
	LISBOA	6	0,08%
	PORTO	6371	87,94%
	SETUBAL	5	0,07%
	VIANA DO CASTELO	48	0,66%
	VILA REAL	76	1,05%
	VISEU	152	2,10%

Figura 4.40: Número de pacientes em lista de espera para consulta organizado por distritos.

Seguidamente, a informação geográfica foi integrada no GeoMap de forma a disponibilizar uma visualização gráfica e geográfica dos dados, sendo possível observar a distribuição do número de pacientes por distritos no mapa de Portugal (Figura 4.41).

Finalmente, e também com o suporte do Pentaho Analyzer EE procedeu-se à construção de diversos tipos de gráficos que permitem uma organização e visualização gráfica da informação (Figura 4.42, sendo que os restantes se encontram no anexo A.9).

A partir da análise da informação analisada e apresentada estruturadamente e de forma gráfica, observou-se uma maior representação do distrito do Porto no número de casos (aproximadamente 88%), tal como já tinha acontecido na lista de espera para bloco. Por outro lado, distritos adjacentes ao Porto apresentam também um elevado número de pacientes alistados (como Aveiro, Braga, Viseu, Vila Real). Este tipo de conhecimento gerado pode-se tornar útil para hospital na medida em que se conhece a origem dos pacientes em lista de espera para consulta.

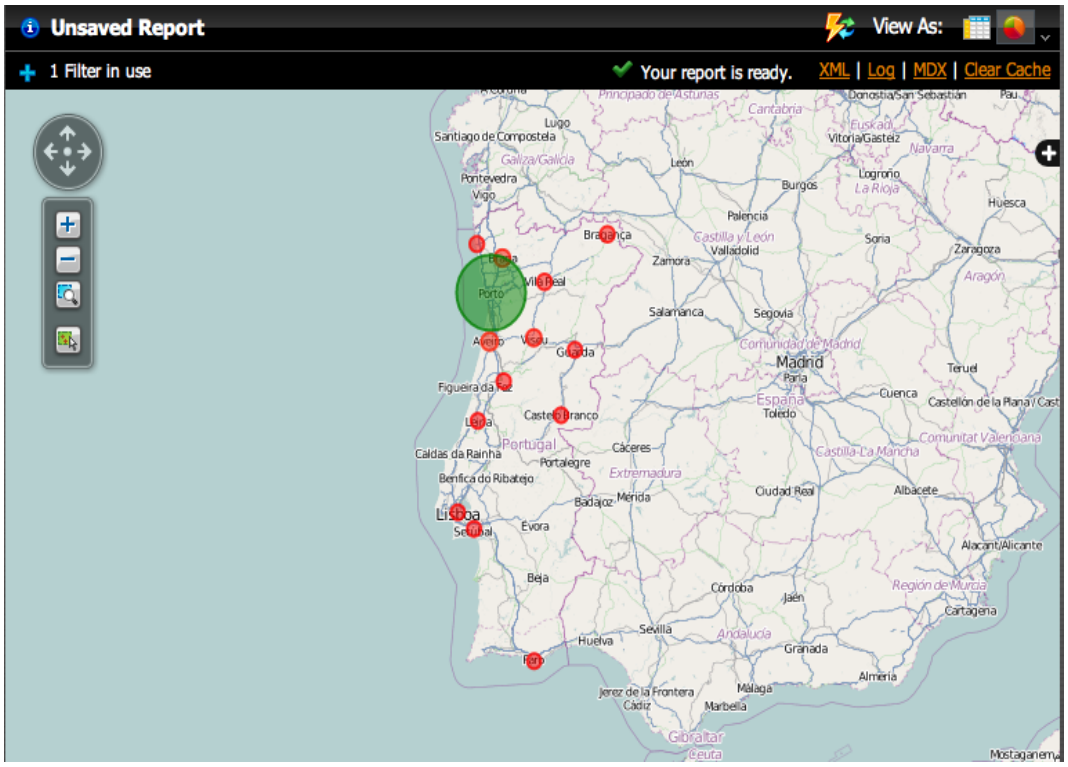


Figura 4.41: Visualização da informação geográfica dos pacientes em lista de espera para consulta utilizando o módulo GeoMap.

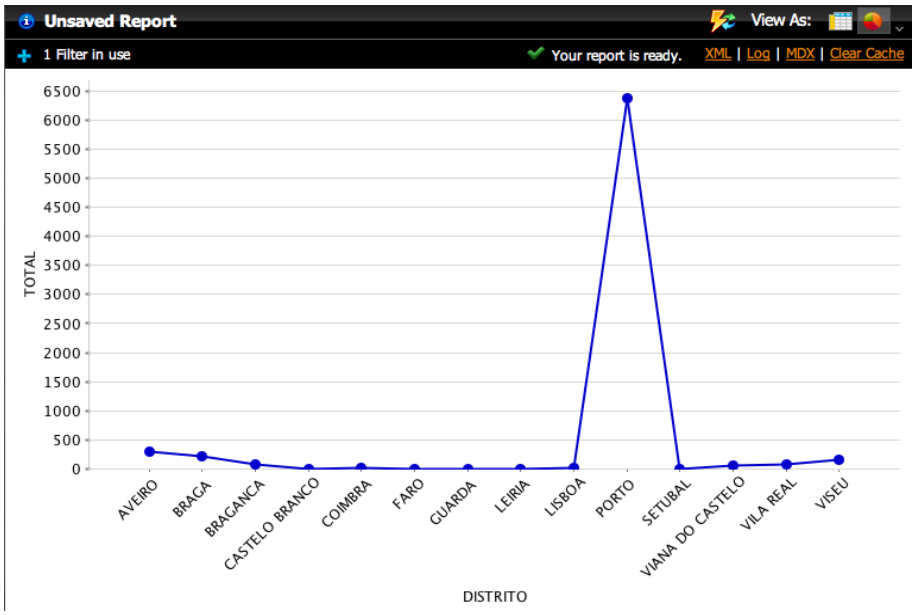


Figura 4.42: Gráfico linear, representando o número de pacientes em lista de espera para consulta por distrito de origem.

4.4.5 Lista de Espera para Consulta por concelhos do distrito do Porto

Para a análise OLAP dos dados da lista de espera para consulta, e de forma análoga ao caso de estudo para bloco, os dados geográficos dos pacientes foram organizados de acordo com o seu concelho de origem. Neste caso particular considerou-se apenas o distrito do Porto, uma vez que se torna o mais relevante para a análise no contexto do hospital em estudo. Repetidamente, utilizou-se o Pentaho Analyzer EE e com o suporte do GeoMap para a estruturação e visualização da informação geográfica (Figura 4.43).

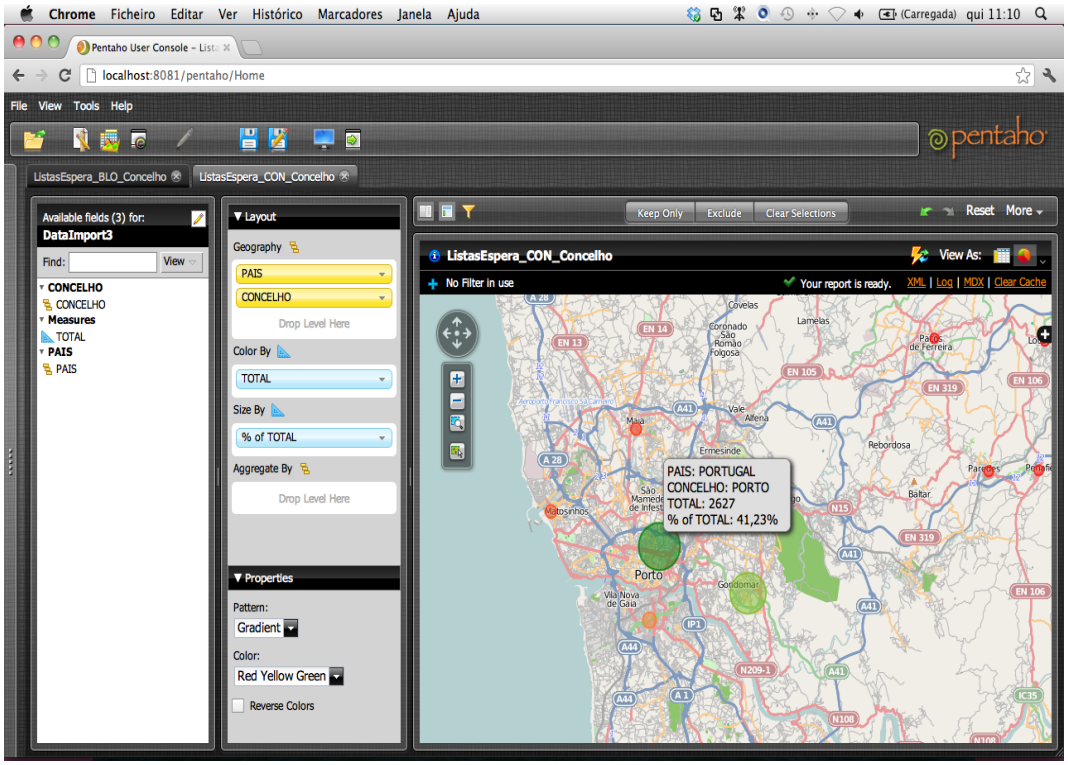


Figura 4.43: Visualização dos dados geográficos dos pacientes em lista de espera para consulta utilizando o módulo GeoMap.

Tal como é possível verificar na Figura 4.43, este módulo permite a interação com o utilizador na medida em que quando este coloca o cursor do rato em cima de cada concelho, é disponibilizada a informação detalhada relativa a esse concelho, como o nome, o total de pacientes originários e a respetiva percentagem.

Seguidamente, estruturou-se a informação por concelhos segundo a Figura 4.44, sendo que primeiramente se procedeu à aplicação de um filtro para canalizar apenas os resultados referentes ao distrito do Porto.

Posteriormente, realizou-se o tratamento da informação de modo gráfico, apresentando-te o gráfico circular (Figura 4.45) e os restantes no anexo A.10.

Concluindo, de um modo global os casos de estudo tendem todos, 2 a 2, para as mesmas inferências e padrões. Os casos relativos à análise por distritos de Portugal permitem concluir que o distrito mais representado é o Porto e, de seguida, Aveiro e Braga tanto para as listas de espera para bloco como para consulta. Este fato é previsível e lógico uma vez que Aveiro e Braga são os distritos mais próximos do Porto. Por outro lado, é possível notar que Faro e Castelo Branco são os distritos menos

PAIS	CONCELHO	TOTAL	% of TOTAL
PORTUGAL	AMARANTE	91	1,43%
	BAIAO	57	0,89%
	FELGUEIRAS	16	0,25%
	GONDOMAR	2191	34,39%
	LOUSADA	20	0,31%
	MAIA	182	2,86%
	MARCO DE CANAVESES	75	1,18%
	MATOSINHOS	257	4,03%
	PACOS DE FERREIRA	23	0,36%
	PAREDES	82	1,29%
	PENAFIEL	74	1,16%
	PORTO	2627	41,23%
	POVOA DE VARZIM	36	0,57%
	SANTO TIRSO	22	0,35%
	TROFA	5	0,08%
	VALONGO	126	1,98%
	VILA DO CONDE	60	0,94%
	VILA NOVA DE GAIA	427	6,70%

Figura 4.44: Número de pacientes em lista de espera para consulta organizado por concelhos.

representados, fato talvez explicado pela distância destes ao distrito do Porto. Em relação à análise realizada por concelhos do distrito do Porto, o Porto é o concelho mais representado seguido de Gondomar e Vila Nova de Gaia. Fatos igualmente previsíveis e com a mesma explicação dos anteriores. Percebe-se também que o concelho da Trofa é o menos representado nas listas de espera para bloco e consulta, fato talvez explicado pela proximidade deste concelho ao distrito de Braga, e por isso, o possível reencaminhamento destes pacientes para o hospital de Braga.

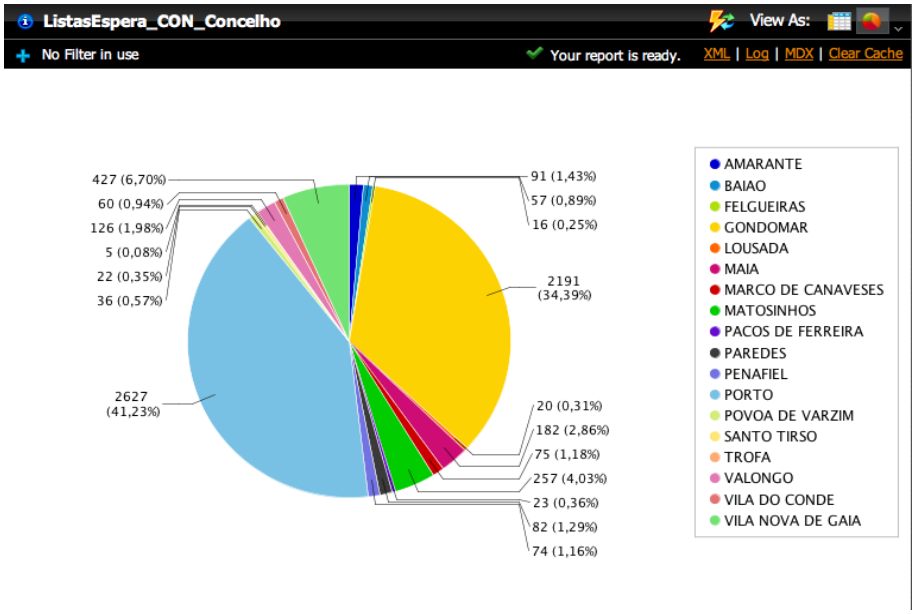


Figura 4.45: Gráfico circular com o número de pacientes em lista de espera para consulta por concelhos do distrito do Porto.

4.4.6 Registo de Óbitos em Lista de Espera para Consulta

O problema apresentado consistiu no tratamento dos casos de morte e na tentativa de interseção e relacionamento destes com outros dados, de forma a detetar possíveis falhas presentes no funcionamento do hospital. A partir da BD hospitalar foi construído um DW, onde se armazenou o conjunto de dados relevantes para o problema apresentado.

Os dados clínicos possibilitaram a interseção de dados relativos aos óbitos e à lista de espera para consulta. Deste modo, através do identificador único de cada paciente (número sequencial) foi possível aceder aos casos em que foram registados óbitos no hospital em estudo e que se encontravam ativos em lista de espera para uma consulta médica. O objetivo principal foi tentar encontrar uma tendência de especialidades em que se registassem mais mortes e tentar descobrir alguma relação de forma a tomar medidas preventivas no estabelecimento de saúde. Desta forma, procedeu-se à criação de um *dashboard* através do módulo CDE do Pentaho CE no qual são apresentados dois tipos de gráficos (barras verticais e circular). Tal como é possível observar na Figura 4.46, nos gráficos estão representadas as especialidades da lista de espera para consulta com mais casos de registo de óbitos. Para tal, foram considerados os casos para o instante de tempo marcado como ativos em lista de espera os que apresentavam uma data de receção de pedido de consulta anterior a uma data de óbito, e que os campos de código de motivo de recusa e número de taxa estivessem a *null*.

No desenvolvimento do *dashboard* as cores correspondentes a cada uma das especialidades foram definidos pelo utilizador. No Pentaho CE quando existem mais do que 10 valores, os gráficos aparecem com cores repetidas tornando-se impossível fazer a correspondência gráfica com a legenda. Assim, definiu-se manualmente a palete de cores a utilizar através da opção *'Colors'*, sendo que inicialmente se optou por código HTML com o intuito de se escolher mais especificamente cada tonalidade. Porém reparou-se que na situação do gráfico circular a definição das cores deste tinha de ser feita através

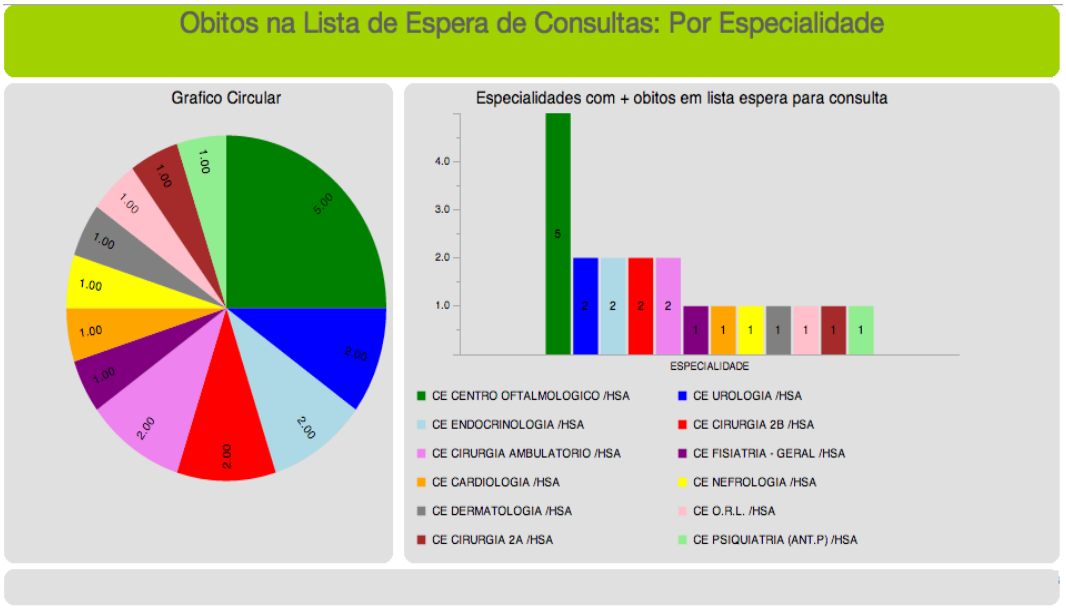


Figura 4.46: Especialidades da Lista de Espera para Consulta com maior número de registo de mortes.

do próprio nome das cores que existem na biblioteca do Pentaho e por isso o desenvolvimento teve que se remeter às cores básicas existentes. De acordo com os resultados obtidos e através do suporte de visualização gráfica disponibilizado pela ferramenta é possível detetar as especialidades de lista de espera com ocorrência de mortes, sendo a mais incidente a 'CE Centro Oftalmológico HSA', podendo então ser considerada a especialidade mais crítica.

Foram, de seguida, desenvolvidos dois relatórios (anexo A.11) com a informação relativa às especialidades médicas da lista de espera em que existiu registo de óbitos. O objetivo prendeu-se com a avaliação e comparação entre o desenvolvimento de um relatório totalmente de raiz por parte do utilizador e do desenvolvimento de um relatório com o suporte do assistente de relatórios (*Wizard Report*). Para tal, utilizou-se a ferramenta PRD.

O desenvolvimento de um relatório de raiz considerou-se relativamente complexo no sentido em que tem que se criar a ligação com a fonte de dados e definir a consulta que se pretende realizar. Também todo o processo de desenvolvimento do relatório é dependente das ações do utilizador desde a definição do título, do rodapé, e do cabeçalho, à atribuição dos espaços do *layout* aos campos, tabelas, linhas e colunas, palete de cores, fonte e tipo de letra, até ao mais ínfimo pormenor. Existe uma dificuldade considerável uma vez que a construção do relatório não apresenta uma filosofia de WYSIWYG, sendo que para se visualizar o resultado final tem que se definir o modo de pré-visualização. Por outro lado, e de um modo muito mais simples, a utilização do assistente de relatórios disponibiliza um conjunto de passos (Figura 4.47) mais ou menos automáticos em que o utilizador apenas vai definindo as diversas hipóteses e no fim tem a opção de pré-visualizar o relatório para confirmar ou não a validade do desenvolvimento.

Primeiramente define-se o tema que se pretende utilizar (o PRD contém uma biblioteca com 5 temas, podendo também importarem-se outros temas); de seguida define-se

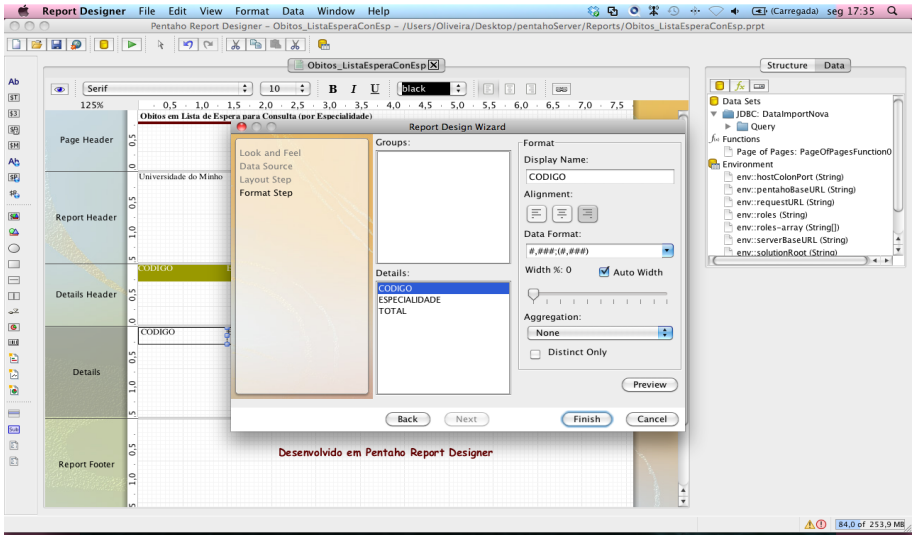


Figura 4.47: Desenvolvimento de um relatório utilizando o assistente de relatórios Wizard.

o modelo e a fonte de dados; posteriormente, selecionam-se e agrupam-se os dados de acordo com o esquema que se pretende no relatório; por fim tal como está apresentado na Figura 4.47 definem-se os detalhes de cada um dos campos como o nome, alinhamento, formato, existindo também a possibilidade de fazer um *preview* antes de finalizar o processo de desenvolvimento. Em ambas as soluções verifica-se que uma das colunas correspondente ao código de especialidade, sendo este representado segundo um campo numérico. O PRD apresenta os número com um ponto na casa dos milhares. Tentou-se retirar o ponto mas não foi possível, considerando-se que a principal justificação para esta falha remete-se para o grande direcionamento do PRD, e do Pentaho em geral, para a análise de custos e vendas, campos em que se torna realmente interessante separar casas numéricas. Em conclusão, para um utilizador que o objetivo seja a criação de um relatório de forma simples e eficaz não sendo muito exigente na definição do tema e das cores a utilizar, bem como em funções mais específicas, verifica-se que os relatórios *Wizard* são os mais adequados e também os mais simples.

Através da ferramenta PRD é também possível introduzirem-se diversos tipos de gráficos e importarem-se imagens. A construção de gráficos é relativamente simples, onde se definem as variáveis, colocam-se os títulos e legenda, opção de 3D, permite a definição da rotação da *label* do eixo das abcissas (uma falha apontada ao módulo CDE do Pentaho CE), entre muitas outras definições. A importação de imagens é bastante simples pois apenas se insere uma *label* para imagens e define-se a diretoria onde se encontra guardada a imagem. O PRD tem inúmeras funcionalidades para além da construção de relatórios, o que se torna uma vantagem pois permite então que o utilizador se forme especificamente nesta ferramenta não necessitando de tomar conhecimento de outros módulos, como por exemplo o CDE para a construção de gráficos.

Posteriormente, prosseguiu-se uma análise mais profunda aos dados explorados na tentativa de interligar a especialidade da lista de espera para consulta e a especialidade onde foi registado como óbito. Porém, e como é óbvio dado se estarem a tratar de casos de morte e por isso de situações graves, a relação é ambígua. A maior parte dos casos

de morte foram registados em unidades críticas como Medicina, Gastroenterologia, Área de Decisão Clínica, Unidade de Cuidados Intensivos, tal como se pode verificar na Figura 4.48. Esta tabela resulta da geração de um relatório através do módulo *Interactive Report* (Pentaho EE).

ESPECIALIDADE_OBITOS	TOTAL_DE_CASOS
INT MEDICINA C /HSA	2
INT GASTRENTEROLOGIA /HSA	2
INT AREA DECISAO CLINICA/SO HSA	2
INT CUIDADOS INTENSIVOS /HSA	1
INT CIRURGIA 1 /HSA	1
INT UROLOGIA /HSA	1
INT UNID.CUID.INTERMEDIOS URGENCIA/HSA	1
INT MEDICINA B /HSA	1
URG-SU ANESTESIOLOGIA /HSA	1

Figura 4.48: Apresentação das especialidades com maior ocorrência de registo de óbitos.

O PRD não é uma ferramenta muito intuitiva, o utilizador deve possuir uma formação ou então a sua formação ser relativamente avançada. A sua utilização e exploração não é simples e a visualização e leitura dos dados apenas é feita na pré-visualização. Todo o seu desenvolvimento consiste em *labels* e campos genéricos. O assistente de relatórios *Wizard* veio aumentar a usabilidade da ferramenta, no sentido em que favorece a facilidade da sua utilização. As soluções apresentadas foram consideradas atrativas e a qualidade gráfica da ferramenta alta. Apresenta simplicidade dos menus e *layout* com boa organização e estruturação. Por outro lado, verifica-se reversibilidade das ações realizadas. Um ponto fraco a apontar é a impossibilidade de apagar algum componente através da tecla *Delete* ou de qualquer atalho do teclado; a eliminação tem que ser feita através do rato. Por fim, constatou-se alguma escassez na documentação teórica existente bem como no suporte técnico de erros apresentados pela ferramenta.

Em relação ao módulo *Interactive Report* do Pentaho EE verificou-se que o desenvolvimento de relatórios é bastante intuitivo, amigável e interativo com o utilizador, baseando-se na filosofia *drag-and-drop*. As soluções são atrativas, mas os temas possíveis de utilizar remetem-se apenas para os existentes na biblioteca da plataforma. Por outro lado o tempo de execução é praticamente instantâneo. A qualidade gráfica da interface é razoável, e a documentação teórica existente é praticamente nula também devido a este módulo ser recente e pertencente à plataforma EE.

Em conclusão ao caso de estudo especificamente, considerando a área de prevenção e aplicação de medidas administrativas para uma gestão mais eficiente da instituição seria importante analisar ao pormenor as especialidades da lista de espera para consulta para perceber se há uma relação direta com o seu número elevado de mortes. Deste modo, uma das medidas possíveis para implementação era o encurtamento do prazo de espera para consultas destas especialidades críticas recrutando, por exemplo, mais profissionais de saúde especialistas nas áreas.

4.4.7 Óbitos em Lista de Espera para Consulta - Análise de DM

Este caso de estudo baseia-se no cruzamento de dados de duas tabelas da BD (lista de espera para consulta e processo clínico eletrónico dos óbitos). Deste modo, pretendeu-se encontrar algum indicador e detetar correlações entre os óbitos registados no hospital e a espera ativa por uma consulta de especialidade. Para tal, foram considerados inicialmente 4 atributos: a especialidade em lista de espera para consulta, a especialidade onde o doente foi registado como óbito, o estado da espera e se faleceu no hospital. Os restantes atributos foram removidos no pré-processamento através de um filtro não supervisionado. Foi utilizada a técnica de DM associação (algoritmo PredictiveApriori).

A computação dos dados de acordo com a técnica de associação resultou em algumas regras. O algoritmo PredictiveApriori procura o melhor N para as regras de associação, com um limite de *threshold* baseado num valor de confiança corrigido. De acordo com o treino do *dataset* foram definidas 28 regras , apresentando-se apenas duas regras a título de exemplo: pode-se afirmar com uma precisão de 0.99479 que os casos em que foram registados óbitos como falecendo no hospital, encontravam-se com um estado ativo na lista de espera. Por outro lado, afirma-se com uma precisão de 0.97141 que os casos registados como óbitos na especialidade de Gastroenterologia apresentavam-se em espera ativa para consulta da especialidade Cirurgia 2B com estado de espera P, e faleceram no hospital.

Seguidamente, foi introduzido mais um atributo (código da patologia) ao *dataset* de forma a computar técnicas de classificação. Os algoritmos utilizados foram RandomTree (Árvores de Decisão), DecisionTable (Regras de Classificação) e IBK (Lazy). Para a computação do novo *dataset* este sofreu um pré-processamento dos dados. As regras de classificação exigem que os dados se encontrem todos discretizados e, para tal, utilizou-se um filtro supervisionado de discretização aplicado a todos os atributos do *dataset*.

O algoritmo Random Tree permite a construção de uma árvore de decisão baseada numa escolha aleatória de K atributos em cada nó. A computação resultou na classificação correta de 85% das instâncias, num total de 20. O valor da estatística Kappa obtido foi de 0.6809. A estatística Kappa mede a conformidade entre as previsões e os valores da classe atual, ou seja, normalmente deve ser um valor próximo de 1, para que as previsões se aproximem o máximo possível das classes dos objetos reais. O erro absoluto médio foi de 0.17, o erro quadrático médio foi de 0.2915, o erro absoluto relativo foi de 35.283% e o erro quadrático relativo foi de 59.5017%. Finalmente, surge o resultado da matriz de classificação que consiste na organização dos resultados de acordo com o problema. Assim sendo, a posição [1,1] corresponde aos verdadeiros positivos (número de instâncias corretamente classificadas positivamente), a posição [1,2] corresponde aos falsos negativos (número de instâncias que foram incorretamente classificadas negativamente), a posição [2,1] corresponde aos falsos positivos (número de instâncias que foram incorretamente classificadas positivamente) e a posição [2,2] corresponde aos verdadeiros negativos (número de instâncias que foram corretamente classificadas negativamente). Da análise da matriz de classificação verifica-se que existe uma percentagem considerável de instâncias corretamente classificadas positivamente (6 instâncias) e negativamente (11 instâncias), para 2 instância classificadas incorreta-

mente negativamente e 1 positivamente.

O algoritmo Decision Table permite a construção e utilização de uma tabela de decisão simples. O treino consistiu num *split* de 66% (ou seja, apenas foi treinado 34% do *dataset*), num total de 7 instâncias. O valor da estatística Kappa foi 0.5882, o erro absoluto médio foi de 0.2881, o erro quadrático médio foi de 0.3009, o erro absoluto relativo foi de 48.0159% e o erro quadrático relativo 47.3078%. Para além do valor da estatística Kappa e dos baixos valores de erro, pode-se observar que este modelo é bom no contexto do problema, verificando-se 85.7143% das instâncias corretamente classificadas (5 positivamente e 1 negativamente).

O IBK é um algoritmo classificador baseado no vizinho mais próximo e num processo de aprendizagem por analogia. Seleciona apropriadamente o valor dos K vizinhos através de validação cruzada, atribuindo pesos à distância entre eles. A computação deste algoritmo resultou em 90% das instâncas classificadas corretamente num total de 20. O valor da estatística Kappa obtido foi 0.7826, o erro absoluto médio foi de 0.1191, o erro quadrático médio foi de 0.223, o erro absoluto relativo foi de 24.7212% e o erro quadrático relativo foi de 45.5084%. Através da análise da matriz de classificação repara-se que a grande parte das instâncias foram corretamente classificadas negativamente (verdadeiros negativos – 12 instâncias) e 6 instâncias corretamente classificadas de forma potitiva.

Concluindo, estes resultados (resultados completos da análise no anexo A.12) não servem para tirar conclusões expressivas, uma vez que se está a lidar com apenas 20 instâncias. Devido à delicadeza do tema não é possível retirarem-se mais conclusões.

4.4.8 Número de Consultas em Lista de Espera Anual - Deteção de Falhas (Análise DM)

A utilização da ferramenta de DM, Weka permitiu realizar um estudo para deteção de falhas no registo de dados de consultas para lista de espera. Para tal, procedeu-se à extração dos dados da tabela de consultas em lista de espera e procedeu-se à sua transformação no sentido de se obterem apenas os anos da data prevista de retorno. Efetuou-se também uma limpeza dos dados nulos que não continham qualquer data de retorno, eliminando-os da análise. Por fim, introduziram-se os dados organizados num ficheiro do tipo arff, formato suportado pelo Weka. Este processo foi realizado através do PDI, ferramenta de ETL. Neste âmbito foram utilizados algoritmos baseados na técnica de *clustering*, entre eles o SimpleKMeans, o EM (*Expectation Maximisation*), o FarthestFirst e o DBScan (resultados da análise no anexo A.13).

O algoritmo SimpleKMeans agrupa dados utilizando o algoritmo K-Means. Para tal, pode utilizar-se a distância euclidiana (por definição) ou a distância de Manhattan (computa os dados de acordo com o valor da mediana, em vez da média). Assim, no contexto do problema foi utilizada a distância de Manhattan (pois tratam-se de valores das listas de espera) e definiu-se como número de *clusters* 5. Os resultados vêm de encontro às conclusões de outras análises, no sentido em que agrupou as instâncias relativas aos anos de 2007, 2008, 2009, 2011 e 2012 como os que são mais prováveis, sendo o ano de 2011 o mais registado com 79% das instâncias agrupadas. Através deste algoritmo, não é possível detetaram-se falhas de registo, mas tomar conhecimento dos dados mais vezes registados e, por isso, mais prováveis. O resultado da computação dos dados através do algoritmo EM permite concluir que existe apenas um *cluster*

que organiza os diferentes anos por número de vezes que cada instância foi registada. Através deste resultado não há divisão dos dados por grupos, porém dá para saber quais as instâncias menos prováveis e assim detetar erros de introdução de dados. Como resultado da computação do algoritmo Farthest First, obteve-se que o grupo central seria a instância 2011 e o ponto mais distante a instância 2055. Daqui se conclui que existe um erro na introdução de dados, onde foram registadas datas referentes ao ano de 2055, o que é previsivelmente impossível. Definindo um maior número de computação de pontos mais distantes encontram-se obviamente mais instâncias que podem ser consideradas falhas. O algoritmo DBScan é baseado em densidade para a descoberta de *clusters* em BDs extensas com ruído. Neste contexto, o ruído corresponde às falhas de registo, sendo que no resultado da computação todas as instâncias com ruído correspondiam realmente a anos introduzidos sem sentido para o contexto, como por exemplo 2055, 2027, 2090, 3012.

Finalmente, foi aplicado um algoritmo baseado na técnica de classificação, o ZeroR. Este algoritmo prevê a categoria maioritária com a construção de uma tabela com as instâncias e os valores de frequência. Como resultado, obteve-se o ano 2011 como o mais frequente no *dataset* classificado. Aproximadamente, 79% das instâncias foram corretamente classificadas num total de 7669 instâncias. Este modelo mostrou-se adequado e com qualidade para o contexto do problema.

4.4.9 Análise dos Tempos de Espera para Consulta

Comummente, realizam-se estudos para explorar os tempos de espera em listas hospitalares para consultar um médico. Neste sentido, efetuam-se análises dos dias de espera dividindo-os por intervalos de tempo de forma a perceber quais os períodos de tempo de espera que acontecem mais vezes.

Neste caso de estudo em particular, analisaram-se os tempos de espera para realização de uma consulta após retorno da marcação da mesma, dividindo-se assim os períodos de dias em intervalos equivalentes a um mês. Assim, foram calculados os valores percentuais relativos ao tempo de espera até 1 mês, entre 1 e 2 meses e assim consecutivamente até à situação em que se está presente em lista de espera mais de 4 meses. Por outro lado, como a coluna da BD tinha muitos registos *null* e se estavam a lidar com valores percentuais foram contabilizados também os valores nulos. Foi criado um DW com os dias de espera já calculados, resultantes da diferença das duas chaves, data de marcação e data de realização. Os resultados foram desenvolvidos e computados com o suporte do módulo CDE, com a construção de dois tipos de gráficos distintos (barras e circular) e com o desenvolvimento de uma tabela, tal como se pode observar na Figura 4.49.

Desconsiderando os valores nulos, é possível afirmar-se que os tempos de espera para a realização de uma consulta após esta se encontrar marcada apresentam valores positivos e razoáveis. 30.33% dos casos esperam menos de um mês para consultar um médico e 55.64% dos casos espera no máximo 3 meses. As medidas preventivas devem incidir sobretudo sobre os 8.43% dos casos em que têm que esperar mais de 4 meses para uma consulta médica.

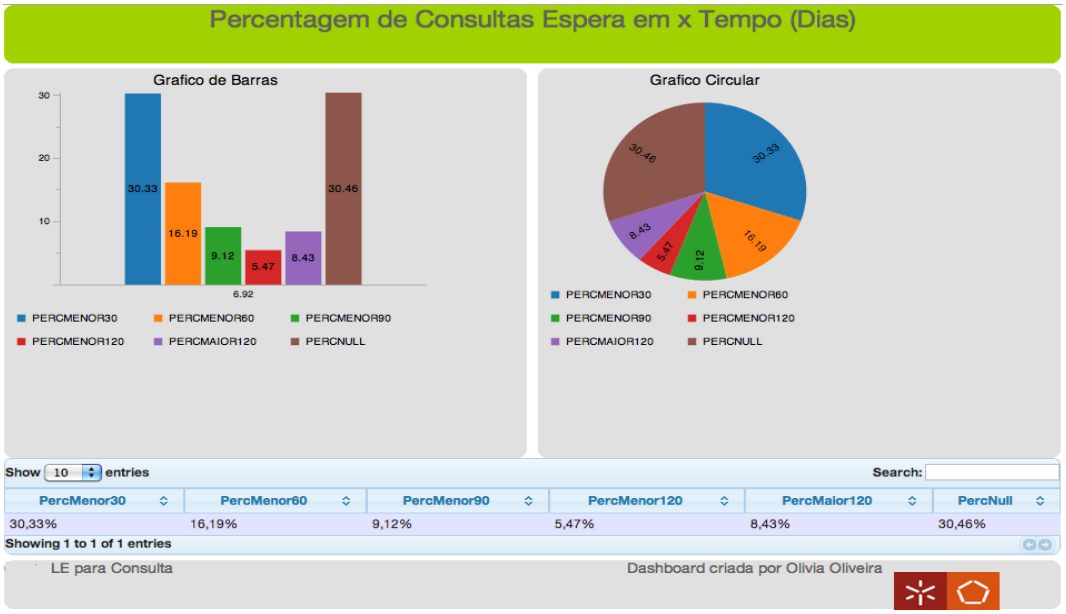


Figura 4.49: Apresentação das percentagens de tempos de espera por períodos mensais.

4.4.10 Dados estatísticos dos dias de espera para Bloco e Consultas

Considerou-se relevante para o projeto explorar os dados estatísticos referentes às listas de espera para cirurgia e consulta. Desta forma, construiu-se um DW apropriado ao problema utilizando a ferramenta PDI (Figura 4.50) para a realização do processo de ETL: extração dos dados da BD original, calcular a diferença (em dias) entre as datas finais e iniciais, filtrar e eliminar todos os resultados negativos os quais não contribuem para as estatísticas e, finalmente, ordenar o resultado final. Utilizaram-se, assim, vários tipos de *steps* como o *Table Input* (para extração de dados da BD), *Calculator* (para calcular a diferença entre data, em dias; este *step* contém várias operações matemáticas), *Filter Rows* (para filtrar linhas de um ficheiro, definindo uma condição de filtragem baseada em verdadeiro ou falso), *Sort Rows* (para ordenação das linhas de acordo com um campo) e *Table Output* e *Text File Output* (para saída e carregamento dos dados transformados). Utilizaram-se assim diferentes tipos de *steps*: de entrada de dados, de saída de dados, estatísticas, de transformação e de utilidades. Como já foi referido, existem inúmeros *steps* no PDI[3]. Após este processo, a computação estatística realizou-se também com o suporte do PDI que possui um conjunto de *steps* adequados para tal. O *step* utilizado foi o *Unvariate Statistics*, onde é possível a partir de uma entrada de dados computar valores estatísticos como a média, mediana, percentis, desvio-padrão, valores mínimo e máximo, e ainda o tamanho da amostra.

Na etapa de computação dos valores estatísticos calculou-se o tamanho da amostra (sample), a média (mean), o desvio-padrão (standard deviation), a mediana (median), os valores máximos e mínimos e o percentil-75. O resultado destes valores foram posteriormente organizados numa tabela de uma relatório, usando o PRD (Figuras 4.51 e 4.52).

De acordo com os resultados obtidos é possível retirarem-se algumas conclusões.

[3] http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+Steps

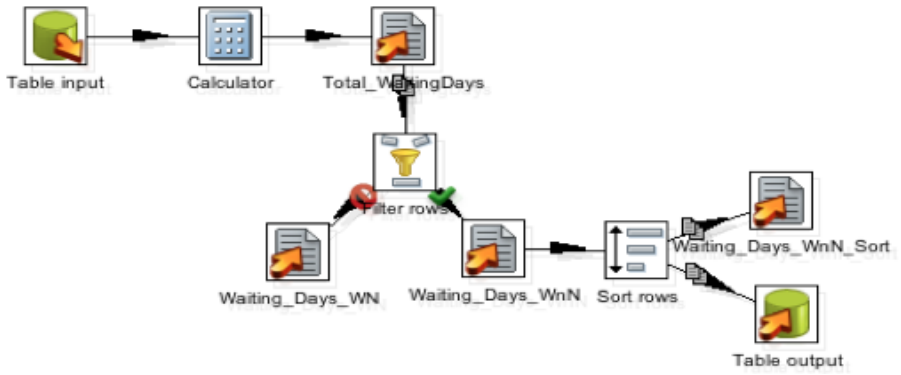


Figura 4.50: Construção do DW na ferramenta PDI.

N (sample)	Mean	StdDev	Min	Max	Median	Percentil-75
110187	88.5	121.68	0	3561	45	128

Figura 4.51: Estatísticas das listas de espera para cirurgia (valores em dias).

Observa-se um maior tempo de espera para cirurgia do que para consulta, sendo os valores médios considerados razoáveis dentro do contexto médico em situações comuns (em média, 3 meses de espera para se ser submetido a uma cirurgia) [52]. Por outro lado, ambos os valores mínimos são nulos, o que é bom mas não se pode considerar um situação ótima de acordo com os fundamentos teóricos subjacentes (secção 2.5). No que diz respeito aos valores máximos para cirurgia e consulta os valores apresentados são considerados maus. Pela observação dos dados, conclui-se que existem pacientes em espera aproximadamente 10 anos (3561 dias) para uma cirurgia e um tempo infinito e inaceitável para uma consulta, sendo que estes valores podem também representar os erros de registos já referidos e detetados noutros casos de estudo.

Em relação aos valores da medida de tendência central, a mediana, pode-se concluir que são valores considerados razoáveis sendo que 50% dos casos de tempo de espera para cirurgia são iguais ou inferiores a 45 dias e para consulta são iguais ou inferiores a 11 dias. Por outro lado, o desvio-padrão, medida de dispersão estatística, é bastante superior nos tempos de espera para consulta (1467.4) do que para cirurgia (121.68), fato também relacionado com as falhas de introdução de dados. Finalmente, os valores do percentil-75 significam que 75% dos casos em lista de espera para cirurgia correspondem no máximo a 128 dias de espera, enquanto que 75% dos casos em lista de espera para consulta correspondem no máximo a 29 dias de espera.

Em forma de conclusão, a computação dos valores estatísticos das listas de espera pode tornar-se em conhecimento útil a ser considerado pelo hospital, com a tomada de consciência da necessidade de implementação de algumas medidas de gestão adicionais, especialmente no caso das listagens para cirurgia, por exemplo com o estabelecimento de um valor máximo limite para o número de dias em espera para cirurgia (programa SIGIC).

Analisando os dias de espera para cirurgia com data de marcação a partir de 2004 (ano em que se implementou o programa SIGIC nas instituições de saúde em Portugal) repara-se que ainda existem alguns casos que o tempo de espera por cirurgia ultrapassa os 365 dias de espera, período de 1 ano. Assim, constata-se que para um total de

N (sample)	Mean	StdDev	Min	Max	Median	Percentil-75
433721	56.5	1467.4	0	365465	11	29

Figura 4.52: Estatísticas das listas de espera para consulta (valores em dias).

110141 casos em espera para cirurgia, 107303 (aproximadamente 97.4% do total da amostra) foram realizados dentro do limite máximo de espera de 1 ano, enquanto que 2838 (aproximadamente 2.6% do total da amostra) ultrapassaram o limite máximo de tempo de espera. Considera-se uma percentagem mínima e até pouco significativa, apresentando-se como possível explicação uma decisão puramente médica de acordo com o caso clínico.

4.5 Monitorização do Funcionamento do Bloco Operatório

O bom funcionamento do bloco operatório traduz-se em redução de custos para o hospital, na boa gestão e funcionamento das listas de espera para cirurgia e no aumento do desempenho de cada profissional. Atrasos no tempo de ocupação das salas do bloco pode traduzir-se num esforço extra por parte dos profissionais em horas extra ou na indisponibilidade da sala para situações de emergência. Por outro lado, cancelamentos inesperados de cirurgias são uma das principais causas para a rentabilização não ótima da utilização do bloco. Desta forma, tornou-se importante monitorizar o funcionamento do hospital em estudo, constituído por 4 unidades hospitalares sendo que no total existem 15 salas operatórias.

4.5.1 Monitorização e Controlo do Funcionamento do Bloco Operatório

Considerou-se importante fazer uma monitorização e controlo do funcionamento do bloco operatório. Para tal, foi criado um *dashboard* com o suporte do módulo Pentaho CDE de modo a tornar gráfica a visualização da informação. Estabeleceu-se a conexão à BD e através de *queries* SQL gerou-se um conjunto de dados relevantes para a análise do problema apresentado. Foram selecionadas a data prevista de realização da cirurgia e a data real de operação. Assim sendo, foram apresentados os resultados em dois tipos distintos de gráficos (barras e circular), tal como se pode verificar na Figura 4.53. As cores dos gráficos são pré-definidas pelo Pentaho, no entanto utilizou-se a opção 'Colors', tornando possível configurar os gráficos de acordo com a tonalidade de cores pretendida. As cores podem ser definidas através de HTML ou então através do seu nome em inglês porém esta última alternativa restringe-se apenas às cores principais. No *dashboard* desenvolvido foram escolhidas as cores azul marinho (#0D1FE0) e rosa claro (#5C595C). Os resultados estão traduzidos segundo uma amostra total de 110150 casos em que as datas de operação e previstas coincidem ou não coincidem, sendo assim possível avaliar a eficiência dos serviços no bloco.

Pela observação da Figura 4.53, pode-se constatar que há uma disparidade de resultados favoráveis em relação aos menos favoráveis no sentido em que o número de casos

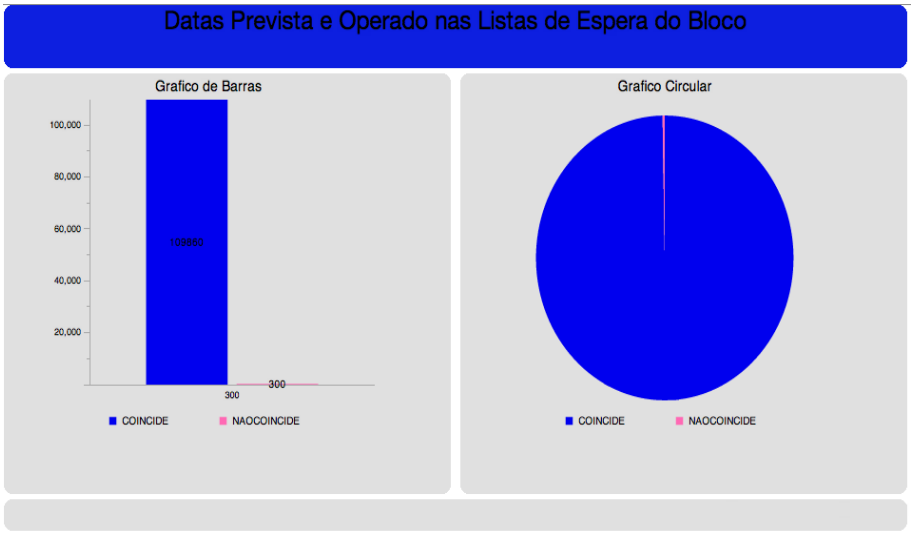


Figura 4.53: Total de casos em lista de espera para o bloco operatório.

em que as cirurgias ocorreram na data prevista é bastante superior (109850 casos) ao número de casos não coincidentes (300 casos). Através de uma análise mais particular dos resultados não coincidentes pode-se perceber que existem mais casos em que a data de operação é posterior à data prevista (222 dos 300 casos), enquanto que a operação apenas foi realizada antes da data prevista em 78 dos 300 casos não coincidentes. Por outro lado, esta discordância de datas apresenta-se como pouco preocupante para a gestão do funcionamento do bloco uma vez que na maioria das situações as datas não coincidem por uma questão de poucos dias.

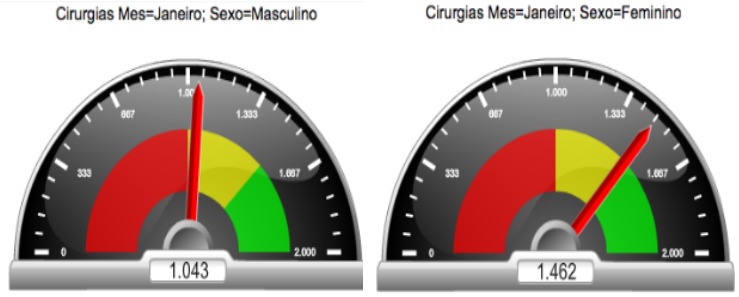
Em forma de conclusão, foi possível realizar-se uma avaliação positiva da eficiência e funcionamento do bloco operatório do hospital, uma vez que em 99.73% dos casos as cirurgias são realizadas no tempo previsto e que apenas uma minoria pouco significativa de 0.27% dos casos não são coincidentes.

4.5.2 Cirurgias Realizadas - Análise Preditiva por Mês e Sexo

As cirurgias realizadas num hospital devem ter em conta o sexo dos pacientes, uma vez que este fator está diretamente dependente da capacidade do hospital por enfermarias para internamento. Assim, pretendeu-se realizar uma análise de forma a encontrar um padrão do número de pessoas operadas, divididas pelo seu género. Desta forma, no caso de existir uma tendência para determinado sexo a instituição de saúde deve então preparar mais enfermarias em função dos resultados. No sentido da análise do problema, foi também efetuada uma análise preditiva por meses, sendo que a partir da informação referente aos meses de Janeiro e Fevereiro seria possível prever-se uma tendência para o mês de Março.

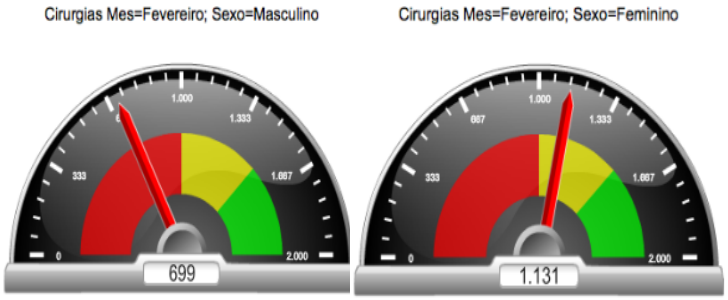
Para o estudo realizado, construiu-se um DW através do PDI de forma a serem selecionados apenas os dados relevantes para a análise (através de *queries* SQL) e a inserção dos mesmos numa nova tabela de forma a ficarem disponíveis todos os dados de modo organizado. Posteriormente, foi desenvolvido um conjunto de *dashboards* com o suporte do Pentaho EE, utilizando gráficos do tipo Dial, Circular e de Barras Verticais. Inicialmente, apresentam-se os dados relativos a Janeiro (Figuras 4.54a e 4.54b) e

Fevereiro (Figuras 4.55a e 4.55b) de acordo com o género das pessoas submetidas a intervenções cirúrgicas nos respetivos meses.



(a) Sexo Masculino (b) Sexo Feminino

Figura 4.54: Gráfico do tipo Dial que representa o número de pessoas por sexo submetidas a cirurgia no mês de Janeiro.



(a) Sexo Masculino (b) Sexo Feminino

Figura 4.55: Gráfico do tipo Dial que representa o número de pessoas por sexo submetidas a cirurgia no mês de Fevereiro.

Da observação dos gráficos é possível concluir-se que o sexo feminino está largamente mais representado do que o sexo masculino. No mês de Janeiro foram operadas mais mulheres (1462 casos) do que homens (1043 casos), representando 58.4% das mulheres em comparação com 41.6% dos homens num total de 2505 casos. No mês de Fevereiro também foram operadas mais mulheres (1131 casos) do que homens (699 casos), representando 61.8% das mulheres em comparação com 38.2% dos homens num total de 1830 casos. Assim, pode-se afirmar que o estabelecimento de saúde em estudo deve ter preparadas um maior número de enfermarias femininas do que masculinas, deslocando também um maior número de profissionais para esses pisos.

De forma a prever e poder agir antecipadamente na preparação das enfermarias para os próximos tempos, tentou-se prever o número de pessoas por género que seriam submetidas a uma intervenção cirúrgica no mês seguinte, Março. No contexto das observações fatuais retiradas, é possível prever-se que no mês de Março o sexo feminino continuará a ser o mais representado, seguindo assim a forte tendência dos meses anteriores.

Para comprovar a veracidade das previsões efetuadas, foram desenvolvidos 2 *dashboards*, um composto por gráficos circulares cada um referente a um mês, e outro composto por gráfico de barras verticais onde se podem observar todos os resultados

dispostos de forma organizada, divididos por mês e por sexo. Apresentam-se, assim, os dados reais relativos a cada um dos meses Janeiro, Fevereiro e Março (Figuras 4.56a, 4.56b e 4.56c), realçando-se o gráfico circular do mês de Março para comprovar ou não a previsão efetuada. Seguidamente, apresentam-se os dados relativos a todos os meses, e por sexo, organizados todos num só gráfico de barras (Figura 4.57).

(a) Janeiro (b) Fevereiro (c) Março

Figura 4.56: Gráficos circulares onde se representam o número de casos cirúrgicos por sexo (1- sexo masculino e 2- sexo feminino) e por mês.

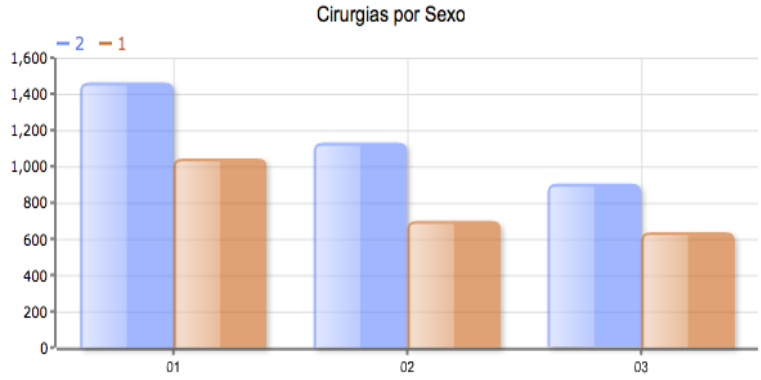


Figura 4.57: Gráfico de barras verticais com o conjunto total de dados apresentado e organizado por meses e por género.

A construção de *dashboards* no Pentaho EE é simples e mais intuitiva do que na plataforma CE. A grande parte do desenvolvimento é automático, ao contrário do CE em que se tem que definir tudo. Por este motivo, o Pentaho EE e os seus módulos estão mais direcionados para utilizadores mais básicos. Apresenta simplicidade e atratividade dos menus e do *layout*, bem como qualidade gráfica da interface e das soluções. Estas são consideradas atrativas e bastante amigáveis do utilizador, apresentando um aspeto muito profissional e permitindo facilidade na leitura dos dados e na interatividade das próprias soluções. Existe reversibilidade das ações e o tempo de execução é muito baixo (poucos segundos). Finalmente, não existe praticamente documentação nem suporte técnico, tal como se verifica para o Pentaho EE de um modo global.

4.5.3 Ocupação do Bloco Operatório

A utilização das salas do bloco operatório significa uma grande porção dos custos de um estabelecimento de saúde. Neste sentido, é importante que exista uma gestão e administração adequadas da utilização das salas operatórias de modo a reduzir o máximo possível os custos associados.

A partir dos dados clínicos do hospital referentes às cirurgias realizadas nos últimos 5 anos (2007-2012), utilizaram-se os tempos de ocupação de sala real e médio. Estes tempos encontram-se armazenados em segundos e o tempo médio de cirurgia baseia-se nos casos dos últimos 3 anos, contendo todos os casos que se encontram entre os percentis 25 e 75. Após a fase de construção do DW, contendo apenas os tempos que interessavam relativos à ocupação das salas e o tempo médio, agrupados por procedimentos e respetivas especialidades, procedeu-se à exploração da informação na tentativa de encontrar conclusões efetivas para a tomada de decisão e implementação de medidas. Para tal, foi construído um conjunto de *dashboards* com o suporte do módulo CDE.

Tal como se pode verificar na Figura 4.58, os gráficos eram do tipo de barras e tinham a opção de empilhados ativa para que fosse possível visualizar os dois tempos numa mesma coluna. Cada coluna representa um procedimento diferente, sendo que cada conjunto de procedimentos pertence a uma determinada especialidade médica. Tal como é possível observar, foi criado um parâmetro para selecionar a especialidade em análise estando, assim, cada *dashboard* associado a uma especialidade médica.

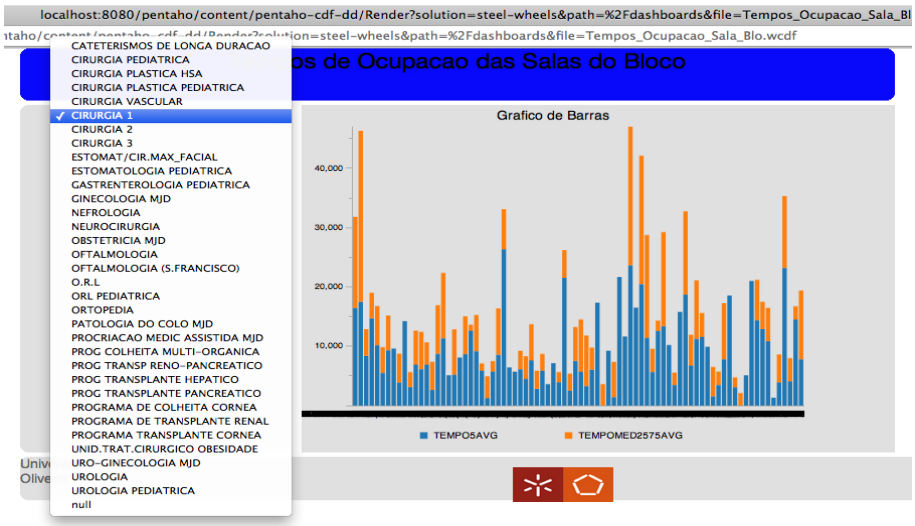


Figura 4.58: Dashboard criado para análise da ocupação das salas operatórias em função da especialidade e dos procedimentos.

A partir da visualização das Figuras 4.59 e 4.60, podem-se retirar conclusões relativas à especialidade de Oftalmologia. Neste caso particular, foi fixada a coluna correspondente ao procedimento Vitrectomia Mecânica NCOP e, por isso, podem-se retirar algumas conclusões sobre os tempos de ocupação relativos a este procedimento. Através da análise das figuras afirma-se que para o procedimento fixado, os tempos de ocupação da sala são viáveis para a instituição apesar de estarem muito próximos, sendo o tempo real (4180.7 s) ligeiramente inferior ao tempo médio (4286.6 s) de ocupação. De um modo global, pode-se concluir que a especialidade de Oftalmologia tem

uma utilização pouco positiva e pouco rentável das salas operatórias, uma vez que os tempos reais de utilização são na maioria superiores aos tempos médios (parte azul das colunas realça-se mais que a parte laranja).

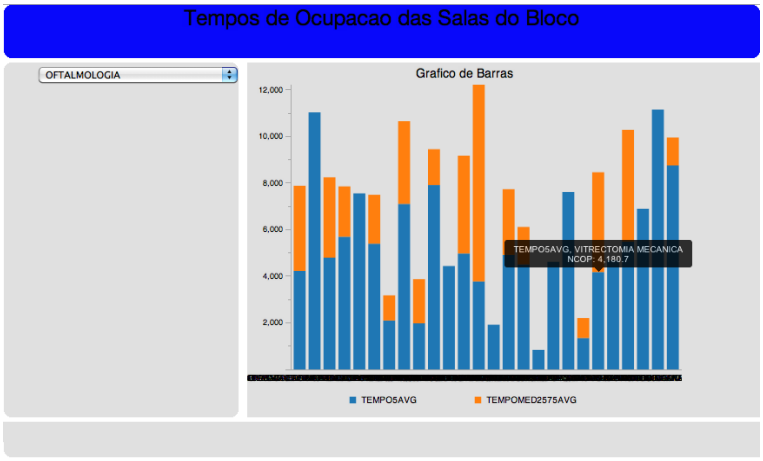


Figura 4.59: Gráfico de barras verticais empilhado com os tempos de ocupação da sala para a especialidade Oftalmologia, com fixação do tempo real de ocupação.

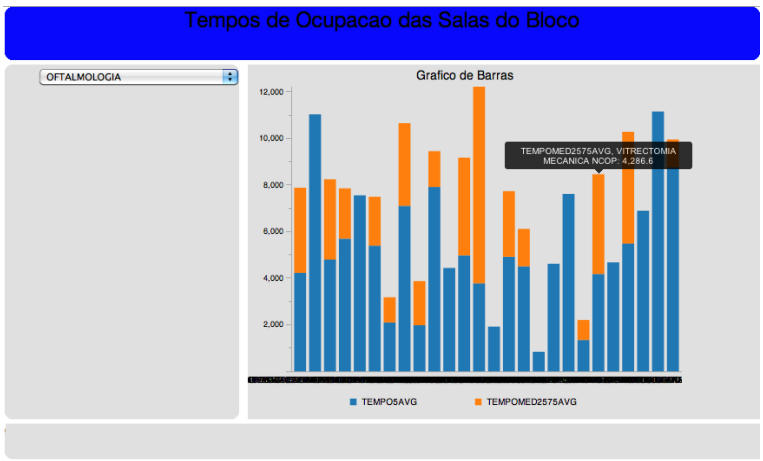


Figura 4.60: Gráfico de barras verticais empilhado com os tempos de ocupação da sala para a especialidade Oftalmologia, com fixação do tempo médio de ocupação.

Tal como é possível verificar-se na Figura 4.58, 4.59 e 4.60, os gráficos gerados pelo Pentaho CE contêm muitos valores em análise (correspondentes aos procedimentos cirúrgicos para cada especialidade) pelo que a visualização dos valores do eixo das abcissas se torna impercetível uma vez que se sobrepõem (barra preta visível). Deste modo, para se tomar conhecimento da relação de cada barra com o respetivo procedimento tem que se colocar o cursor do rato em cima de cada barra e visualizar a informação de forma interativa através do quadro que vai surgindo com a informação referente a cada situação.

Em relação à análise seguinte, estão apresentadas de seguida as Figuras 4.61 e 4.62 onde foi selecionada a especialidade Procriação Médica Assistida MJD. Esta especialidade tem apenas 3 procedimentos possíveis. Analisando em particular o procedimento Aspiração de Ovário, é possível verificar-se que o tempo real de utilização (2357.6 s)

é superior ao tempo médio de utilização (1753 s), o que significa que está a existir uma sobreocupação das salas operatórias no que diz respeito às últimas intervenções cirúrgicas deste procedimento da especialidade médica em questão.

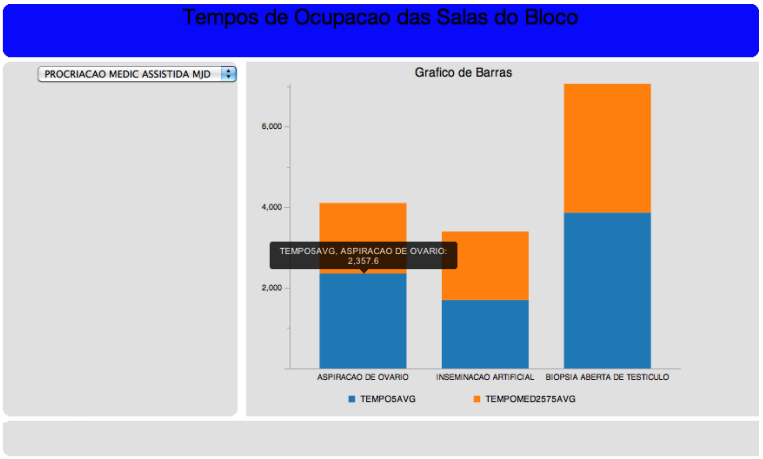


Figura 4.61: Gráfico de barras verticais empilhado com os tempos de ocupação da sala para a especialidade Procriação Médica Assistida MJD, com fixação do tempo real de ocupação.

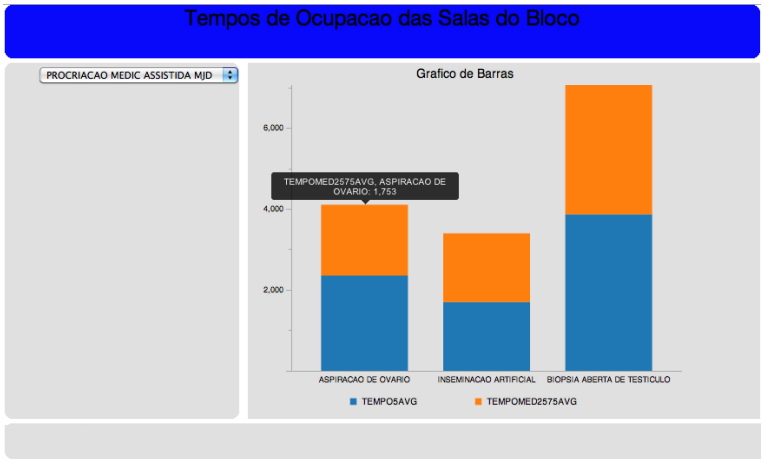


Figura 4.62: Gráfico de barras verticais empilhado com os tempos de ocupação da sala para a especialidade Procriação Médica Assistida MJD, com fixação do tempo médio de ocupação.

Como conclusão final, pode-se afirmar que existem algumas medidas de prevenção possíveis de serem implementadas no seio da instituição. Falando particularmente do caso de estudo, pode-se ter em atenção a especialidade de Oftalmologia e tentar perceber o motivo das últimas intervenções cirúrgicas estarem a ocupar o bloco durante um maior período de tempo que as intervenções anteriores. Talvez reunir com o especialistas da área e alertar para o fato constatado seria uma atitude estimuladora da análise e repensamento do desempenho. Em analogia a este caso de estudo, era importante realizar uma análise pormenorizada para cada uma das especialidades, de forma a agir naquelas que estão a sobrecarregar a ocupação das salas operatórias e, que por conseguinte, estão a traduzir um aumento de custos associados para o hospital.

Os gráficos desenvolvidos são considerados apelativos e amigáveis ao utilizador. Por outro lado, as cores são atrativas e permitem a fácil distinção entre os diferentes tempos uma vez que se está a utilizar um tipo de gráfico empilhado. A construção de gráficos empilhados no CE é simples, necessitando para isso de colocar apenas a opção de empilhado como verdadeira. A utilização de parâmetros também torna os resultados mais atrativos sem haver a necessidade de se gerarem muitos *dashboards*. Assim, e apenas com um *dashboard* pode-se selecionar a especialidade médica que se pretende visualizar e analisar os dados.

4.5.4 Análise OLAP da Utilização das Salas Operatórias

Este caso de estudo tem como principal objetivo estudar a utilização de cada uma das salas em diferentes meses (Janeiro, Fevereiro e Março) do ano 2012. A tabela que contém a informação relativa às cirurgias contém muito informação e armazena milhares de dados, tendo-se considerado pertinente proceder à construção de um DW para organizar apenas os dados relevantes numa nova tabela da BD. Para o processo de ETL foi utilizada a ferramenta PDI. A extração inicial dos dados foi realizada através de uma consulta por SQL à tabela original onde foram selecionadas as colunas correspondentes à especialidade, ao procedimento, ao tempo de ocupação da sala em cada caso e ao tempo médio de utilização da sala para cada procedimento e especialidade. Posteriormente, efetuou-se uma transformação dos dados, no sentido em que se calculou o valor médio do tempo de ocupação da sala operatória por procedimento e por especialidade. Finalmente, procedeu-se ao carregamento da informação integrada e organizada para um novo esquema e uma nova tabela da BD. O esquema gráfico utilizado no PDI apresenta-se na Figura 4.63.

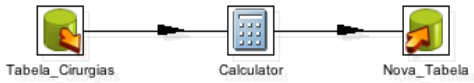


Figura 4.63: Construção do DW através do módulo PDI.

A análise dimensional foi efetuada com o suporte do Pentaho Analysis. Inicialmente, definiram-se as medidas e as dimensões de análise através do navegador OLAP (Figura 4.64). Considerou-se que as salas operatórias e os meses eram dimensões que constituam as linhas e o valor total de utilização de cada sala era a medida de análise. Neste caso de estudo não foi utilizado qualquer filtro uma vez que a informação foi diretamente extraída através de uma *query* SQL.

Consequentemente, é construída uma tabela de forma automática onde está organizada a informação definida no cubo (Figura 4.65). Para a construção da tabela utilizou-se a função *Drill Replace*, que permite a substituição de um membro pelos seus sub-membros, detalhando assim todas as salas operatórias existentes.

A *query* MDX resultante da informação consultada encontra-se transcrita na Figura 4.66.

A partir da *query* definida para a consulta dos dados, construiu-se de forma automática um gráfico, tal como se pode observar na Figura 4.67.

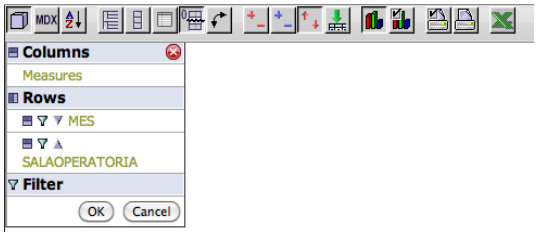


Figura 4.64: Navegador OLAP para a definição e estruturação do cubo multidimensional.

		Measures
MES	SALAOPERATORIA	X
All MESs	SALA D B.ED.CLASSICO	18
	SALA-A B.ED.CLASSICO	162
	SALA-A BLOCO CENTRAL	355
	SALA-B B.ED.CLASSICO	312
	SALA-B BLOCO CENTRAL	250
	SALA-C B.ED.CLASSICO	72
	SALA-C BLOCO CENTRAL	352
	SALA-D BLOCO CENTRAL	261
	SALA-E BLOCO CENTRAL	417
	SALA-F BLOCO CENTRAL	567
	SALA-G BLOCO CENTRAL	302
	SALA-H BLOCO CENTRAL	219
	SALA-I BLOCO CENTRAL	178
	SALA-N1 BL NEUROCIRU	156
	SALA-N2 BL NEUROCIRU	182

Figura 4.65: Tabela gerada de forma automática a partir da definição do cubo OLAP.

A Figura 4.68 traduz as propriedades do gráfico construído. Foi utilizado o tipo de gráfico de barras vertical com opção 3D, definiu-se a fonte da letra, as legendas dos eixos e as dimensões do gráfico.

Seguidamente, utilizou-se a função *Swap Axes* do Pentaho Analysis (Figura 4.69), que permite a articulação dos dados permutando as linhas e as colunas.

Consequentemente, foi gerado um novo gráfico representado na Figura 4.70.

Finalmente, utilizou-se a função *Drill Through* (Figura 4.71), de forma a se visualizarem os dados detalhados das células selecionadas.

Em resultado da utilização da função de *Drill Through* foi possível a visualização detalhada da informação da sala operatória Sala-A Bloco Central (Figura 4.72) e da Sala-A B. Ed. Classico (Figura 4.73), com o total de vezes que foram utilizadas as salas em cada mês.

A realização da análise OLAP é vantajosa no sentido em que o uso de gráficos e cubos OLAP produz um maior dinamismo e interatividade com a manipulação dos dados. A partir da análise realizada é possível retiram-se conclusões globais que se possam traduzir em conhecimento útil para a tomada de decisões na instituição de saúde. Foi possível analisarem-se as salas operatórias mais utilizadas ao longo dos 3

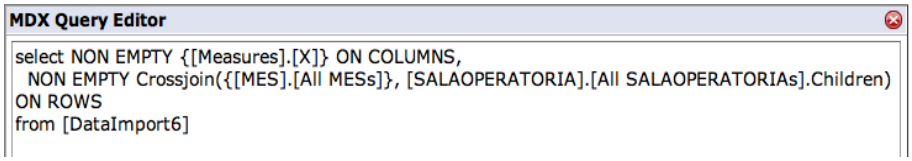


Figura 4.66: Visualização da query MDX utilizada para o caso de estudo.

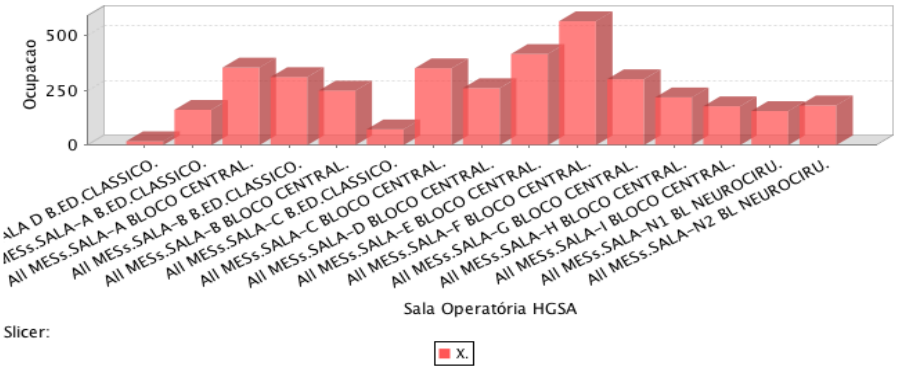


Figura 4.67: Visualização do gráfico criado pelo Pentaho Analysis.

meses (Sala-F Bloco Central; Sala-E Bloco Central; Sala-A Bloco Central, por ordem decrescente) e a sala menos utilizada no mesmo período de tempo (Sala D B. Ed. Classico). Este conhecimento é importante para a equipa administrativa no sentido em que as salas mais utilizadas devem ser submetidas a uma inspeção e manutenção mais frequentes, e os seus equipamentos devem ser permanentemente atualizados. Por outro lado, fazendo uma análise mais detalhada dos dados pode-se perceber em que mês é que determinadas salas são mais ou menos utilizadas bem como a sua utilização global, e daí tentar desviar intervenções cirúrgicas de salas sobrecarregadas para salas com menor taxa de ocupação.

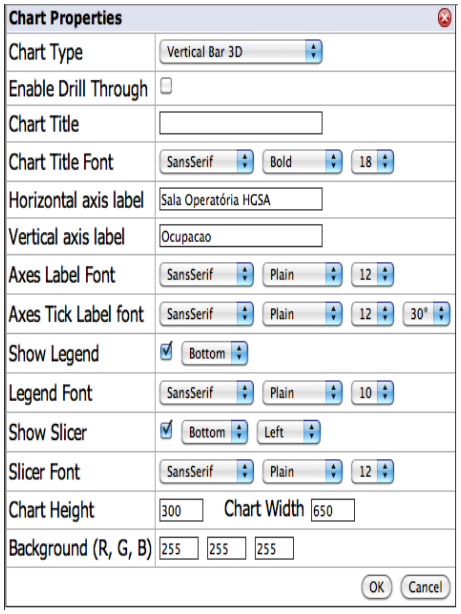


Figura 4.68: Janela de definição das propriedades dos gráficos.

Figura 4.69: Barra de ferramentas do Pentaho Analysis com o botão de Swap Axes realçado.

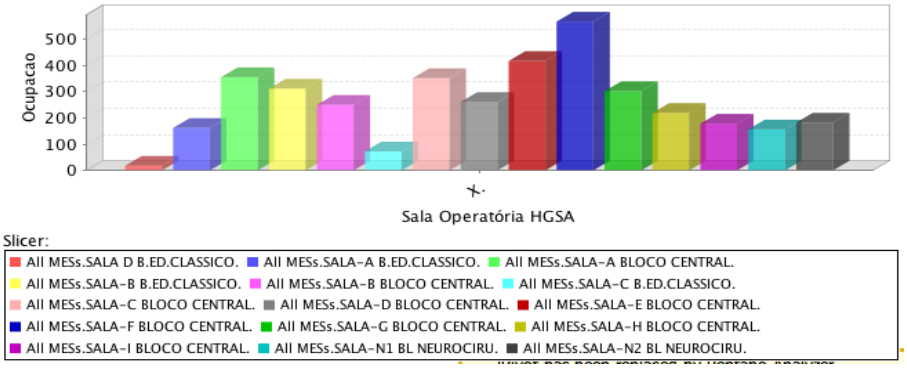


Figura 4.70: Visualização do gráfico resultante da utilização da função *Swap Axes*.

Figura 4.71: Barra de ferramentas do Pentaho Analysis com o botão de *Drill Through* realçado.

Drill Through Table for X

MES	SALAOPERATORIA	X
01	SALA-A BLOCO CENTRAL	151,00
02	SALA-A BLOCO CENTRAL	99,00
03	SALA-A BLOCO CENTRAL	105,00

Page 1/1 Rows/page 10

Figura 4.72: Visualização detalhada dos dados relativos à Sala-A Bloco Central através da utilização da função de *Drill Through*.

Drill Through Table for X

MES	SALAOPERATORIA	X
01	SALA-A B.ED.CLASSICO	53,00
02	SALA-A B.ED.CLASSICO	63,00
03	SALA-A B.ED.CLASSICO	46,00

Page 1/1 Rows/page 10

Figura 4.73: Visualização detalhada dos dados relativos à Sala-A B. Ed. Clássico através da utilização da função de *Drill Through*.

4.5.5 Ocupação do Bloco Operatório – Análise Preditiva

Os dados relativos às intervenções cirúrgicas realizadas no hospital em estudo foram organizados por salas operatórias e meses de realização, com o intuito principal de se conseguir efetuar uma análise preditiva dos dados. Neste âmbito, os dados são referentes aos 3 primeiros meses do ano de 2012 (Janeiro, Fevereiro e Março), sendo que através da análise dos dados relativos a Janeiro e Fevereiro se tentou encontrar tendências para o mês de Março. Para tal, foi construído um DW apenas com a informação transformada e selecionada de acordo com o problema, procedendo-se posteriormente ao desenvolvimento de *dashboards* com o suporte do Pentaho EE. A construção do DW realizou-se através da utilização de *queries* SQL para consulta e seleção dos dados relevantes, utilizando-se de seguida o PDI para criação e inserção de um nova tabela com os dados limpos e organizados, estando desta forma preparados para a análise. Por sua vez, os *dashboards* foram construídos de acordo com o esquema de dados criado e baseiam-se em gráficos de barras verticais (estilo *open flash chart*), o tipo de gráfico considerado mais adequado para a análise de uma grande quantidade de dados, como é possível verificar para este problema. Consequentemente, foram então desenvolvidos os gráficos referentes à ocupação das diversas salas operatórias nos meses de Janeiro (Figura 4.74) e Fevereiro (Figura 4.75).

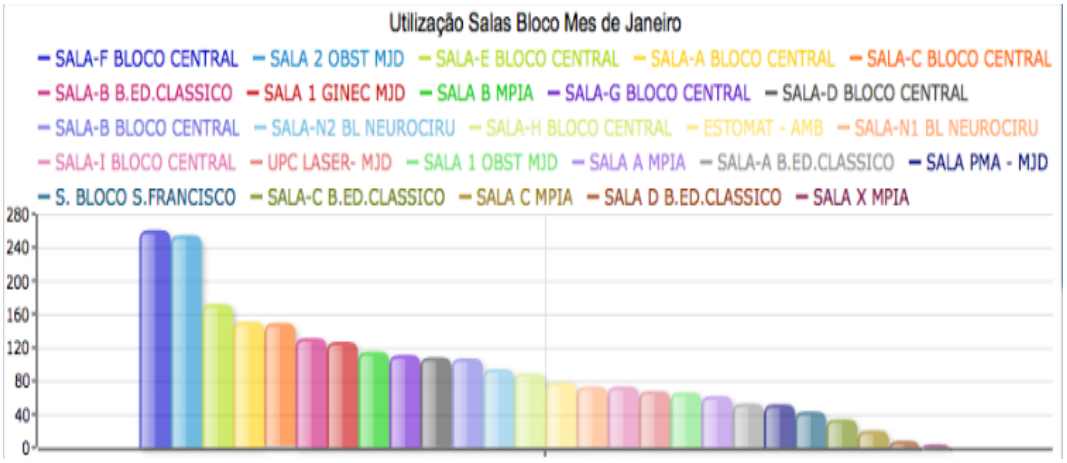


Figura 4.74: *Dashboard* desenvolvido com o suporte do Pentaho EE, onde se apresenta a ocupação do bloco operatório em Janeiro de 2012.

Pela observação dos gráficos, repara-se que as salas operatórias mais utilizadas no mês de Janeiro foram, por ordem decrescente, a Sala-F do Bloco Central, a Sala 2 de Obstetrícia da MJD (Maternidade Júlio Dinis) e a Sala-E do Bloco Central. As duas salas menos utilizadas foram, por ordem decrescente, a Sala D do B. Ed. Clássico e a Sala X do Maria Pia. Em relação à informação referente ao mês de Fevereiro, repara-se que as salas operatórias mais utilizadas foram, por ordem decrescente, Sala 2 de Obstetrícia da MJD, Sala-F do Bloco Central e a Sala-E do Bloco Central. Por sua vez, as salas menos utilizadas foram, por ordem decrescente, a Sala D do B. Ed. Clássico e a Sala X do Maria Pia. Através da análise desta informação é possível prever que as salas operatórias referidas como as mais utilizadas em princípio serão as que apresentam também maior taxa de ocupação no mês de Março, seguindo-se também o mesmo raciocínio para as salas menos utilizadas. A informação real relativa

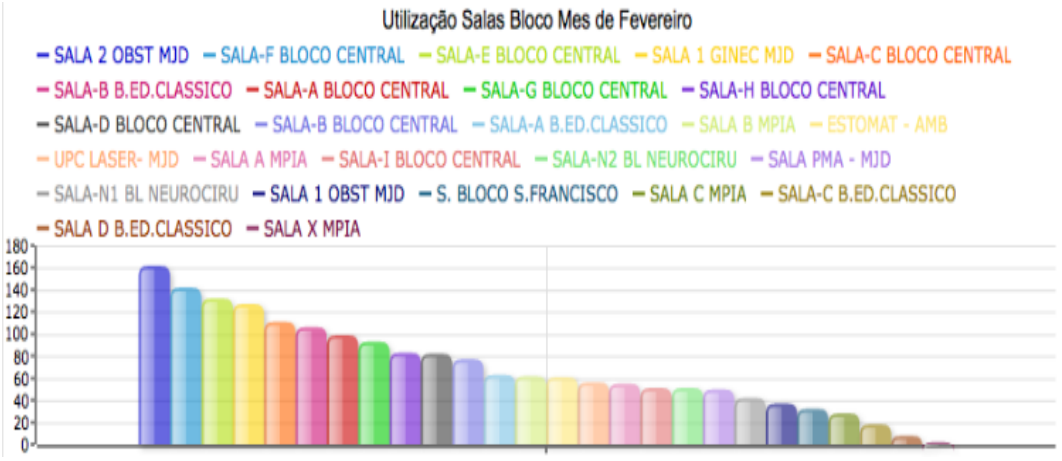


Figura 4.75: *Dashboard* desenvolvido com o suporte do Pentaho EE, onde se apresenta a ocupação do bloco operatório em Fevereiro de 2012.

à ocupação do bloco no mês de Março está apresentada na Figura 4.76, onde se podem então confrontar os dados reais com os que foram inicialmente previstos.

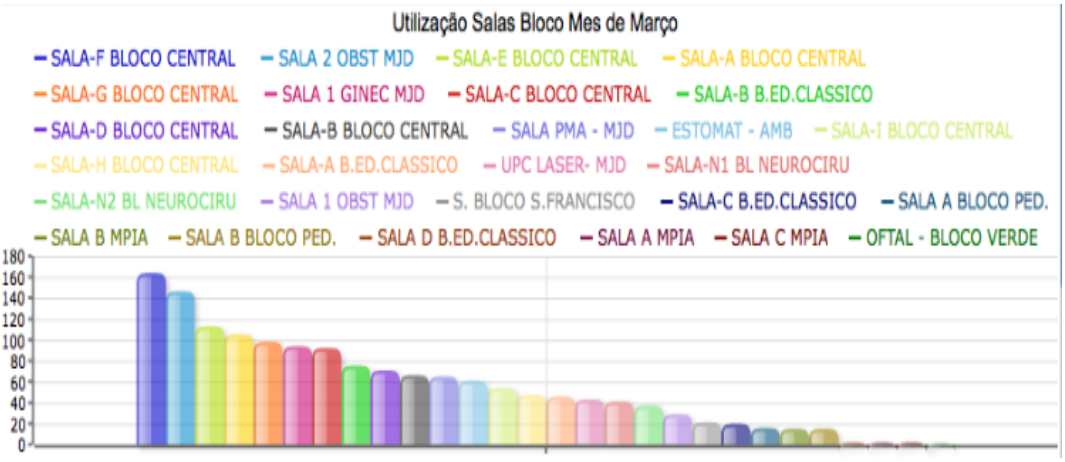


Figura 4.76: *Dashboard* desenvolvido com o suporte do Pentaho EE, onde se apresenta a ocupação do bloco operatório em Março de 2012.

Da observação do *dashboard*, é possível verificar-se que as salas mais utilizadas do bloco operatório mantêm-se as mesmas (Sala-F do Bloco Central, Sala 2 de Obstetrícia da MJD e a Sala-E do Bloco Central) verificando-se que a análise preditiva foi realizada com sucesso neste aspeto. Contudo, em relação às salas menos utilizadas já não se verificou uma continuidade de menor utilização das registadas em Janeiro e Fevereiro. Esta análise era baseada em valores muito pequenos o que faz também com que haja uma grande variabilidade da ordenação na lista destas salas que ocupam os últimos lugares em relação à taxa de ocupação.

4.5.6 Utilização do Bloco Operatório – Análise DM

A análise realizada baseou-se nas técnicas de segmentação (*clustering*) e de classificação. Os dados foram computados segundo os algoritmos de *clustering* EM, SimpleKMeans e FarthestFirst. Os dados utilizados para análise foram: o dia e o mês

de cirurgia, a sala operatória, o tempo de sala ocupado e o início da intervenção. A extração dos dados e a consequente criação do ficheiro arff foi efetuada através do PDI. Os valores de tempo de ocupação da sala e de início de intervenção encontram-se em horas. Os resultados completos da análise encontram-se no anexo A.14.

O algoritmo de EM obteve como resultado 8 *clusters* sendo que o que obteve maior percentagem de instâncias agrupadas (23%) apresenta as seguintes caraterísticas: realização das cirurgias no mês de Janeiro, maioritariamente no dia 11, com a utilização frequente das Sala-A do Bloco Central e Sala-F do Bloco Central, com uma média de ocupação de sala de 2.08 e desvio-padrão 0.8, e com o início da intervenção às 11.6 (valor médio). Em resultado da computação do algoritmo SimpleKMeans foram definidos 2 *clusters* com as seguintes caraterísticas. *Cluster* 1(563 instâncias – 57%): mês de Janeiro, dia 16, Sala 2 de Obstetrícia da MJD, tempo de ocupação de 1.7 (média) e início das intervenções às 12.4 (média); *Cluster* 2(432 instâncias – 43%): mês de Fevereiro, dia 2, Sala 2 de Obstetrícia da MJD, tempo de ocupação de 1.8 (média) e início das intervenções às 13 (média). Por fim, o algoritmo FarthestFirst definiu como *cluster* central com 594 instâncias agrupadas (60%) o seguinte: mês de Janeiro, dia 10, Sala-C do Bloco Central, ocupação de 1.3 (média) e início da intervenção às 9.9 (média). O *cluster* considerado mais distante com 401 instâncias (40%) apresenta as seguintes caraterísticas: mês de Fevereiro, dia 29, Sala-N1 do Bloco de Neurocirurgia, tempo de ocupação de 10.8 (média) e início às 8.5 (média).

Em termos de técnica de classificação, o algoritmo utilizado foi o J48. Este algoritmo é uma implementação Weka do algoritmo C4.5 e gera árvores de decisão. Em cada nó da árvore, o C4.5 escolhe um atributo dos dados que divida de forma mais eficaz o conjunto de amostra em subconjuntos de uma classe ou de outra. Assim, os nós iniciais são os dias do mês, depois divide-se por mês e depois dependendo do tempo de ocupação ou do início da intervenção classifica-se uma sala operatória. A árvore tem 308 folhas e o seu tamanho é de 586. Aproximadamente 55.7% das instâncias foram bem classificadas num total de 995, e obteve-se uma valor de estatística Kappa de 0.5299. Os valores seguintes do modelo são: erro absoluto médio, erro quadrático médio, erro absoluto relativo e erro quadrático relativo. O erro absoluto médio é 0.04, o erro quadrático médio é 0.1414, o erro absoluto relativo é 55% e o erro quadrático relativo é 71.4%.

4.6 Apreciação Global da Ferramenta Pentaho BI

As principais conclusões acerca da utilização e exploração da ferramenta de BI Pentaho encontram-se sintetizadas de seguida nas Tabelas 4.1, 4.2, 4.3, 4.4, 4.5 e 4.6.

Plataforma BI	
Community Edition	• Incorporável e composta por vários componentes; • Não se trata de um sistema único, mas vários integrados; • Baseada em tecnologias livres (J2EE, JDBC, JNDI, JSR-168, JSP, HTML, XSL e XML); • Portabilidade, escalabilidade e integração; • Permite ligações a diversas fontes de dados; • Definição total e pormenorizada de todo o desenvolvimento; • Tempo de processamento ligeiramente superior ao da EE no desenvolvimento das soluções; • Direcionada para pessoal mais técnico; • Irreversibilidade de ações; • Possui uma ferramenta de administração; • Não há necessidade de um *Data Warehouse*, mas pode-se utilizar.
Enterprise Edition	• Permite ligações a diversas fontes de dados; • Filosofia *drag-and-drop*; • Construção de soluções de forma bastante intuitiva e interativa; • Direcionada para o utilizador final; • Interação dinâmica com o utilizador; • Soluções com aspeto profissional; • Licença de custo reduzido.

Tabela 4.1: Conclusões gerais da Plataforma de BI.

Pentaho Reporting	
Interative Report - EE	• Simples e intuitivo; • Filosofia *drag-and-drop*; • Funções de paginação e data automáticas;
Pentaho Report Designer	• Criação flexível de relatórios; • Mais complexo e pouco intuitivo; • Ambiente gráfico atrativo para o desenvolvimento e projeto de relatórios; • Ambiente de desenvolvimento não traduz o resultado final (capacidade de ocultação dos objetos do relatório com necessidade de pré-visualização); • Possibilidade de construção de gráficos e de importação de imagens (disponibilização de mais de 15 tipos de gráficos diferentes); • Acesso a dados relacionais, OLAP e XML; • Formatos de exportação: XLS / PDF / DOCX / CSV / RTF / XML / HTML; • Totalmente dependente do utilizador; • Capacidade de incorporar HTML e JavaScript para controlos dinâmicos e interativos de relatórios; • Cruzamento da plataforma (cliente e servidor) em Mac, Linux/Unix e Windows; • Portabilidade, escalabilidade e integração total com Java; • Possibilidade de publicação no servidor de BI.
Report Wizard	• Desenvolvimento rápido de relatórios; • Simples: seguimento de um conjunto de etapas; • Processo semi-automático; • Funções de paginação e data automáticas.

Tabela 4.2: Conclusões gerais do Pentaho Reporting.

Dashboarding

Módulo CDE	• Opções flexíveis de desenvolvimento; • Permite ligações com inúmeras fontes de dados (MDX,SQL,MQL,...); • Definição do *layout* totalmente dependente do utilizador; • Grande variedade de componentes visuais: diversos tipos de gráficos, tabelas, texto, seleção, parâmetros,...; • Integração com PRD e Analysis; • Interatividade com o utilizador na própria solução; • Irreversibilidade de ações; • Fecho de uma janela de edição automático (não aparece uma mensagem a perguntar se quer gravar o trabalho realizado até ao momento, perdendo-se assim num simples clique); • O *Span Size* (representa a largura do painel na definição do *layout*) tem um valor máximo de 24, sendo que caso o objetivo passe pelo desenvolvimento de um *dashboard* de maiores dimensões não se pode aplicar nenhum *template* pré-definido; • Simplicidade e qualidade gráfica; • Pouca documentação técnica de suporte ao desenvolvimento.
Dashboards - EE	• Desenvolvimento de *dashboards* mais intuitivo e praticamente automático; • Soluções mais atrativas e profissionais; • Documentação teórica e de suporte escassa ou praticamente nula.

Tabela 4.3: Conclusões gerais do Dashboarding.

Integração de Dados

Pentaho Data Integration	• Ferramenta ETL com aproximação a modelos e metadados; • Acesso a diversos tipos de dados /*inputs*; • Desempenho e escalabilidade empresarial; • Integração de dados e programação de transformações (capacidade de agendamento); • Inúmeras funções de transformação de dados (junção, estatística, filtragem, ordenação,...); • Grande variedade de saídas /*outputs*; • Possibilidade de transformações complexas sem uso de código; • Integração com PRD; • Necessidade de conhecimentos teóricos por parte do utilizador; • Bom ambiente gráfico; • Projeção totalmente gráfica das transformações e dos *jobs*; • Interface gráfica atrativa (cada *step* tem um ícone diferente para identificar a sua função); • Integração, relatórios interativos e análise de dados; • Sugere o que fazer e não como fazer; • Ferramenta 100% Java, com suporte multi-plataforma (permitindo que a ferramenta seja executada em diferentes sistemas operativos); • Arquitetura extensível de fácil desenvolvimento; • Facilidade de reutilização de *queries* de consulta e componentes de transformações; • Baseado em repositórios; • Gestão estruturada de modelos, conexões e *logs*; • Transforma números decimais da BD (*number*) num ficheiro .arff com ',' a separar as casas decimais, provocando ambiguidade com a separação dos atributos.

Tabela 4.4: Conclusões gerais do Pentaho Data Integration.

Análise OLAP

Pentaho Analysis - CE	• Análises interativas e dinâmicas, baseadas na filosofia *drag-and-drop*; • Arquitetura J2EE, baseada em *standards*; • Pouca qualidade gráfica da interface; • Conetividade com JDBC e JNDI; • Consulta de dados baseada em SQL e suporte com MDX e XML (possibilidade de editar *query* MDX); • Sem integração com o Google Maps; • Gráficos e tabelas com navegação (*Member, Position, Through, Down*); • Manipulação dos dados dinâmica, e através da barra de ferramentas (relativamente intuitiva); • Gráficos configuráveis (janela de propriedades); • Escalabilidade e desempenho de acesso baseados em Excel e *Web*; • Possibilidade de criação de relatórios de análise (não se conseguiu a exportação para PDF ou Excel); • Integração com o Aggregation Designer e o Schema Workbench; • Segurança integrada, agendamento de tarefas, alertas, integração do portal e metadados; • Documentação escassa.
Pentaho Analyzer - EE	• Soluções atrativas; • Integração com o Google Maps (GeoMap), traduzido num sistema inteligente de localização; • Análises OLAP com componente geográfico; • Filosofia *drag-and-drop*; • Gráficos da análise gerados automaticamente e com diversos tipos; • Exportação para PDF, Excel ou CSV (exceto o mapa); • Não existe documentação de suporte.

Tabela 4.5: Conclusões gerais da Análise OLAP.

Data Mining	
Pentaho Weka	• Análise preditiva de padrões ocultos e desempenho futuro através de algoritmos de classificação, segmentação e associação; • Avaliação estatística dos modelos de aprendizagem; • Visualização e pré-processamento dos dados de entrada e computação do resultado da aprendizagem; • Contém bastantes algoritmos; • Permite deteção de falhas e erros de registo; • Conclusões retiradas com pouca relevância (em muito devido à limitação de dados utilizados).

Tabela 4.6: Conclusões gerais do Pentaho Data Mining.

Capítulo 5

Conclusões e Trabalho Futuro

O processo de EC emergiu de um campo de investigação que foca essencialmente o desenvolvimento de teorias, metodologias e práticas de extração de informação e conhecimento útil a partir de enormes e extensas BDs. Está integralmente centrado no utilizador uma vez que este controla a seleção e integração dos dados, a limpeza e a transformação dos mesmos, a escolha dos métodos de análise e a interpretação dos resultados, sendo assim considerado um processo de aprendizagem interativo e indutivo. Contudo, e apesar de já existirem provas suficientes da credibilidade e das vantagens do processo de EC para as organizações e instituições, convencer os profissionais de saúde a alterar os seus hábitos baseados apenas nas evidências pode tornar-se num processo complexo e moroso. Principalmente a classe médica recusa alterações nas políticas hospitalares mesmo que confrontados com evidências de melhorias. A grande maioria dos médicos prefere ouvir uma segunda opinião de outro médico do que se basearem nos resultados de EC. Outra das principais questões deste processo no âmbito da área da saúde reside na privacidade dos dados tratados e na ética de utilização de informação clínica real de pacientes. Estes são considerados os principais obstáculos na evolução das técnicas de EC.

A BI é, atualmente, uma área com elevado potencial tanto na área da investigação como da indústria. A aquisição de dados é um processo cada vez mais simples e os modelos de gigantes DW com dimensões de *terabytes* são cada vez mais comuns. Por outro lado, alterações ao nível do *hardware* bem como o decréscimo dos custos da memória principal têm influenciado a forma como se arquitetam os DW nos dias que decorrem. Também o enraizamento dos serviços de armazenamento de dados em rede (*cloud*) tem provocado uma nova forma de raciocínio e, por isso, novas alterações são expectáveis. Finalmente, existe um aumento de pedidos para integrar sistemas de BI nas próximas gerações de dispositivos móveis, sendo que a ferramenta Pentaho já possui alguns dos seus sistemas integrados para Android, iPhone e iPad.

Para a construção e utilização de um sistema de BI como o Pentaho são obviamente necessários conhecimentos específicos do conceitos de BI e de como se processa a tecnologia. Ao nível do *design* das várias etapas do processo de BI são necessárias determinadas decisões sendo que os conhecimentos técnicos do desenvolvedor influenciam diretamente a qualidade das decisões tomadas e, por isso, também a capacidade de resposta às necessidades da organização. Para além dos conhecimentos técnicos, o desenvolvedor deve também identificar as necessidades da organização e o que é esperado com a integração de um sistema de BI, antes mesmo deste começar a desenvolver

soluções. Que tipos de relatórios são interessantes para a organização? Quais as análises exigidas para a criação desses relatórios e se há dados disponíveis no contexto da análise. Deve existir uma ideia inicial clara de quais os objetivos fundamentais da organização.

Em relação à escolha da ferramenta *open source* de BI, o Pentaho é uma ferramenta extremamente vocacionada e orientada para a indústria, ou seja, para a análise de mercado, custos, vendas, produtos e consumidores o que torna a sua aplicação à área da saúde mais complicada. Para tal, é necessária uma adequação da ferramenta aos dados clínicos, tendo sido mesmo assim detetadas algumas limitações e entraves na construção de determinados resultados. Porém, mostrou ser uma possível solução para as instituições de saúde no sentido em que proporciona soluções de BI capazes de serem integradas nos processos clínicos, médicos e administrativos. O Pentaho quando aplicado à saúde pode contribuir para: o desenvolvimento e o acompanhamento de métricas para medir a qualidade dos serviços; a melhoria da segurança dos pacientes e a redução dos erros médicos; o suporte de *standards* HL7; a redução de custos operacionais; e a racionalização da utilização de recursos. Conclui-se que a escolha da ferramenta de BI ideal para uma determinada organização ou instituição é relativa e um pouco subjetiva, não existindo uma ferramenta universal considerada a melhor. Uma vez que as instituições seguem diferentes estratégias e políticas a opção da ferramenta deve ter em conta os objetivos da organização e assim adequar às soluções de BI. Atualmente, encontram-se inúmeras soluções *open source* de BI, tornando-se claro que as ferramentas que contêm uma versão empresarial (tal como o Pentaho) apresentam melhores capacidades de visualização e focam-se sobretudo nos utilizadores finais. Por outro lado, a versão da comunidade adapta-se melhor a utilizadores técnicos de informática ou com alguns conhecimentos específicos.

Numa conclusão global do Pentaho, este mostrou ser uma ferramenta de BI com uma elevada quantidade de funcionalidades, bom suporte técnico por parte da comunidade, muito estável e madura. O Pentaho permite o desenvolvimento de diversas e variadas soluções: ETL, análises OLAP, relatórios, *dashboards* e *data mining*. A ferramenta demonstra funcionalidades como: critérios básicos (como modelagem visual, suporte com SQL Server, suporte *worflow* no ETL, relatórios ad-hoc, gráficos, *dashboards*, suporte português); arquitetura (como multi-plataforma, arquitetura escalável, disponível na internet/intranet, customização funcional de componentes); ETL (como função de agrupamento, função de extração de dados, função de ordenação); relatórios (como relatórios personalizados, exportação para PDF, exportação para formato livre – ODT); usabilidade (como facilidade de uso, atratividade, interface personalizável, bom suporte técnico e documentação); administração (como permitir agendamento de tarefas, permitir gestão centralizada, perfil de utilizador); e produto (como custo zero, amadurecimento, integração).

As principais vantagens do Pentaho assentam, assim, na usabilidade apresentada, na habilidade de padrões utilizados que facilitam integrações, a orientação a ferramentas integradas, a possibilidade de customização e extensão de toda a infra-estrutura, a focalização em desenvolvimento de relatórios e análises adequados aos processos empresariais, a extração de dados, a facilidade de licença a custo zero (Pentaho CE) e os tempos de processamento de poucos segundos.

Num registo conclusivo a nível clínico referente aos casos de estudo, verifica-se uma ligeira melhoria contínua, ao longo dos últimos anos, na gestão e monitorização das

listas de espera. Realçam-se em particular as listas de espera para cirurgia nas quais é possível detetar uma tentativa de diminuir o tempo máximo de espera, muito provavelmente um resultado consequente da implementação do programa SIGIC, porém os dados analisados e considerados no âmbito do projeto não são suficientes para retirar conclusões efetivas a este nível. Por outro lado, observa-se que a monitorização da ocupação das salas operatórias poderia ser mais eficiente, principalmente ao nível dos tempos médios de ocupação de sala em determinadas especialidades e do tempo de início de cada intervenção com principal destaque para as primeiras cirurgias do dia. Assim, um bom funcionamento do bloco operatório pode traduzir-se num maior aproveitamento e numa melhor utilização das salas operatórias e, por isso, numa diminuição do número de casos em lista de espera para cirurgia.

Seria interessante num trabalho futuro fazer uma comparação do Pentaho (ferramenta *open source*) com uma ferramenta proprietária. Nos objetivos iniciais deste projeto previa-se a comparação com o Oracle BI, porém por questões de tempo não foi possível. Um outro trabalho já mais extenso mas igualmente interessante seria a comparação com algumas ferramentas *open source* concorrentes do Pentaho no mercado atual. Em relação ao projeto Pentaho ficou, nesta dissertação, por explorar a ferramenta Mondrian, igualmente pelo motivo de falta de tempo e porque a esta ferramenta foi-lhe atribuída uma prioridade de estudo mais baixa no sentido em que o módulo de análise de dados da plataforma satisfaz as necessidades do utilizador e cumpre os objetivos na perfeição. Por outro lado, considerou-se relevante o suporte de um maior número de dimensões para análise, possibilitando assim a construção de cubos OLAP mais complexos.

Em relação aos casos de estudo da monitorização das listas de espera para cirurgia em que se analisam os tempos de espera foram considerados os tempos entre a data de marcação da cirurgia e a data de operação, não tendo sido considerados os casos de cancelamento. Estes enquanto estão em lista de espera até à data de cancelamento traduzem uma ocupação em lista e consequentemente custos para o hospital, tornando-se portanto conveniente numa análise futura incluir também os casos de cancelamento e considerá-los como parte da lista de espera para cirurgia.

Bibliografia

[1] El-Sappagh et al. A proposed model for data warehouse etl processes. *Journal of King Saud University - Computer and Information Sciences 23*, pages 91–104, 2011.

[2] Pedro Pita Barros. As listas de espera para intervenção cirúrgica em portugal, 2008.

[3] A. D. Binti Rumi. An implementation of pentaho in reporting management module. In *Centre for Advanced Software Engineering, Universiti Teknologi Malaysia*, 2008.

[4] Johanna Viitanen, Hannele Hypponen, Tinja Laaveri, Jukka Vanska, Jarmo Reponen, and Ilkka Winblad. National questionnaire study on clinical ict systems proofs: Physicians suffer from poor usability. *International Journal of Medical Informatics*, 80(10):708 – 725, 2011.

[5] Maribel Yasmina Santos and Isabel Ramos. *Business Intelligence Tecnologias da Informação na Gestão de Conhecimento*. FCA - Editora de Informática, segunda edition, 2009.

[6] Sarah E. Reed, Daniel Y. Na, Theodore C. Mayo, Lance W. Shapiro, Joseph B. Duty, James H. Conklin, and Donald E. Brown. Implementing and analyzing a data mart for the arlington county initiative to manage domestic violence offenders. In *Proceeedings of the 2010 IEEE Systems and Information Engineering Design Symposium*, 2010.

[7] Mounir Ben Ayed, Hela Ltifi, Christophe Kolski, and Adel M. Alimi. A user-centered approach for the design and implementation of kdd-based dss: A case study in the healthcare domain. *Decision Support Systems*, 50(1):64 – 78, 2010.

[8] Filipe Portela, Pedro Gago, Manuel F. Santos, Álvaro Silva, Fernando Rua, José Machado, António Abelha, and José Neves. Knowledge discovery for pervasive and real-time intelligent decision support in intensive care medicine. In *KMIS*, pages 241 – 249. SciTePress, 2011.

[9] Sumana Sharma, Kweku-Muata Osei-Bryson, and George M. Kasper. Evaluation of an integrated knowledge discovery and data mining process model. *Expert Systems with Applications*, 39(13):11335 – 11348, 2012.

[10] Oscar Romero and Alberto Abelló. A framework for multidimensional design of data warehouses from ontologies. *Data & Knowledge Engineering*, 69(11):1138 – 1157, 2010.

[11] Andrew M Wilson, Lehana Thabane, and Anne Holbrook. Application of data mining techniques in pharmacovigilance. *Br J Clin Pharmacol.*, 57(2):127 – 134, 2004.

[12] Ruben D. Canlas. Data mining in healthcare: Current applications and issues. *Carnegie Mellon University*, 2009.

[13] Jeremy Mennis and Diansheng Guo. Spatial data mining and geographic knowledge discovery - an introduction. *Computers, Environment and Urban Systems*, 33(6):403 – 408, 2009.

[14] Richard Lenz and Manfred Reichert. It support for healthcare processes - premises, challenges, perspectives. *Data and Knowledge Engineering*, 61(1):39 – 58, 2007.

[15] N. Matheson. Things to come: postmodern digital knowledge management and medical informatics. *J Am Med Inform Assoc*, 2(2):73 – 78, 1995.

[16] Nathan C. Hulse, Guilherme Del Fiol, Richard L. Bradshaw, Lorrie K. Roemer, and Roberto A. Rocha. Towards an on-demand peer feedback system for a clinical knowledge base: A case study with order sets. *Journal of Biomedical Informatics*, 41(1):152 – 164, 2008.

[17] Huseyin Guruler, Ayhan Istanbullu, and Mehmet Karahasan. A new student performance analysing system using knowledge discovery in higher educational databases. *Computers and Education*, 55(1):247 – 254, 2010.

[18] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uturusamy. Advances in knowledge discovery and data mining. *Cambridge, MA: The MIT Press*, 1996.

[19] Abeer Khan, Nadeem Ehsan, Ebtisam Mirza, and Sheikh Zahoor Sarwar. Integration between customer relationship management (crm) and data warehousing. *Procedia Technology*, 1(0):239 – 249, 2012.

[20] Álvaro Rebuge and Diogo R. Ferreira. Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, 37(2):99 – 116, 2012.

[21] K. Thangavel, P. P. Jaganathan, and P. O. Easmi. Data mining approach to cervical cancer patients analysis using clustering technique. *Asian Journal of Information Technology*, 4(5):403 – 417, 2006.

[22] João Ferreira, Miguel Miranda, António Abelha, and José Mahcado. O processo etl em sistemas data warehouse. In *INForum 2010 - II Simpósio de Informática*, pages 757 – 765, 2010.

[23] Mohammed M I Awad, Mohd Syazwan Abdullah, and Abdul Bashah Mat Ali. Extending etl framework using service oriented architecture. *Procedia Computer Science*, 3(0):110 – 114, 2011.

[24] Shaker H. Ali El-Sappagh, Abdeltawab M. Ahmed Hendawi, and Ali Hamed El Bastawissy. A proposed model for data warehouse etl processes. *Journal of King Saud University - Computer and Information Sciences*, 23(2):91 – 104, 2011.

[25] Tim A. Majchrzak, Tobias Jansen, and Herbert Kuchen. Efficiency evaluation of open source etl tools. *SAC'11 March 21-25*, pages 287 – 294, 2011.

[26] Marleen de Mul, Peter Alons, Peter van der Velde, Ilse Konings, Jan Bakker, and Jan Hazelzet. Development of a clinical data warehouse from an intensive care clinical information system. *Computer Methods and Programs in Biomedicine*, 105(1):22 – 30, 2012.

[27] Salvatore T. March and Alan R. Hevner. Integrated decision support systems: A data warehousing perspective. *Decision Support Systems*, 43(3):1031 – 1043, 2007.

[28] Thomas Thalhammer, Michael Schrefl, and Mukesh Mohania. Active data warehouses: complementing olap with analysis rules. *Data & Knowledge Engineering*, 39(3):241 – 269, 2001.

[29] Andrew Grant, Andriy Moshyk, Hassan Diab, Philippe Caron, Fabien de Lorenzi, Guy Bisson, Line Menard, Richard Lefebvre, Patricia Gauthier, Richard Grondin, and Michel Desautels. Integrating feedback from a clinical data warehouse into practice organisation. *International Journal of Medical Informatics*, 75(5):232 – 239, 2006.

[30] S. J. Welch, S. S. Jones, and T. Allen. Mapping the 24-hour emergency department cycle to improve patient flow. *Jt Comm J Qual Patient Saf.*, 33(5):247 – 255, 2007.

[31] Surajit Chaudhuri, Umeshwar Dayal, and Vivek Narasayya. An overview of business intelligence technology. *Communications of the ACM*, 54(8):88 – 98, 2011.

[32] Jan Fabian Ehmke, Daniel Grosshans, Dirk Christian Mattfeld, and L. Douglas Smith. Interactive analysis of discrete-event logistics systems with support of a data warehouse. *Computers in Industry*, 62(6):578 – 586, 2011.

[33] Alberto A. G. S. Júnior and Catarina C. Bernardino. Proposta de um sistema de business intelligence para exploração de indicadores de gerência de redes. Faculdade de Tecnologia da Universidade de Brasília, 2009.

[34] Luis F. Tapia and Ricardo V. Pinto. Incorporación de elementos de inteligencia de negocios en el proceso de admisión y matrícula de una universidad chilena. *Revista chilena de ingeniería*, 18(3):383 – 394, 2010.

[35] Alisson Ferreira Silva. Business intelligence®: auxílio na tomada de decisão. Brasília, 2010. Departamento de Administração, Faculdade de Economia, Administração e Contabilidade, Universidade de Brasília.

[36] Hsu-Hao Tsai. Global data mining: An empirical study of current trends, future forecasts and technology diffusions. *Expert Systems with Applications*, 39(9):8172 – 8181, 2012.

[37] Francesco Gagliardi. Instance-based classifiers applied to medical databases: Diagnosis and knowledge extraction. *Artificial Intelligence in Medicine*, 52(3):123 – 139, 2011.

[38] Mohamed Z. Elbashir, Philip A. Collier, and Michael J. Davern. Measuring the effects of business intelligence systems: The relationship between business process and organizational performance. *International Journal of Accounting Information Systems*, 9(3):135 – 153, 2008.

[39] J. Hurst and L. Siciliani. Tackling excessive waiting times for elective susrgery: A comparison of policies in twelve oecd countries. *Health Policy*, 72(2):201 – 215, 2003.

[40] A. Fernandes, J. Perelman, and C. Mateus. Health and healthcare in portugal: Does gender matter? Technical report, Instituto Nacional Ricardo Jorge, Lisboa, Portugal, 2010.

[41] Tribunal de Contas. Auditoria ao acesso aos cuidados de saúde do sns - sistema integrado de gestão de inscritos para cirurgia sigic. In *Relatório no 25/07 - 2a. S*, pages 1 – 63, Lisboa, Portugal, 2007.

[42] OPSS Observatório Português dos Sistemas de Saúde. Saúde: que rupturas? In *Relatório de Primavera*, pages 1 – 124, Lisboa, Portugal, 2003.

[43] UCGIC Unidade Central de Gestão de Inscritos para Cirurgia. Manual de gestão de inscritos para cirurgia - processo de gestão de utente. Technical report, Ministério da Saúde, Administração Central do Sistema de Saúde, IP, Lisboa, Portugal, 2005.

[44] Balaji Janamanchi, Evangelos Katsamakas, Wullianallur Raghupathi, and Wei Gao. The state and profile of open source software projects in health and medical informatics. *International Journal of Medical Informatics*, 78(7):457 – 472, 2009.

[45] Teresa Waring and Philip Maddocks. Open source software implementation in the uk public sector: Evidence from the field and implications for the future. *International Journal of Information Management*, 25(5):411 – 428, 2005.

[46] Joshua Greenbaum. Realizing the pentaho agile bi opportunity: Bi for the masses and customer success. Technical report, Enterprise Applications Consulting, 2010.

[47] Raquel Benbunan-Fich. Using protocol analysis to evaluate the usability of a commercial web site. *Information and Management*, 39(2):151 – 163, 2001.

[48] Erik Liljegren. Usability in a medical technology context assessment of methods for usability evaluation of medical equipment. *International Journal of Industrial Ergonomics*, 36(4):345 – 352, 2006.

[49] Younghwa Lee and Kenneth A. Kozar. Understanding of website usability: Specifying and measuring constructs and their relationships. *Decision Support Systems*, 52(2):450 – 463, 2012.

[50] Andre W. Kushniruk, Marc M. Triola, Elizabeth M. Borycki, Ben Stein, and Joseph L. Kannry. Technology induced error and usability: The relationship between usability problems and prescription errors when using a handheld application. *International Journal of Medical Informatics*, 74:519 – 526, 2005.

[51] Ghang Lee, Charles M. Eastman, Tarang Taunk, and Chun-Heng Ho. Usability principles and best practices for the user interface design of complex 3d architectural design and engineering tools. *International Journal of Human-Computer Studies*, 68:90 – 104, 2010.

[52] Huw Dixon and Luigi Siciliani. Waiting-time targets in the healthcare sector: How long are we waiting? *Journal of Health Economics*, 28(6):1081 – 1098, 2009.

Apêndice A

Apêndice

A.1 Registo de Casos na Lista de Espera para Cirurgia

Especialidades com Totais de Espera para Cirurgia

DES_GRUPO	TOTAL
OFTALMOLOGIA	27636
ORTOPEDIA	18156
CIRURGIA VASCULAR	13257
CIRURGIA AMBULATORIO	12578
UROLOGIA	8145
GINECOLOGIA MJD	8108
O.R.L	6647
ESTOMAT/CIR.MAX_FACIAL	6439
NEUROCIRURGIA	6141
DERMATOLOGIA	4800
CIRURGIA 1	4266
CIRURGIA 2	4114
CIRURGIA PEDIATRICA MPIA	3695
CIRURGIA 3	3419
ORL PEDIATRICA MPIA	2918
NAO UTILIZAR -GINECOLOGIA HGSA	2280
PATOLOGIA DO COLO MJD	1713
ORL PEDIATRICA	1691
CIRURGIA PEDIATRICA	1587
UROLOGIA PED MPIA	1238
PROCRIACAO MEDIC ASSISTIDA MJD	1225
URO-GINECOLOGIA MJD	1091
OFTALMOLOGIA (S.FRANCISCO)	944
OBSTETRICIA MJD	919
CIRURGIA PLASTICA MPIA	903
NEUROPATOLOGIA	735
ESTOMATOLOGIA PEDIATRICA MPIA	725
NEFROLOGIA	553
UROLOGIA PEDIATRICA	353
CATETERISMOS DE LONGA DURACAO	336
CIRURGIA PLASTICA PEDIATRICA	322

DES_GRUPO	TOTAL
ESTOMATOLOGIA PEDIATRICA	266
UNID.TRAT.CIRURGICO OBESIDADE	231
CIRURGIA PLASTICA HSA	207
ORTOPEDIA PED. MPIA	132
ORTOPEDIA PEDIATRICA	61
PROGRAMA DE TRANSPLANTE RENAL	29
OFTALMOLOGIA MPIA	7
NEUROCIRURGIA PED. MPIA	4
NEUROCIRURGIA PEDIATRICA	4
GASTRENTEROLOGIA PEDIATRICA	1
PROGRAMA TRANSPLANTE CORNEA	1
NEFROLOGIA PEDIATRICA	1

A.2 Registos Ativos em Lista de Espera para Cirurgia

Figura A.1: Gráfico de pontos desenvolvido no Pentaho CE, onde estão representadas as 10 especialidades com o maior número de registos ativos naquele momento.

A.3 Cirurgias em Lista de Espera (períodos mensais)

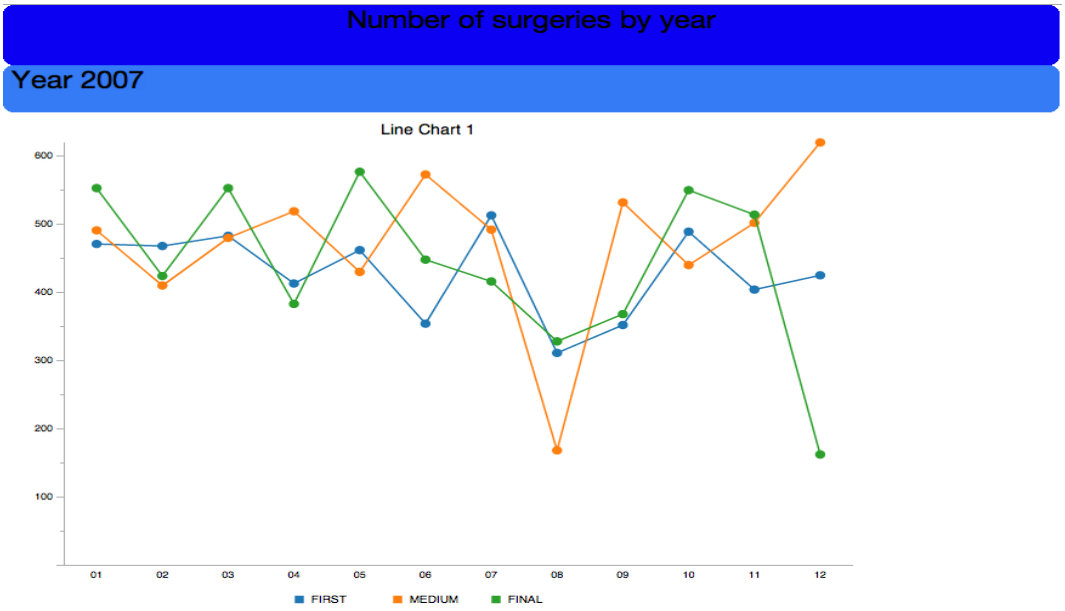


Figura A.2: Número de cirurgias em lista de espera ao longo do ano 2007.

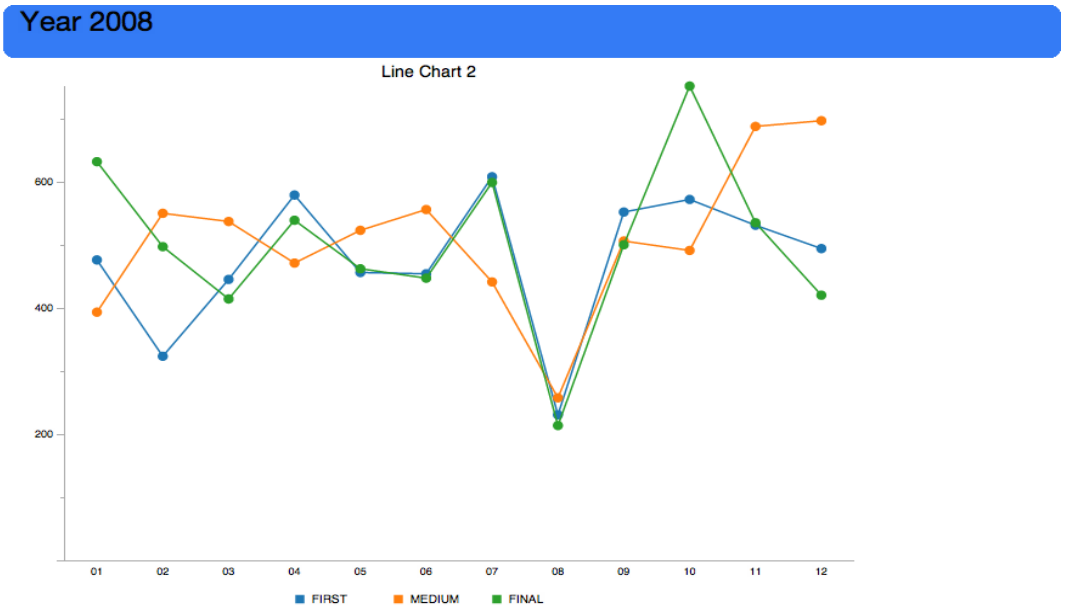


Figura A.3: Número de cirurgias em lista de espera ao longo do ano 2008.

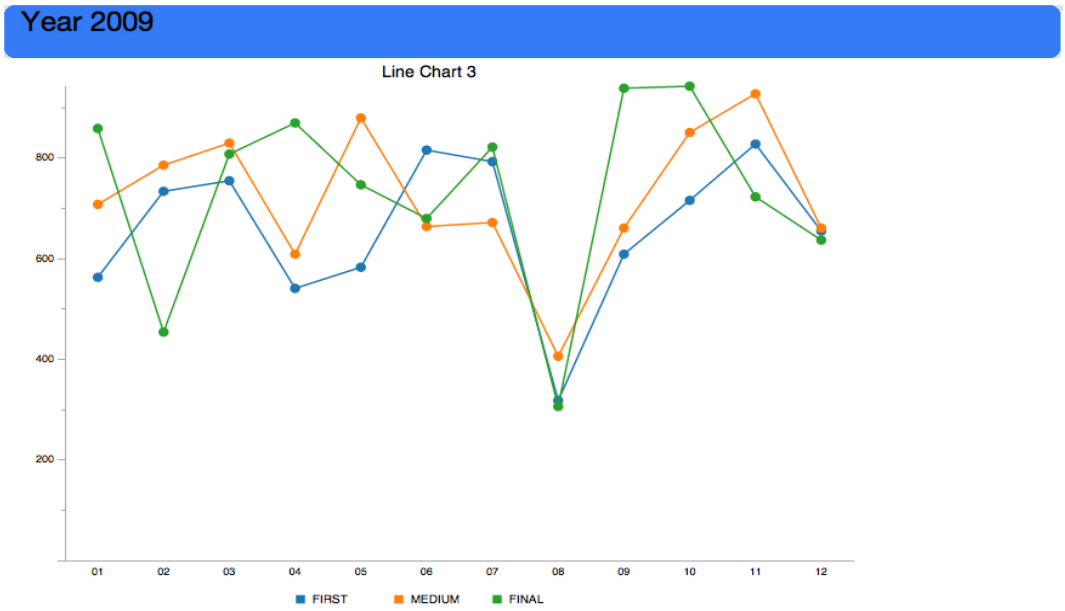


Figura A.4: Número de cirurgias em lista de espera ao longo do ano 2009.

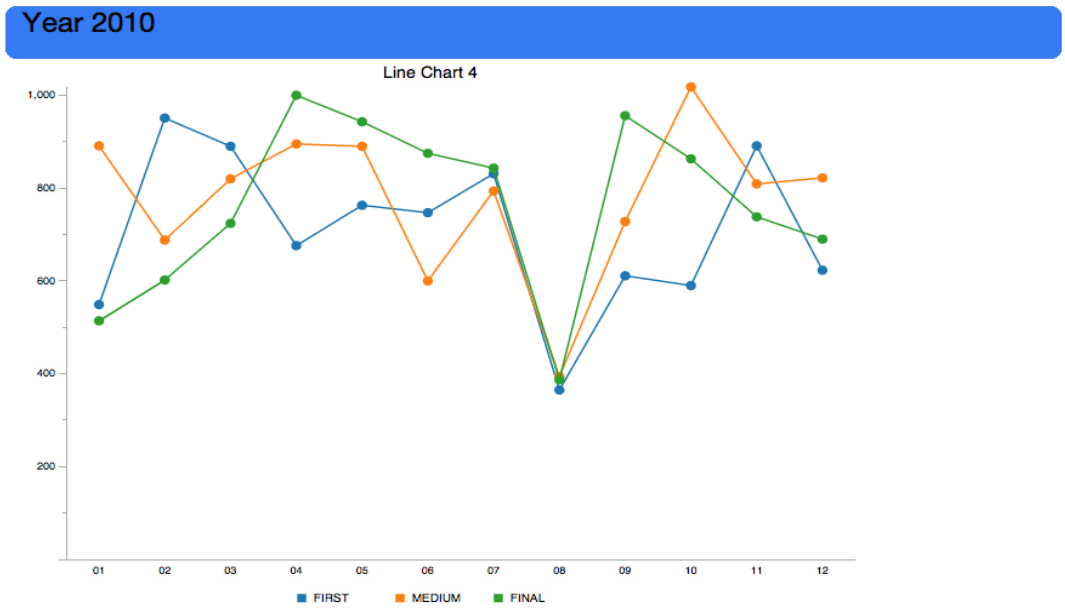


Figura A.5: Número de cirurgias em lista de espera ao longo do ano 2010.

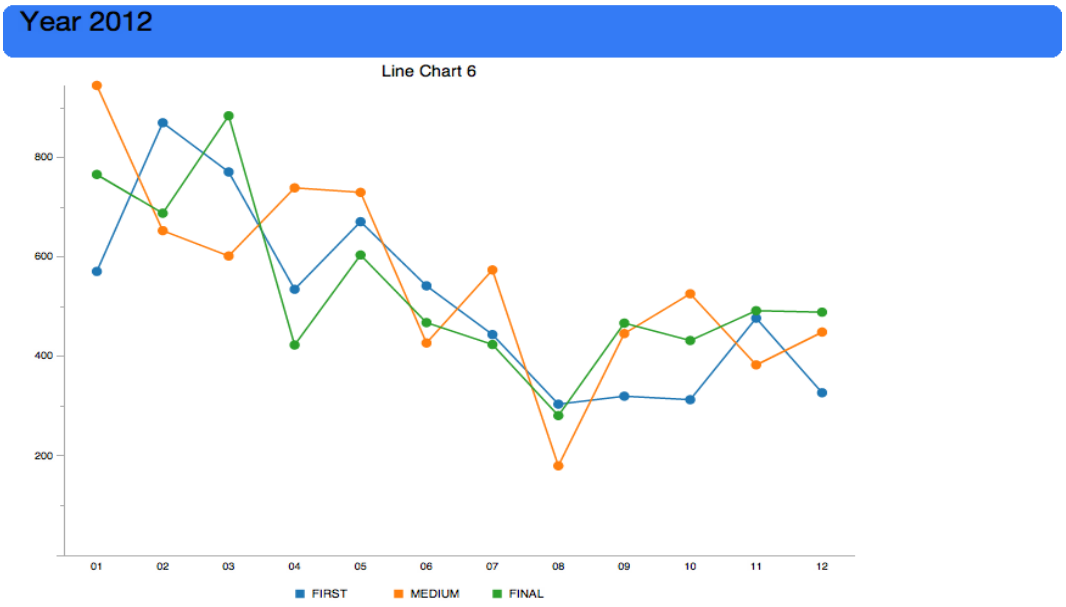


Figura A.6: Número de cirurgias em lista de espera ao longo do ano 2012.

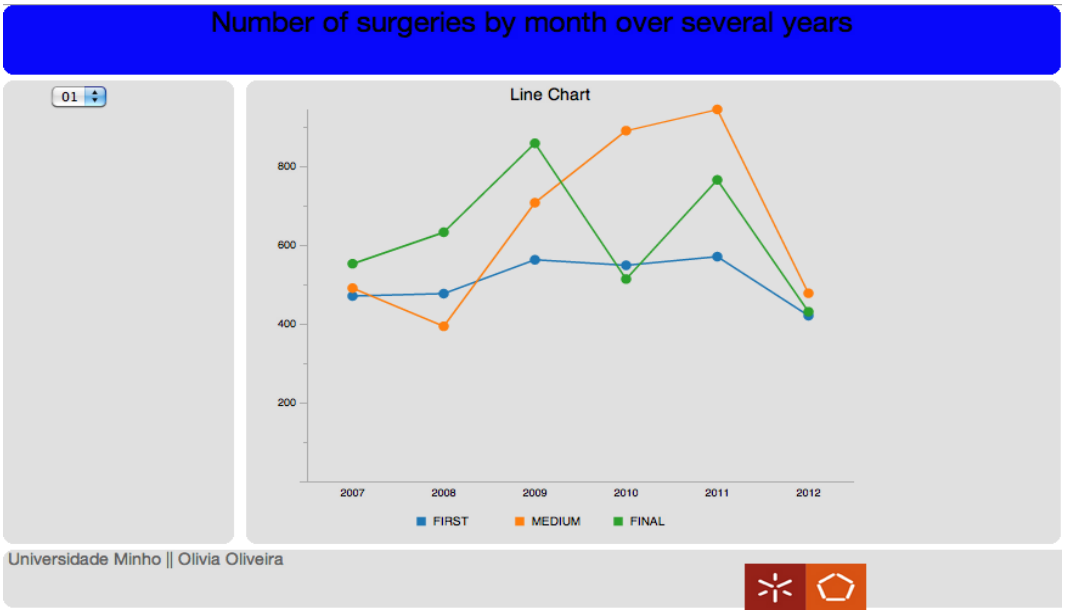


Figura A.7: Número de cirurgias em lista de espera durante vários anos no mês de Janeiro.

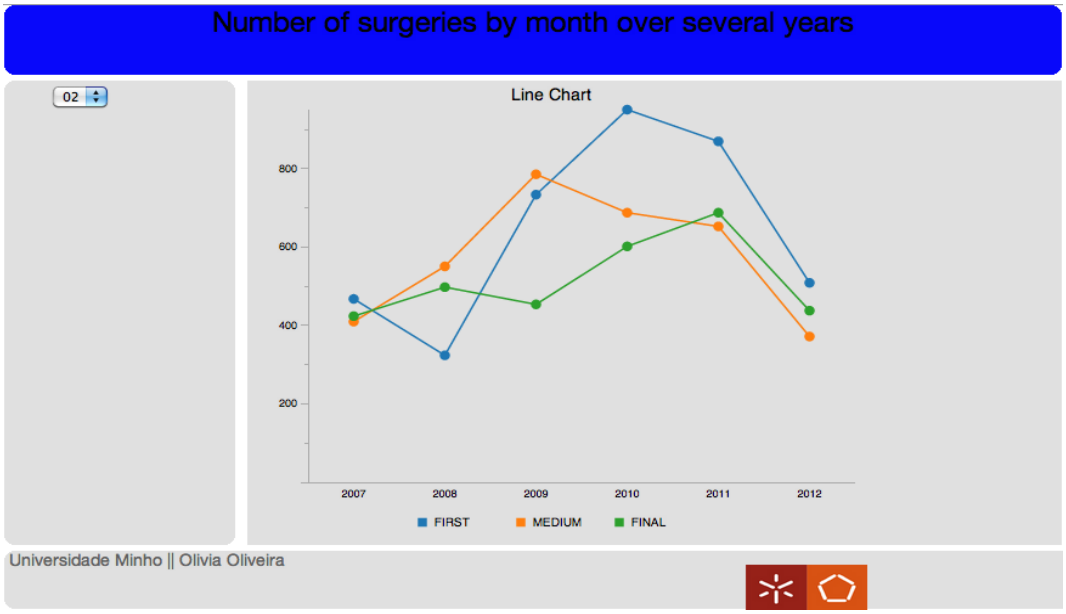


Figura A.8: Número de cirurgias em lista de espera durante vários anos no mês de Fevereiro.

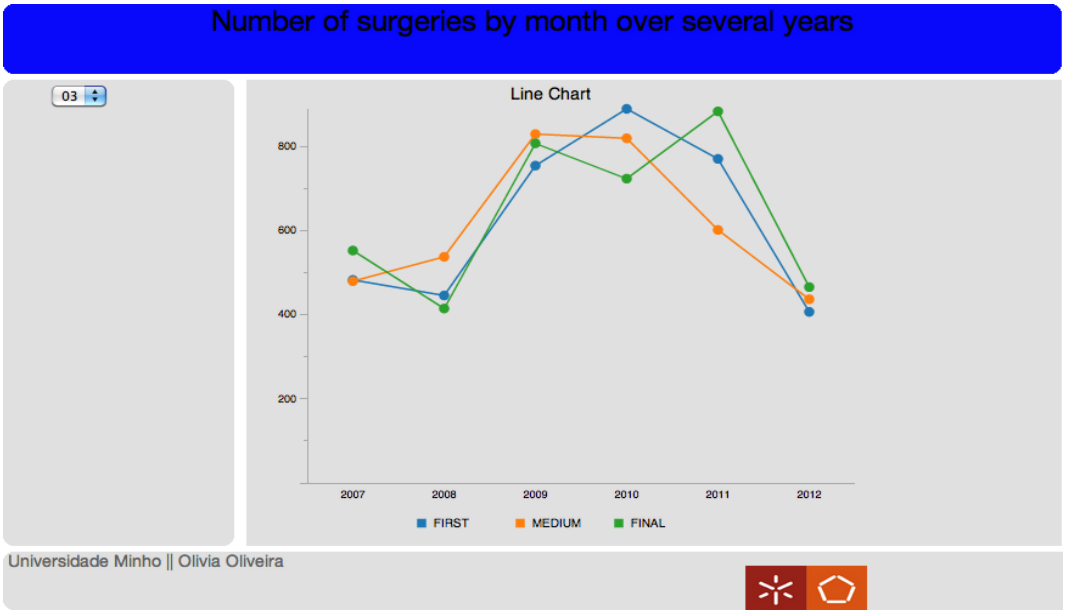


Figura A.9: Número de cirurgias em lista de espera durante vários anos no mês de Março.

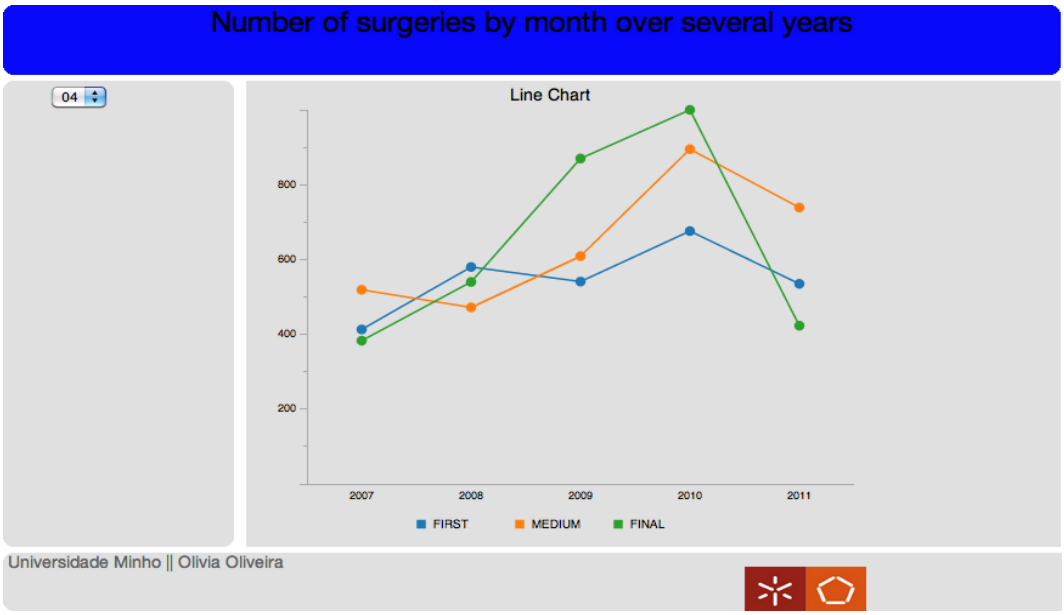


Figura A.10: Número de cirurgias em lista de espera durante vários anos no mês de Abril.

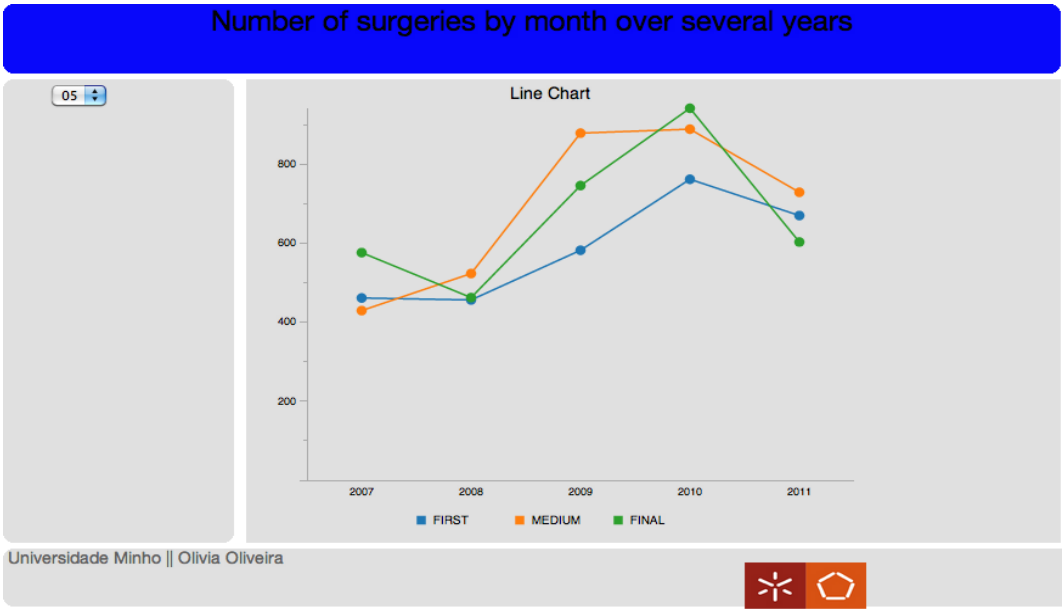


Figura A.11: Número de cirurgias em lista de espera durante vários anos no mês de Maio.

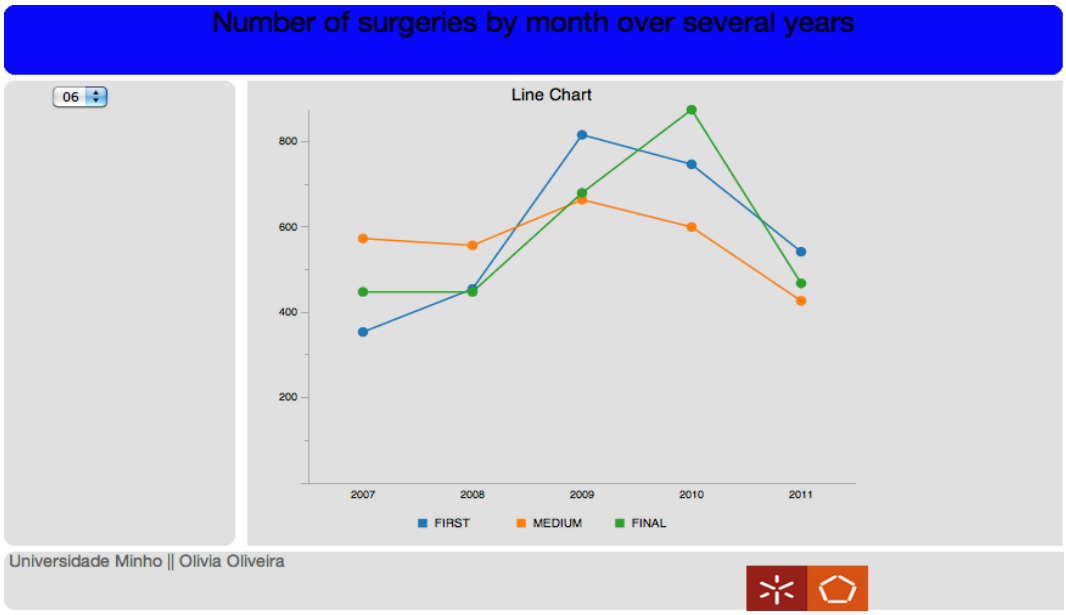


Figura A.12: Número de cirurgias em lista de espera durante vários anos no mês de Junho.

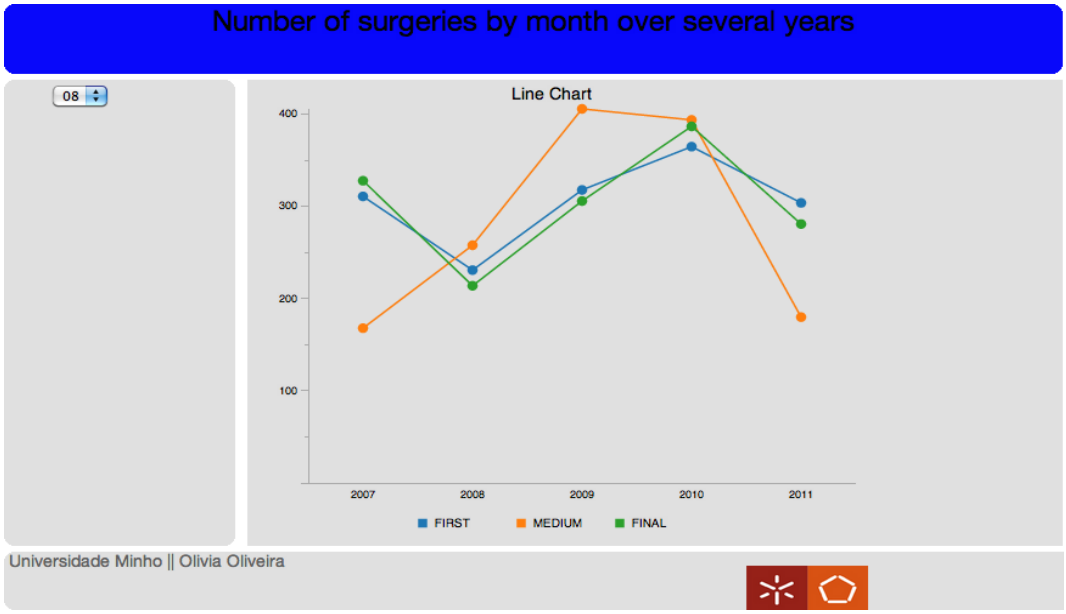


Figura A.13: Número de cirurgias em lista de espera durante vários anos no mês de Agosto.

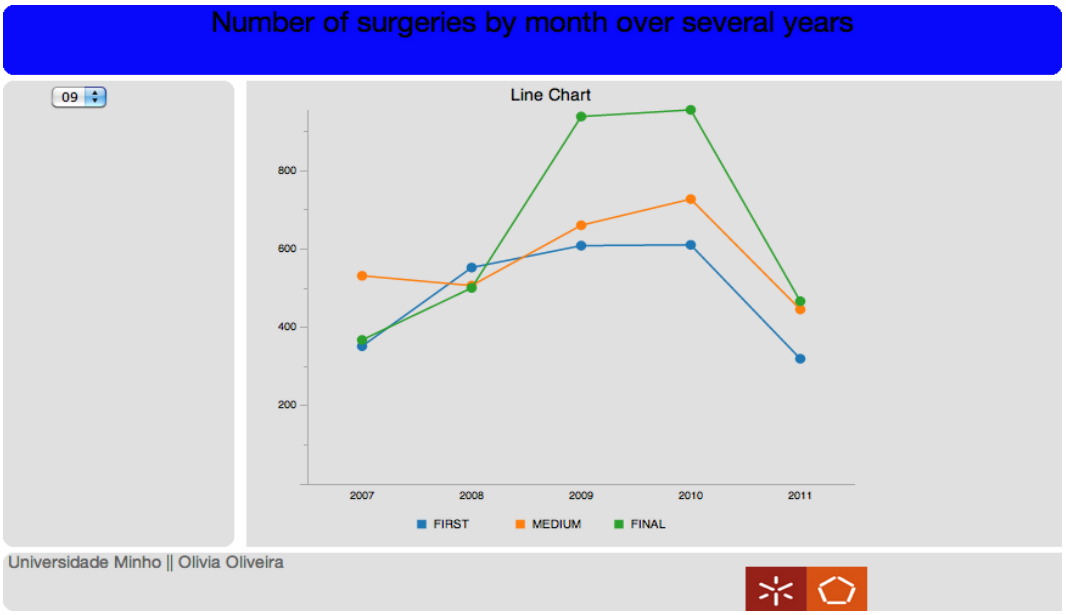


Figura A.14: Número de cirurgias em lista de espera durante vários anos no mês de Setembro.

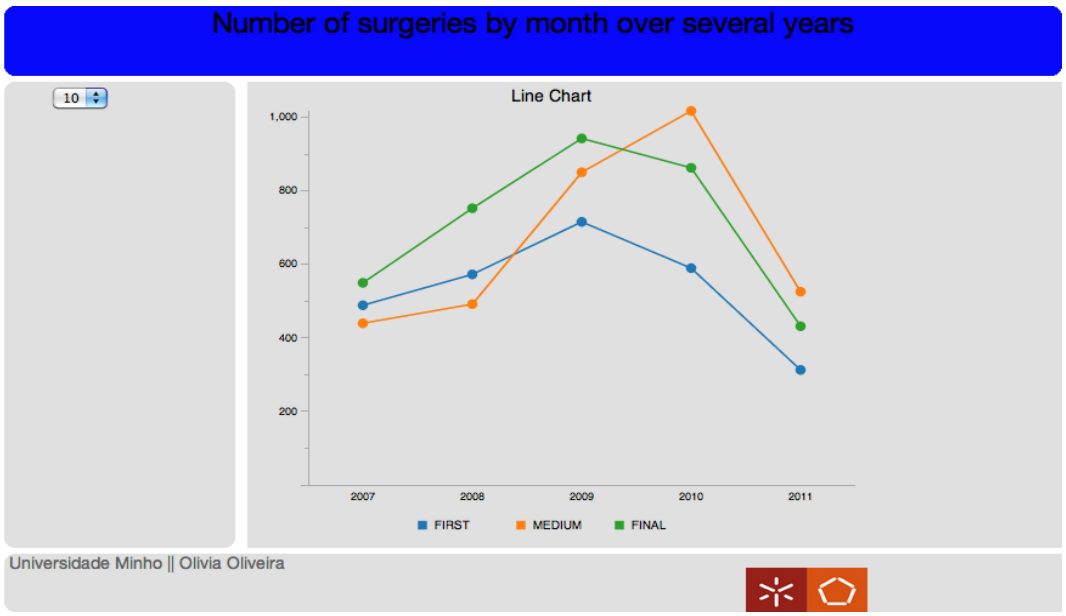


Figura A.15: Número de cirurgias em lista de espera durante vários anos no mês de Outubro.

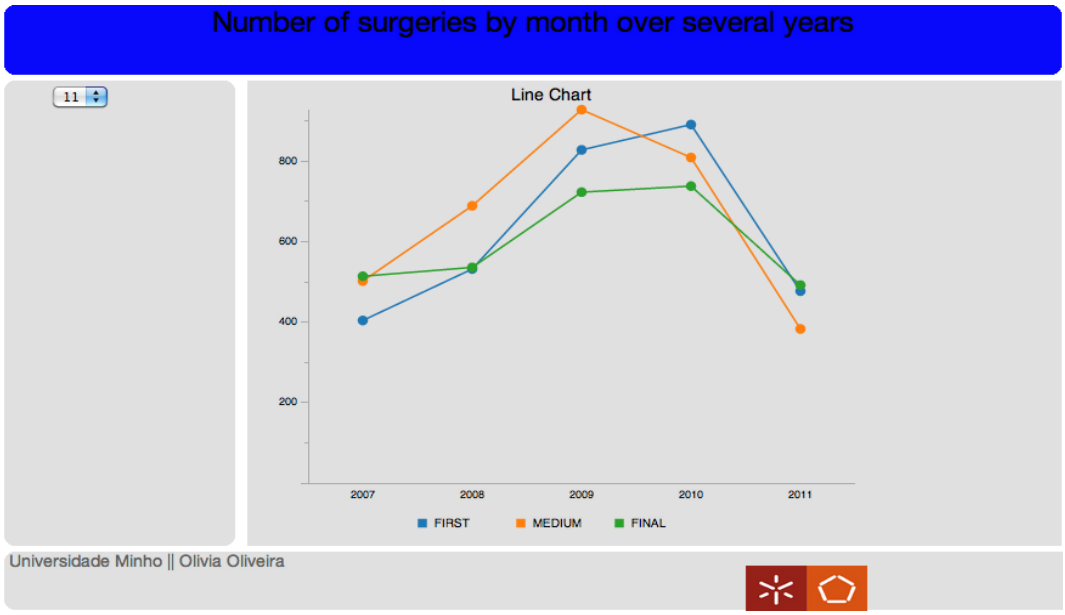


Figura A.16: Número de cirurgias em lista de espera durante vários anos no mês de Novembro.

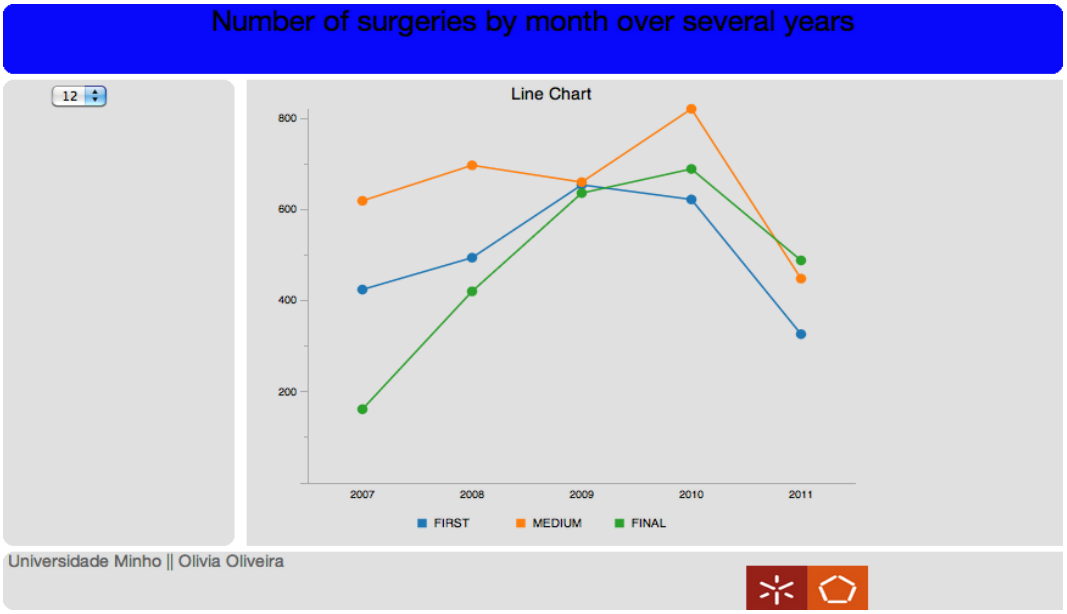


Figura A.17: Número de cirurgias em lista de espera durante vários anos no mês de Dezembro.

A.4 Pentaho Google Maps (GeoMap)

A.4.1 Lista de Espera para Bloco por Distritos

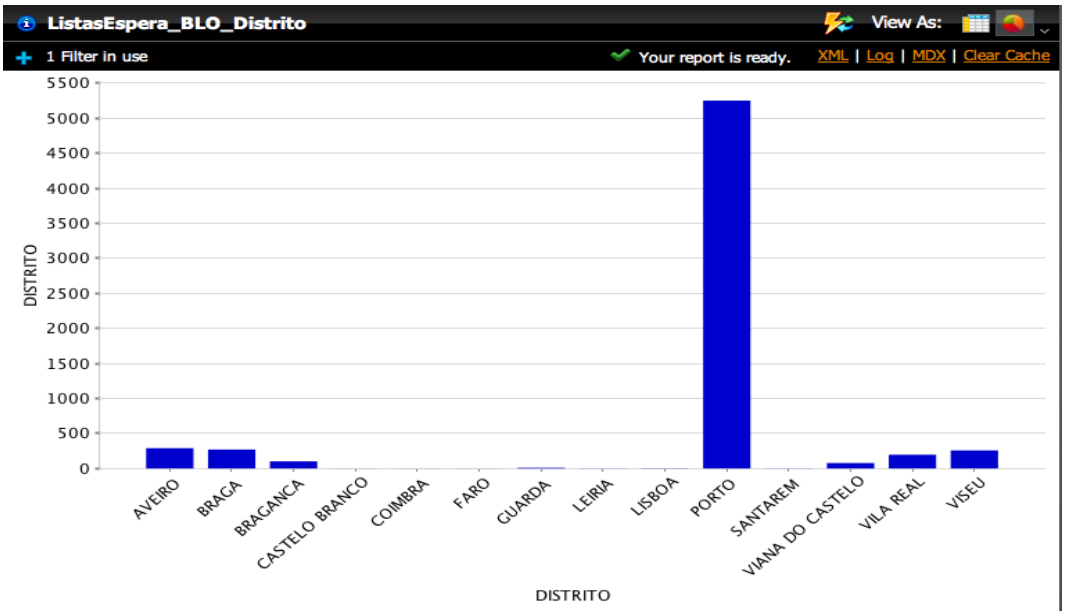


Figura A.18: Gráfico de barras verticais com o número de pacientes em lista de espera para cirurgia por distritos.

Lista de Espera para Bloco por Distritos

PAIS	DISTRITO	TOTAL
PORTUGAL	AVEIRO	289
PORTUGAL	BRAGA	270
PORTUGAL	BRAGANCA	101
PORTUGAL	CASTELO BRANCO	1
PORTUGAL	COIMBRA	1
PORTUGAL	FARO	1
PORTUGAL	GUARDA	10
PORTUGAL	LEIRIA	2
PORTUGAL	LISBOA	3
PORTUGAL	PORTO	5252
PORTUGAL	SANTAREM	2
PORTUGAL	VIANA DO CASTELO	77
PORTUGAL	VILA REAL	196
PORTUGAL	VISEU	257

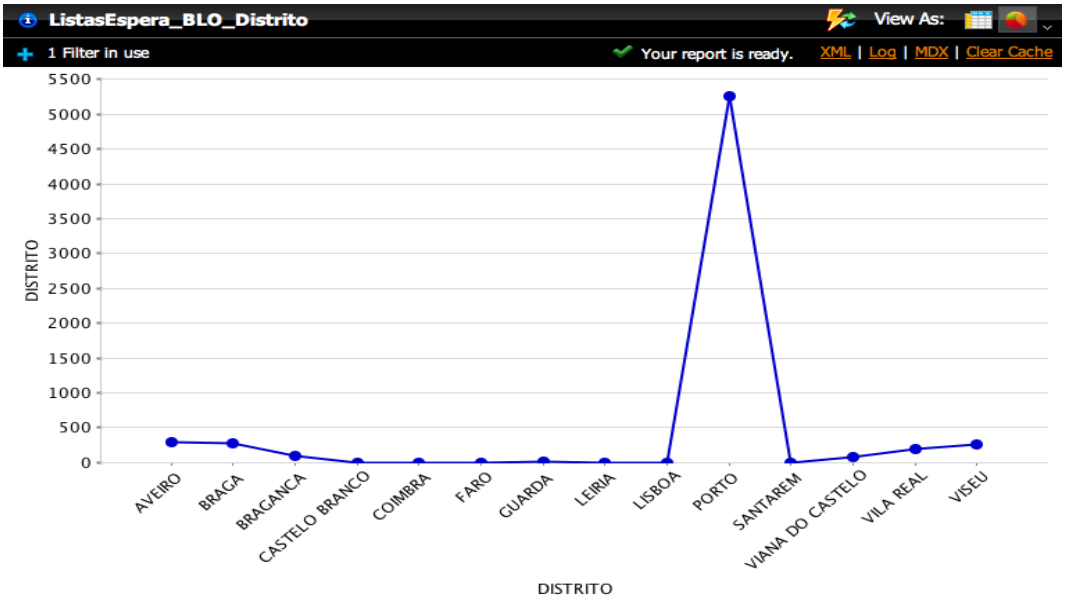


Figura A.19: Gráfico linear, representando o número de pacientes em lista de espera para cirurgia por distritos.

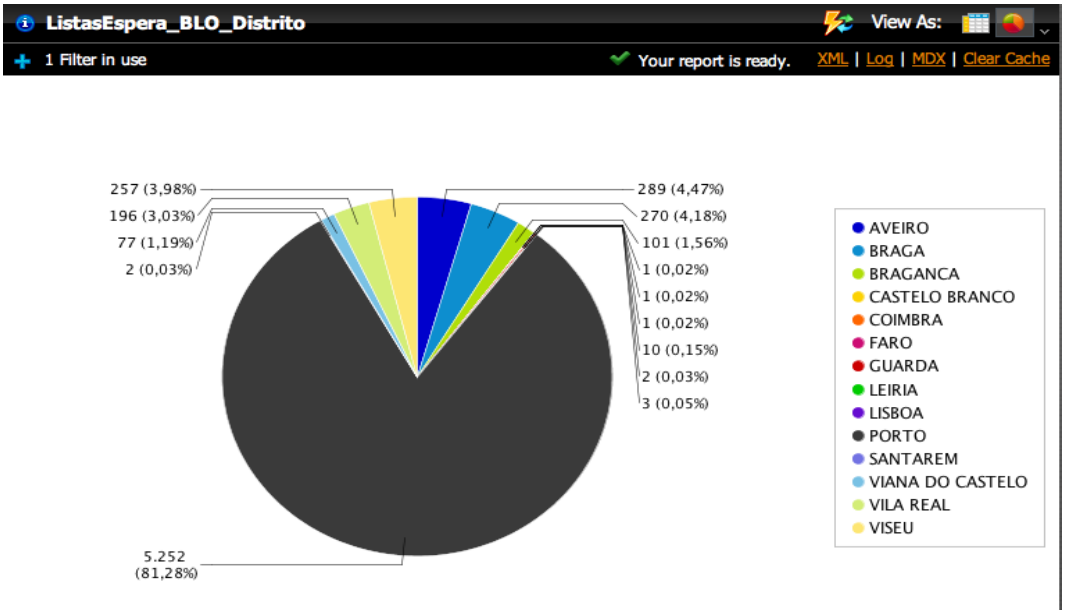


Figura A.20: Gráfico circular com o número de pacientes em lista de espera para cirurgia por distritos.

A.4.2 Lista de Espera para Bloco por concelhos do distrito do Porto

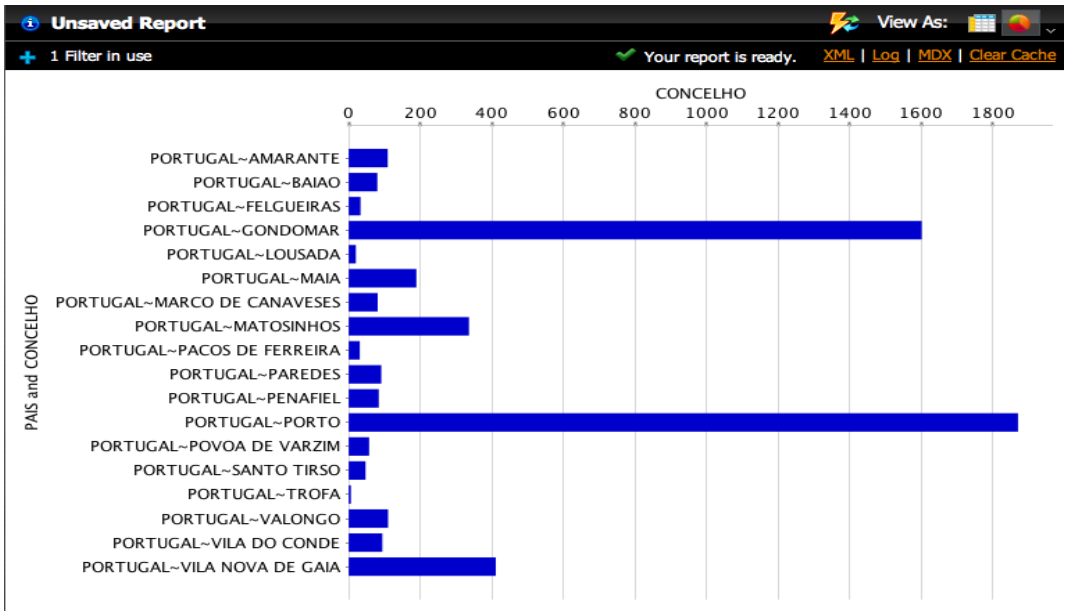


Figura A.21: Gráfico de barras verticais com o número de pacientes em lista de espera para cirurgia por concelhos do distrito do Porto.

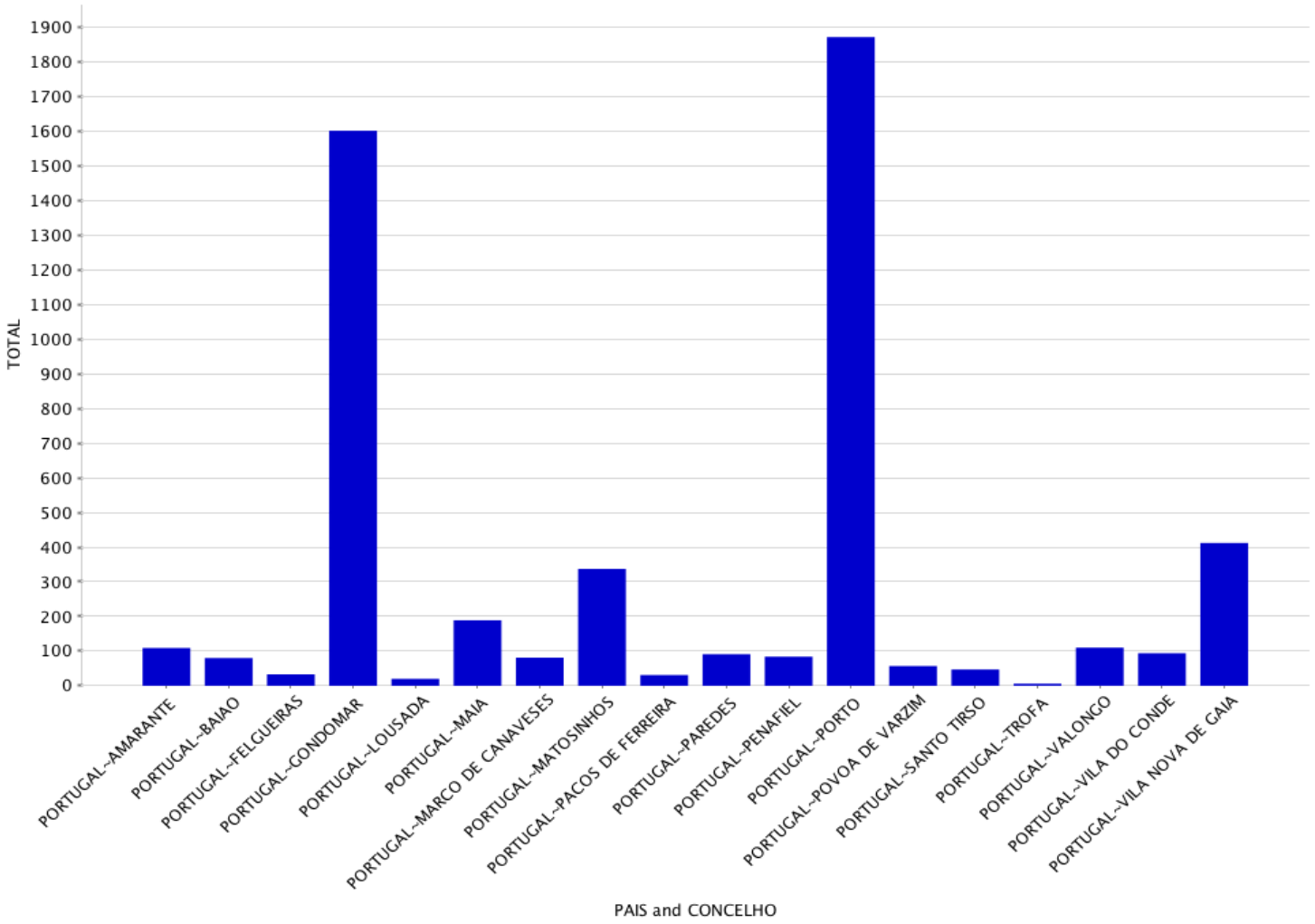
Lista de Espera para Bloco por Concelhos do Porto

CONCELHO	CONCELHO
PORTO	1870
GONDOMAR	1601
VILA NOVA DE GAIA	411
MATOSINHOS	336
MAIA	189
VALONGO	110
AMARANTE	109
VILA DO CONDE	94
PAREDES	91
PENAFIEL	84
MARCO DE CANAVESES	81
BAIAO	80
POVOA DE VARZIM	57
SANTO TIRSO	47
FELGUEIRAS	33
PACOS DE FERREIRA	31
LOUSADA	20
TROFA	6

ListasEspera_BLO_Concelho

By joe

ListasEspera_BLO_Concelho



ListasEspera_BLO_Concelho

PAIS	CONCELHO	TOTAL
PORTUGAL	AMARANTE	109
	BAIAO	80
	FELGUEIRAS	33
	GONDOMAR	1601
	LOUSADA	20
	MAIA	189
	MARCO DE CANAVESES	81
	MATOSINHOS	337
	PACOS DE FERREIRA	31
	PAREDES	91
	PENAFIEL	84
	PORTO	1871
	POVOA DE VARZIM	57
	SANTO TIRSO	47
	TROFA	6
	VALONGO	110
	VILA DO CONDE	94
	VILA NOVA DE GAIA	412

ListasEspera_BLO_Concelho

About this Report

Report Name:ListasEspera_BLO_Concelho

Description:No Description

Report Creator:joe

Report Location:steel-wheels/analysis/ListasEspera_BLO_Concelho.xanalyzer

Created on:1/Out/2012 14:36:39

Cube:DataImport

Filter Summary

CONCELHO includes AMARANTE, BAIAO, FELGUEIRAS, GONDOMAR, LOUSADA, MAIA, MARCO DE CANAVESES, MATOSINHOS, PACOS DE FERREIRA, PAREDES, PENAFIEL, PORTO, POVOA DE VARZIM, SANTO TIRSO, TROFA, VALONGO, VILA DO CONDE and VILA NOVA DE GAIA

Fields Used

PAIS
Original Name:PAIS
Description:No Description

CONCELHO
Original Name:CONCELHO
Description:No Description

TOTAL
Original Name:TOTAL
Description:No Description

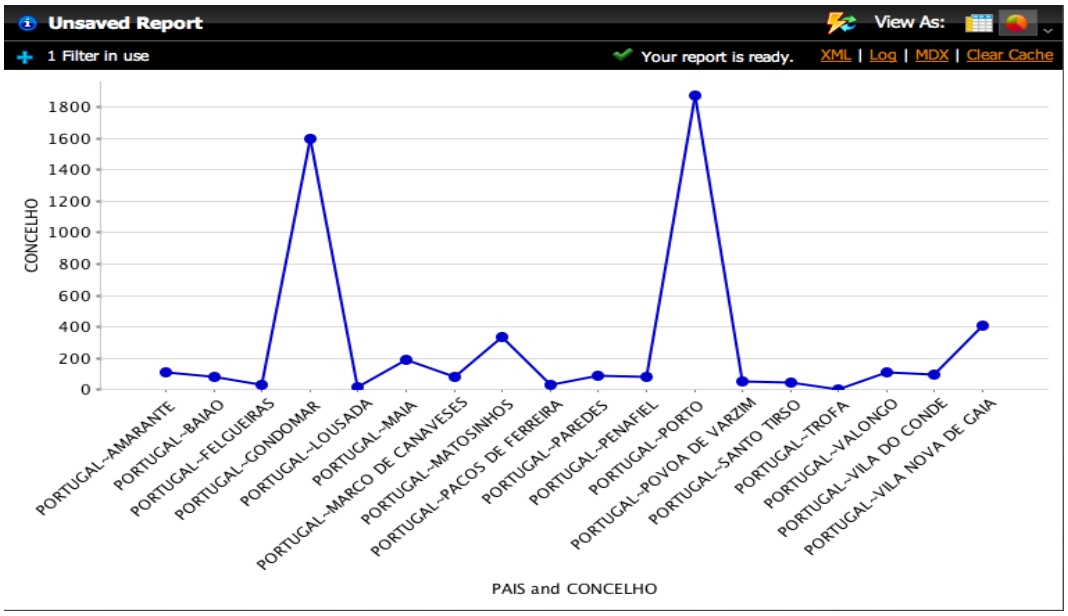


Figura A.22: Gráfico linear, representando o número de pacientes em lista de espera para cirurgia por concelhos do distrito do Porto.

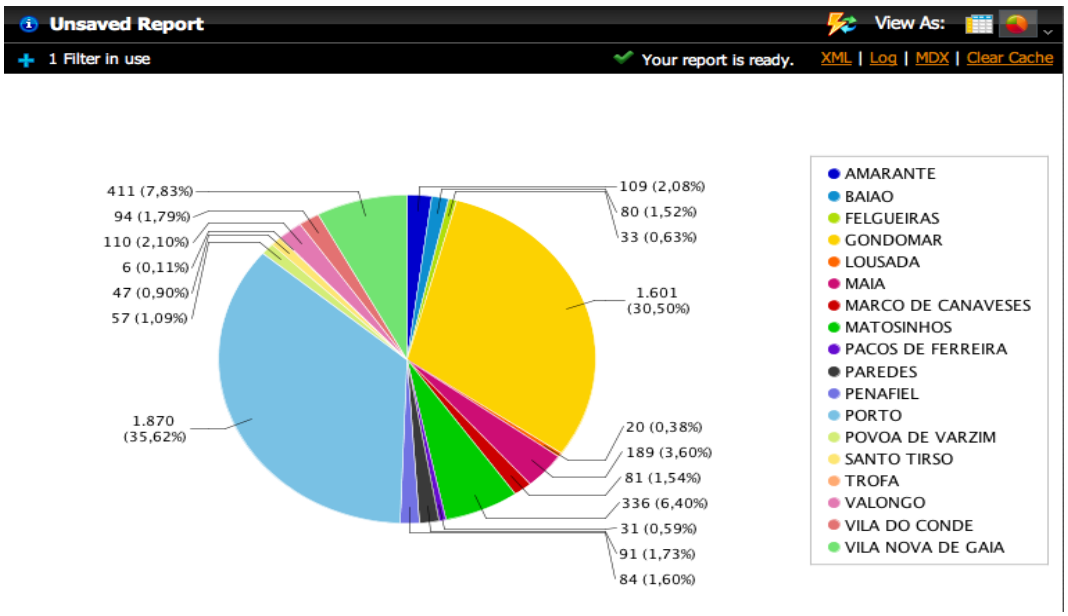


Figura A.23: Gráfico circular com o número de pacientes em lista de espera para cirurgia por concelhos do distrito do Porto.

A.5 Lista de Espera para Bloco - Análise DM

=== Run information ===

Scheme:weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100
Relation: NewRelation-weka.filters.unsupervised.attribute.Remove-R1,3-16,18-20,22-23,25-37,39-41
Instances: 552
Attributes: 5
DES_GRUPO
COD_PATOLOGIA
PRIORIDADE
TIPO_CIRURGIA
DISTRITO
Test mode:evaluate on training data

=== Model and evaluation on training set ===

EM
==

Number of clusters selected by cross validation: 3

	Cluster		
Attribute	0	1	2
	(0.17)	(0.73)	(0.11)
[total]	106.9092	414.5819	72.5089

Time taken to build model (full training data) : 1.79 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 91 (16%)
1 404 (73%)
2 57 (10%)

Log likelihood: -7.88127

=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation: NewRelation-weka.filters.unsupervised.attribute.Remove-R1,3-16,18-20,22-23,25-37,39-41
Instances: 552

```
Attributes:   5
              DES_GRUPO
              COD_PATOLOGIA
              PRIORIDADE
              TIPO_CIRURGIA
              DISTRITO
Test mode:evaluate on training data

=== Model and evaluation on training set ===

kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 933.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute                 Full Data       0       1
                            (552)    (520)     (32)
=============================================================
=
DES_GRUPO                 ORTOPEDIA ORTOPEDIA ORTOPEDIA
COD_PATOLOGIA                 71515    71515     7373
PRIORIDADE                        1        1        1
TIPO_CIRURGIA                     1        1        1
DISTRITO                     PORTO    PORTO    BRAGA

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      520 ( 94%)
1       32 (  6%)
```

```
=== Run information ===

Scheme:weka.clusterers.FarthestFirst -N 2 -S 1
Relation:     NewRelation
Instances:    552
Attributes:   41
              DES_GRUPO
              PRIORIDADE
              TIPO_CIRURGIA
              DISTRITO
```

Test mode:evaluate on training data

=== Model and evaluation on training set ===

FarthestFirst
==========

Cluster centroids:

Cluster 0
	ORTOPEDIA 1 1 PORTO
Cluster 1
	OFTALMOLOGIA 1 2 VISEU

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	504 (91%)
1	 48 (9%)

A.6 Registo de Casos na Lista de Espera para Consulta

Especialidades com Totais de Espera para Consulta

DES_ESPECIALIDADE	TOTAL
CE DERMATOLOGIA /HSA	27217
CE ORTOPEDIA /HSA	25281
CE CENTRO OFTALMOLOGICO /HSA	21585
CE O.R.L. /HSA	18791
CE CIRURGIA VASCULAR /HSA	15258
CE CIRURGIA AMBULATORIO /HSA	11341
CE NEUROLOGIA /HSA	10962
CE OFTALMOLOGIA ADICIONAL / HSA	10599
CE UROLOGIA /HSA	9710
CE ESTOMATOLOGIA /HSA	9589
CE ENDOCRINOLOGIA /HSA	8169
CE NEUROCIRURGIA /HSA	7353
CE GINECOLOGIA/HSA	6195
CE CIRURGIA PEDIATRICA/MPIA	5858
CE GASTRENTEROLOGIA /HSA	5669
CE OTORRINO PEDIATRICA/MPIA	4794
CE OFTALMOLOGIA /HSA	4680
CE PEDIATRIA /HSA	4504
CE PEDOPSIQUIATRIA MGL	4441
CE CARDIOLOGIA /HSA	4393
CE GINECOLOGIA GERAL/MJD	4109
CE CIRURGIA 2A /HSA	3580
CE FISIATRIA - MUSCULO ESQUELETICOS/HSA	3373
CE NUTRICAO /HSA	3323
CE NEFROLOGIA /HSA	3043
CE MEDICINA 2B /HSA	2903
CE OBSTETRICIA BAIXO RISCO/MJD	2843
CE ESTOMATOLOGIA PEDIATRICA/MPIA	2795
CE CIRURGIA 2B /HSA	2671
CE HEMATOLOGIA CLINICA /HSA	2642
CE CIRURGIA 1 /HSA	2628

DES_ESPECIALIDADE	TOTAL
CE ORTOPEDIA PEDIATRICA /HSA	2356
CE SONO /HSA	2326
CE PSIQUIATRIA (ANT.P) /HSA	2296
CE IMUNOALERGOLOGIA PEDIATRICA/MPIA	2140
CE ANESTESIOLOGIA/HSA	2042
CE PLANEAMENTO FAMILIAR/MJD	2005
CE PRE-TRASPLANTE RENAL (NEFRO)/HSA	1968
CE DOR /HSA	1739
CE UROLOGIA PEDIATRICA/MPIA	1721
CE NEUROLOGIA PEDIATRICA/MPIA	1526
CE MEDICINA 1D /HSA	1399
CE PEDIATRIA MEDICA /MPIA	1307
CE GRUPO MEDICINA SEXUAL /HSA	1151
CE CARDIOLOGIA PEDIATRICA/MPIA	1148
CE PSICOLOGIA UNID LIGACAO MPIA	1109
CE FISIATRIA - GERAL /HSA	1067
CE MEDICINA 1A /HSA	976
CE PSICOLOGIA /HSA	927
CE PEDOPSIQUIATRIA UNID. LIGACAO MPIA	918
CE MEDICINA VIAJANTE HJU	907
CE APNEIA DO SONO-CUIDADOS INTENSIVOS 1	892
CE FISIATRIA PEDIATRICA /HSA	877
CE OBSTETRICIA/HSA	859
CE GASTRENTEROLOGIA PEDIATRICA/MPIA	780
CE FISIATRIA - LESOES MEDULARES /HSA	739
CE CIRURGIA MAXILO-FACIAL /HSA	719
CE CIRURGIA PLASTICA/MPIA	711
CE PATOLOGIA DO COLO/MJD	673
CE GRUPO EDUCACAO DM2 /HSA	673
CE PSIQUIATRIA GERAL	672
CE DESABITUACAO TABAGICA /HSA	657
CE ENDOC-GRUPO TIROIDE /HSA	598
CE PEDOPSIQUIATRIA 1.INFANCIA -MGL	571

DES_ESPECIALIDADE	TOTAL
CE URO-GINECOLOGIA/MJD	553
CE FISIATRIA - LESOES ENCEFALICAS /HSA	536
CE MED.FISICA REAB.PEDIATRICA/MPIA	513
CE OBSTETRICIA RISCO/MJD	471
CE PARAMILOIDOSE /HSA	458
CE GESTACAO TERMO/MJD	428
CE APOIO A FERTILIDADE/MJD	427
CE NEFRO PEDIATRICA/MPIA	400
CE DOENTES AUTOIMUNES /HSA	363
CE MEDICINA 2C /HSA	344
CE FISIATRIA - REUMATISMO /HSA	329
CE NEUROLOGIA PED/EPILEPSIA MPIA	310
CE DERMATOLOGIA PEDIATRICA/MPIA	301
CE ENDOCRINOLOGIA - PE DIABETICO /HSA	298
CE CIRURGIA AMB ADICIONAL/HSA	288
CE PSICOLOGIA - MGL	285
CE ORTOPEDIA ADICIONAL /HSA	265
CE PERI-OPERATORIA /MJD	260
CE ENDOC-GRUPO HIPOFISE /HSA	260
CE IMUNOALERGOLOGIA /HSA	249
CE GINEC PEDIAT E ADOLESCENCIA/MJD	245
CE I.G.O./MJD	241
CE NUTRICAO / MPIA	241
CE NEUROLOGIA PED/CEFALEIAS MPIA	233
CE ORTOPEDIA PEDIATRICA/MPIA	225
CE ENDOCRINOLOGIA PEDIATRICA/MPIA	221
CE FISIATRIA - AMPUTADOS /HSA	220
CE HEMATOLOGIA PEDIATRICA/MPIA	173
CE RAYNAUD / CAPILAROSCOPIA HSA	169
CE NEUROCIRURGIA PEDIATRICA/MPIA	162
CE ESTOMATOLOGIA HJU	154
CE PEDIATRIA-PNEUMOLOGIA MPIA	149
CE PSIQUIATRIA GERAL / GONDOMAR	148

DES_ESPECIALIDADE	TOTAL
CE PEDOPSIQUIATRIA P.C.ALIMENTAR -MGL	134
CE PNEUMOLOGIA GERAL HJU	133
CE ANESTESIOLOGIA/MJD	129
CE PATOLOGIA MAMA/MJD	122
CE GASTRO-NUTRICAO/MPIA	122
CE ENDOCRINOLOGIA PED/ HSA	120
CE OBSTETRICIA HIPERTENSAO/MJD	116
CE PEDIATRIA DOEN METABOLICAS/MPIA	116
CE CIRURGIA PLASTICA HSA	116
CE OBSTETRICIA DIABETES/MJD	113
CE PSICOLOGIA GERAL	98
CE OFTALMOLOGIA PEDIATRICA/MPIA	96
CE PSIQUIATRIA HJU	95
CE OBSTETRICIA ADOLESCENTES/MJD	90
CE PROCRIACAO MEDICAMENTE ASSISTIDA/MJD	86
CE PRE-CONCEPCAO/MJD	79
CE NEUROLOGIA PED/DOENCAS MOVIM MPIA	77
CE RAC/MJD	70
CE MULT. ENDOCRINOLOGIA PED/HSA	65
CE PRE-TRANSPLANTE HEPATICO/HSA	61
CE PEDIATRIA GENETICA/MPIA	58
CE PEDIATRIA IMUNOLOGIA/MPIA	55
CE NUTRICAO HJU	55
CE NEUROLOGIA PED/DOENCAS NEUROM MPIA	53
CE HEMAT-GRUPO LINFOMAS CUTANEOS /HSA	43
CE PEDIATRIA DESENVOLVIMENTO/MPIA	43
CE D.P.C.A. (NEFROLOGIA) / HSA	42
CE PSICOLOGIA HJU	40
CE IMUNO ALERGIA FARMACOS PED / MPIA	40
CE MENOPAUSA/MJD	37
CE UROLOGIA/MJD	36
CE GINECOLOGIA ONC/MJD	33
CE GASTRO-HEPATOLOGIA MPIA	30

DES_ESPECIALIDADE	TOTAL
CE ANESTESIA / PECLEC /HSA	29
CE IMUNO ALERGIA ALIMENTAR PED/MPIA	28
CE INFECCIOLOGIA GERAL HJU	28
CE GENETICA /MJD	27
CE PATOLOGIA DIGESTIVA PEDIATRICA/MPIA	26
CE INFECCIOLOGIA C HJU	25
CE ENDOC-GRUPO C.TIROIDE /HSA	25
CE DIAG. PRE-NATAL/MJD	23
CE PSICOLOGIA GERAL / GONDOMAR	23
CE ESTOMATERAPIA /HSA	22
CE IMUNOLOGIA CLINICA 2 / HSA	20
CE PSICOLOGIA /MJD	19
CE PEDOPSIQUIATRIA CRISE -MGL	17
CE PEDIATRIA-DOEN NEUROMUSCULARES/MPIA	16
CE PATOLOGIA FETAL/MJD	14
CE OTORRINO PEDIATRICA-SURDEZ/MPIA	13
CE MULTIDISCIPLINAR OBESIDADE /HSA	13
CE MULTID CARDIOPATIAS CONGENITAS	13
CE GASTRO-OBESIDADE MORBIDA/MPIA	11
CE MEDICINA FAMILIAR /HSA	11
CE PATOLOGIA ENDOCRINA GRAVIDEZ MJD	11
CE ANESTESIOLOGIA PEDIATRICA/MPIA	10
CE CESSACAO TABAGICA /HSA	10
CE ACONSELHAMENTO GENETICA/MJD	10
CE ANALGESIA/MJD	10
CE MULT. TRAT. DIABETES TIPO1 POR BOMBAS	8
CE REABILITACAO RESPIRATORIA HJU	8
CE ANTI-TABAGICA HJU	8
CE GASTRENTEROLOGIA ADICIONAL/HSA	7
CE TRANSPLANTES RENAIS (NEFRO)/HSA	6
CE PEDIATRIA FIBROSE QUISTICA /MPIA	5
CE SAP-INFECCIOLOGIA HJU	5
CE NEFRO PED/ESPINHA BIFIDA /MPIA	5

DES_ESPECIALIDADE	TOTAL
CE NEURO PED/DEGEN METAB MPIA	5
CE CIRURGIA PEDIATRICA /HSA	5
CE MED.FISICA REAB.ESPINHA BIFIDA/MPIA	4
CE GRUPO PATOLOGIA MAMARIA /HSA	4
CE AVALIACAO RISCO OBST/MJD	3
CE SAP-PNEUMOLOGIA HJU	3
CE PRIMEIRO TRIMESTRE /HSA	2
CE NEFRO PED/DIAG.PRENATAL MPIA	2
CE NUTRICAO ADICIONAL/HSA	2
CE AMTCO (AV.MULT.TRAT.CIRUR.OBESIDADE)	2
CE DESENVOLVIMENTO R.N./MJD	2
CE INFECCIOLOGIA B HJU	2
CE TROMBOSE HEMOSTASE/MJD	1
CE DIABETES/GRAVIDEZ /HSA	1
CE PED NEFROLOGIA /HSA	1
CE ANESTESIOLOGIA PED/HSA	1
NAO UTILIZAR- EPILEPSIA/GRAVIDEZ /HSA	1
CE ANTIRABICO HJU	1
CE NEFRO PED/TRANSPLANTE RENAL MPIA	1
CE NEUROLOGIA/NEUROPSICOLOGIA HSA	1

A.7 Registos Ativos em Lista de Espera para Consulta

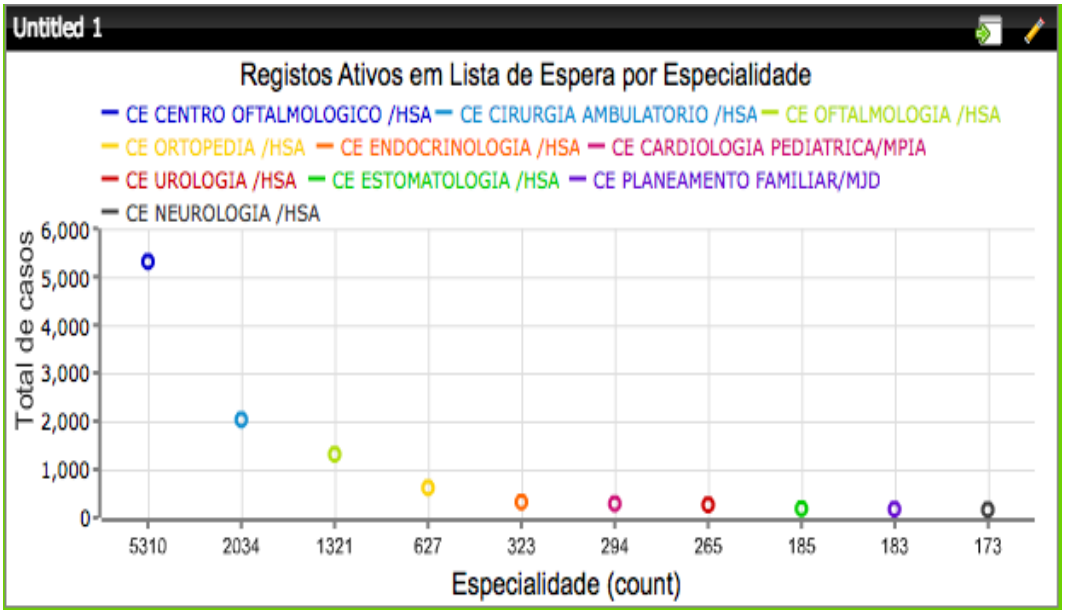


Figura A.24: Gráfico de pontos desenvolvido no Pentaho EE, onde estão representadas as 10 especialidades com o maior número de registos ativos naquele momento.

A.8 Consultas em Lista de Espera (períodos mensais)

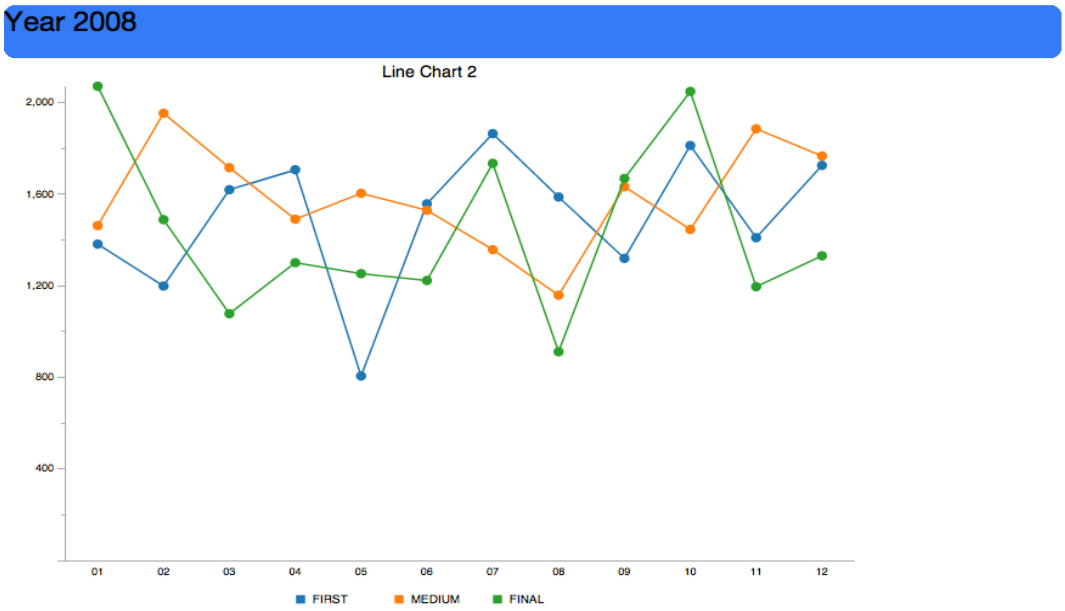


Figura A.25: Número de consultas em lista de espera ao longo do ano 2008.

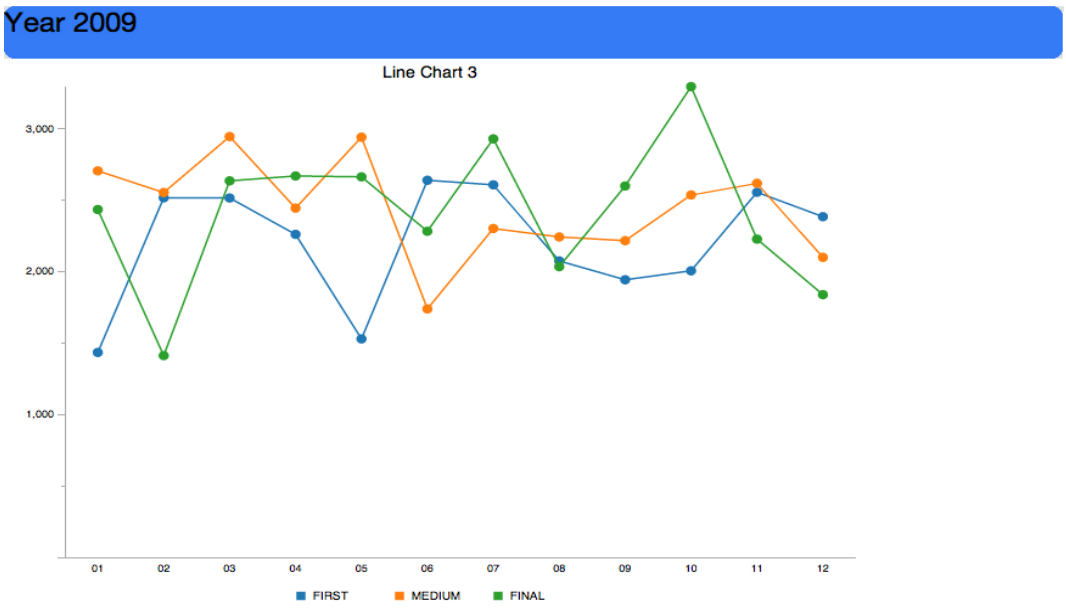


Figura A.26: Número de consultas em lista de espera ao longo do ano 2009.

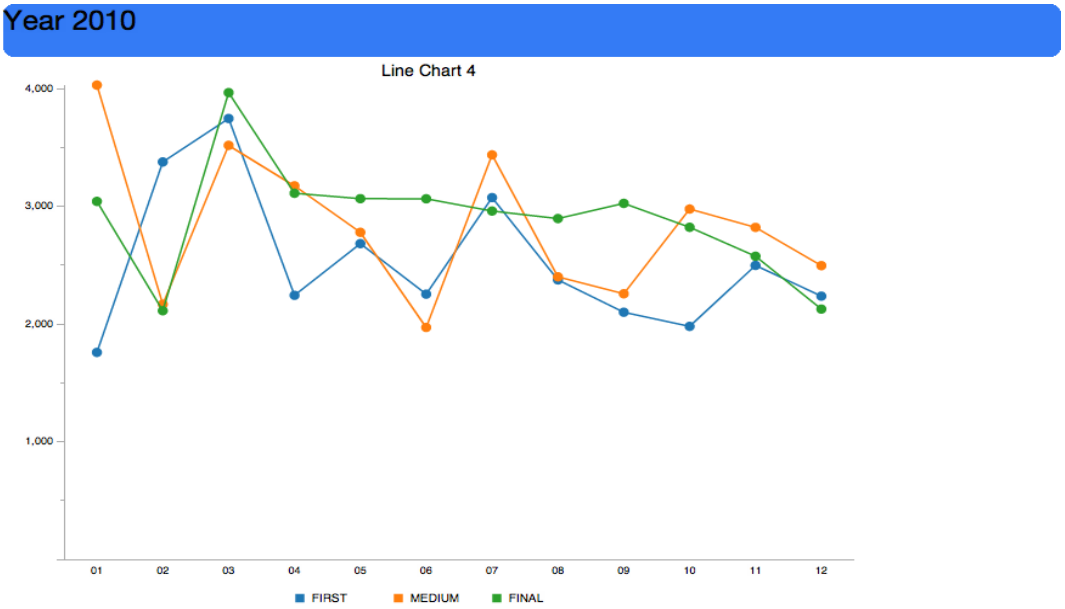


Figura A.27: Número de consultas em lista de espera ao longo do ano 2010.

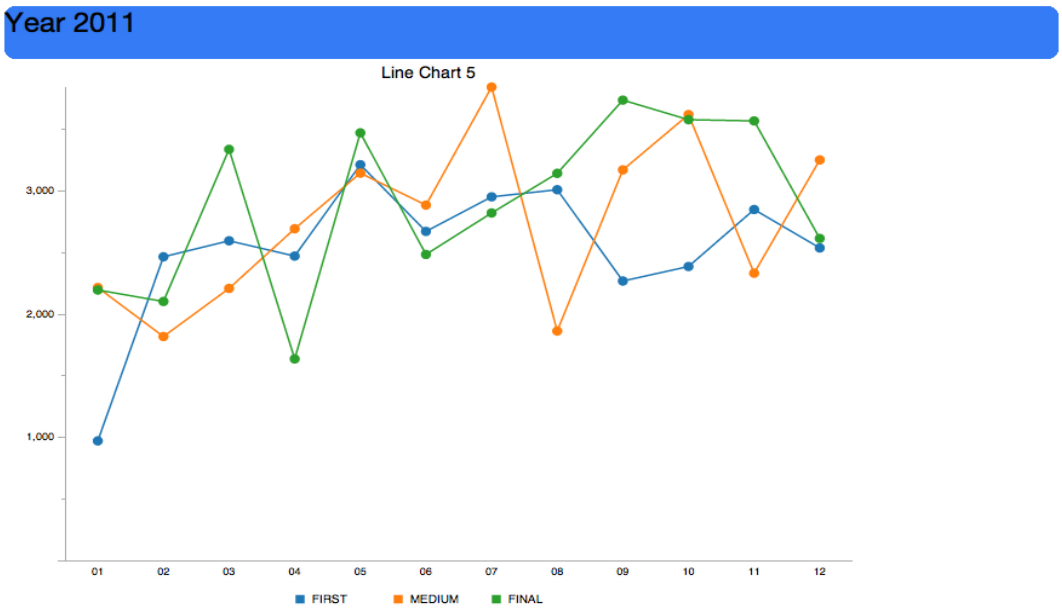


Figura A.28: Número de consultas em lista de espera ao longo do ano 2011.

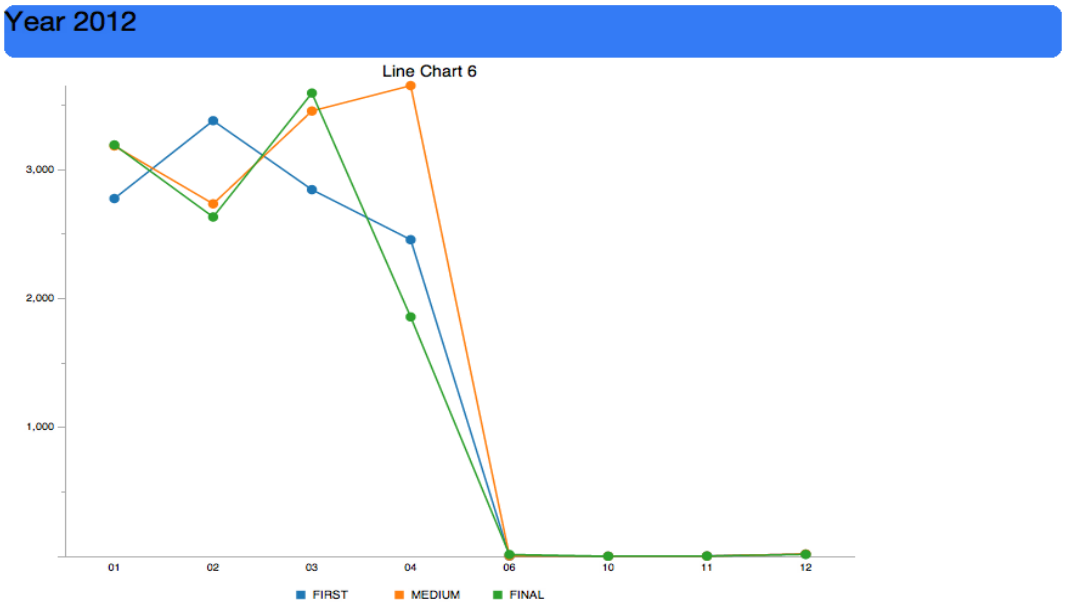


Figura A.29: Número de consultas em lista de espera ao longo do ano 2012.

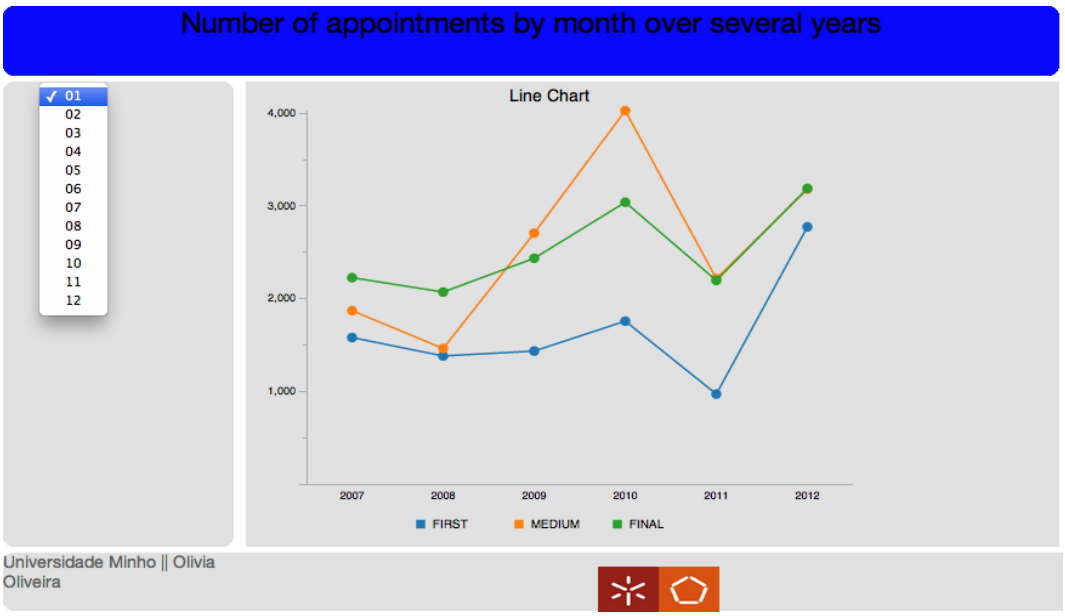


Figura A.30: Número de consultas em lista de espera durante vários anos com o parâmetro '01' selecionado (mês de Janeiro).

Figura A.31: Número de consultas em lista de espera durante vários anos no mês de Fevereiro.

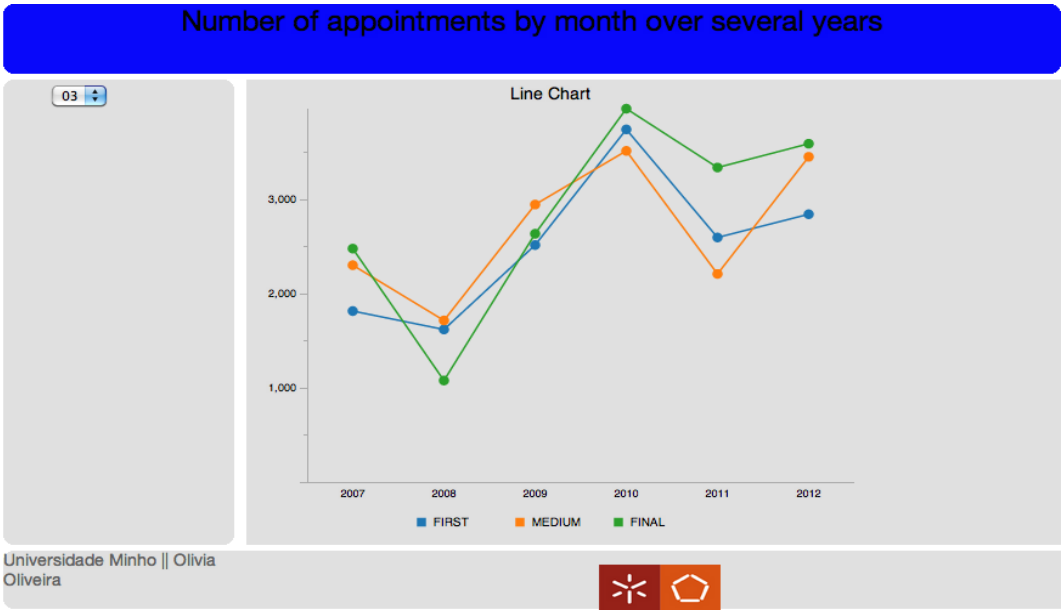


Figura A.32: Número de consultas em lista de espera durante vários anos no mês de Março.

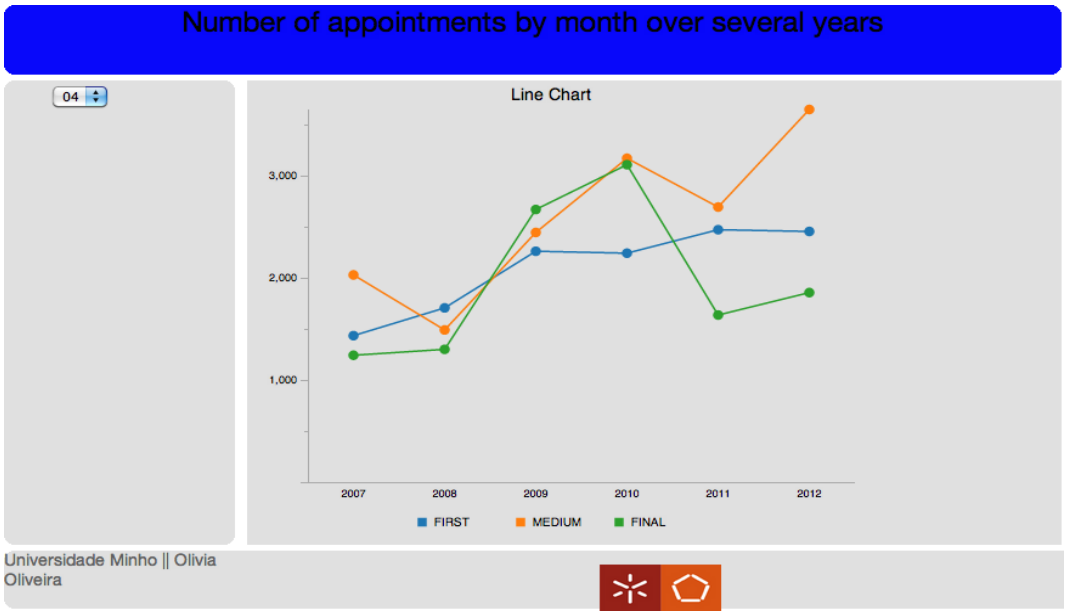


Figura A.33: Número de consultas em lista de espera durante vários anos no mês de Abril.

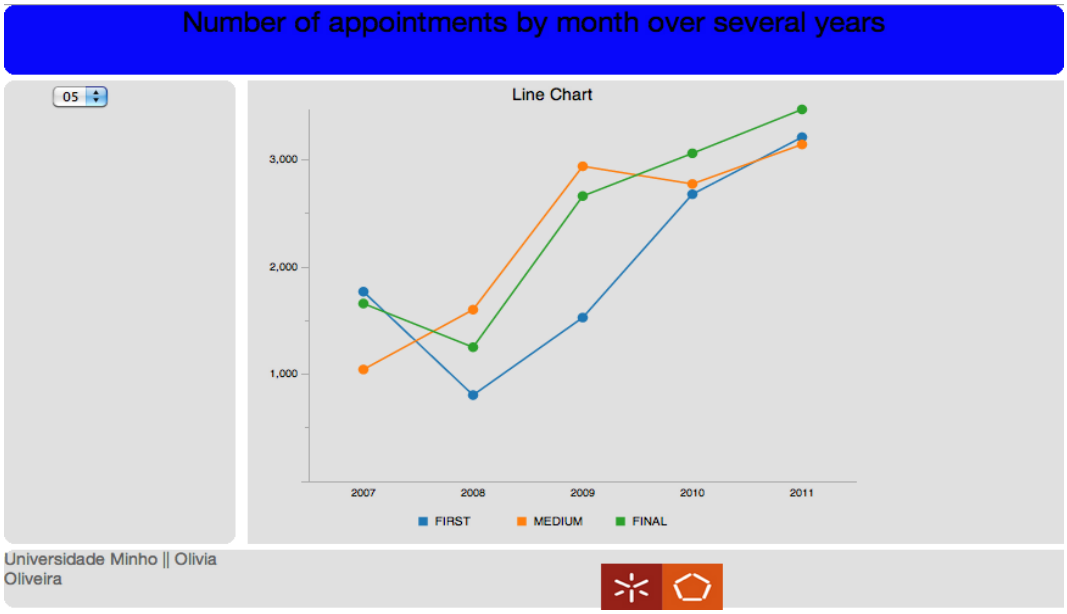


Figura A.34: Número de consultas em lista de espera durante vários anos no mês de Maio.

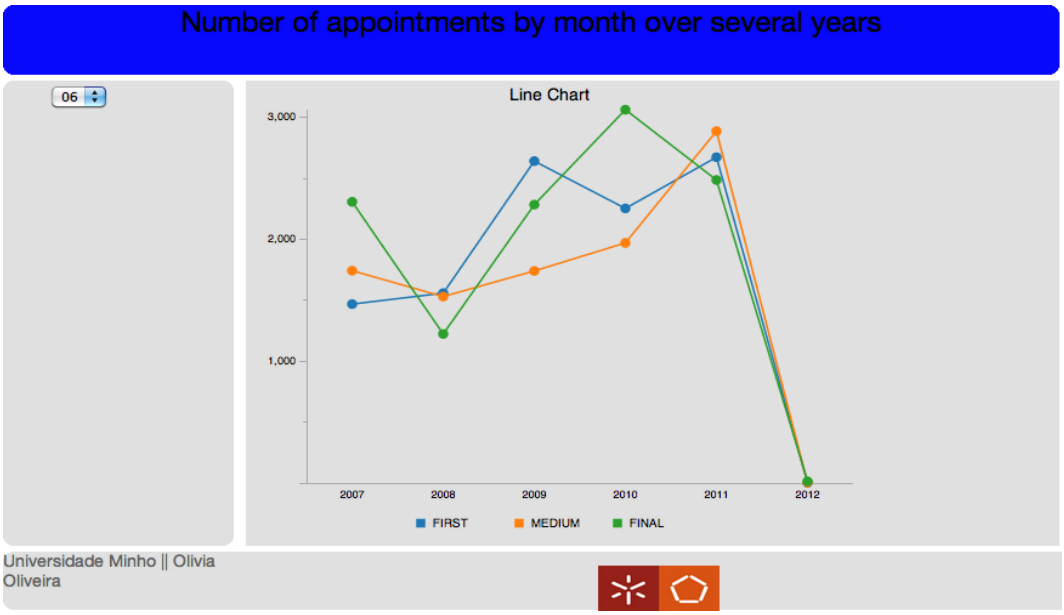


Figura A.35: Número de consultas em lista de espera durante vários anos no mês de Junho.

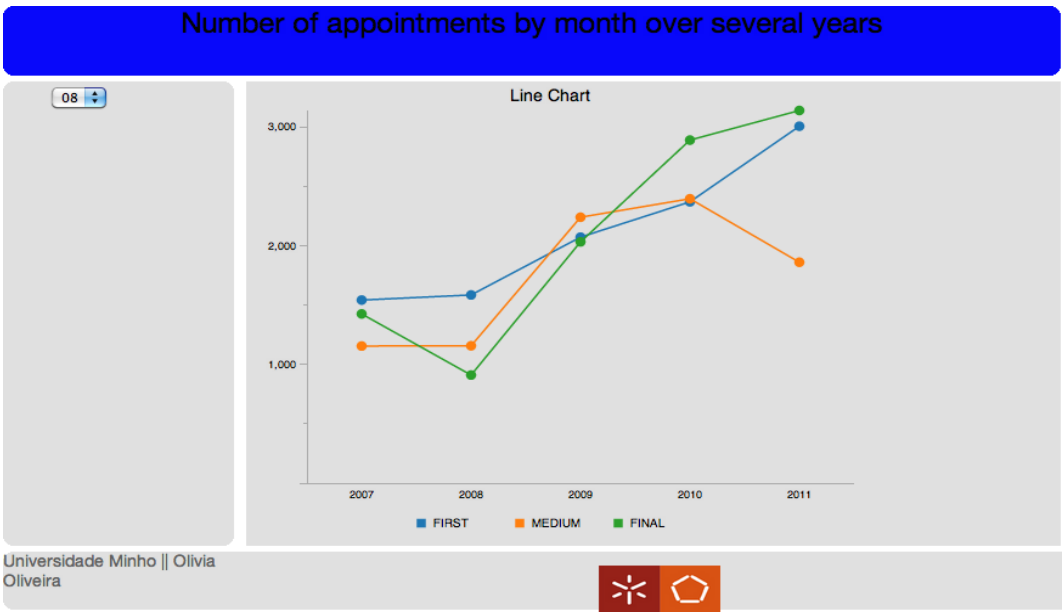


Figura A.36: Número de consultas em lista de espera durante vários anos no mês de Agosto.

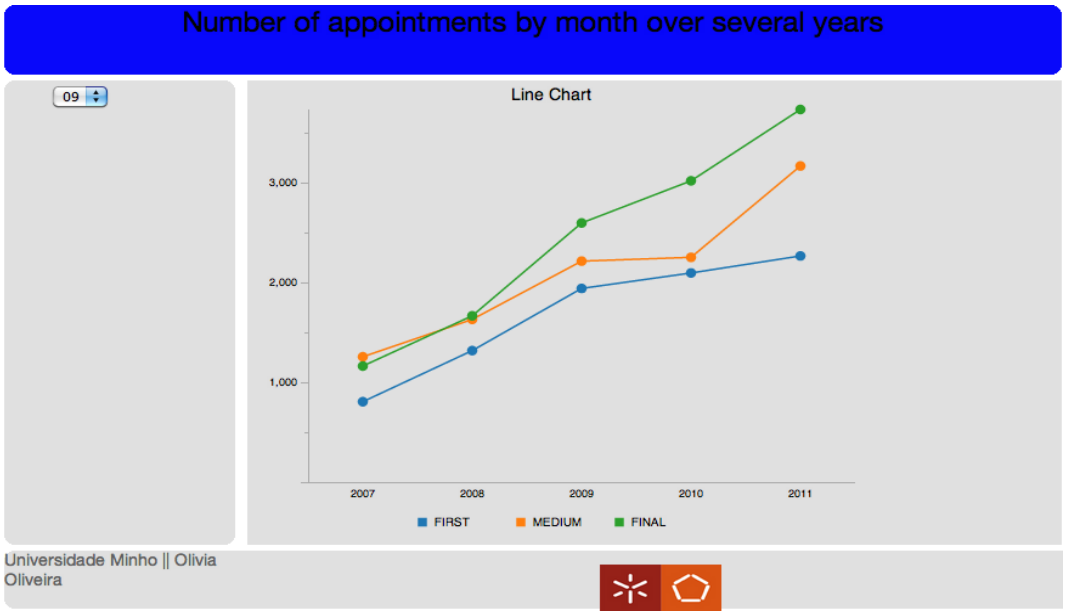


Figura A.37: Número de consultas em lista de espera durante vários anos no mês de Setembro.

Figura A.38: Número de consultas em lista de espera durante vários anos no mês de Outubro.

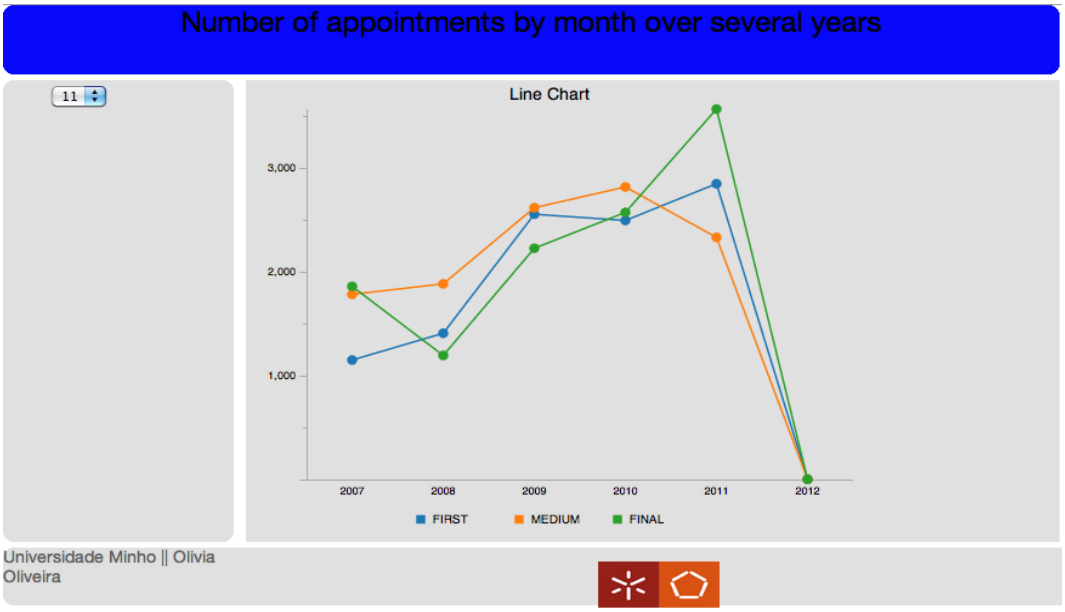


Figura A.39: Número de consultas em lista de espera durante vários anos no mês de Novembro.

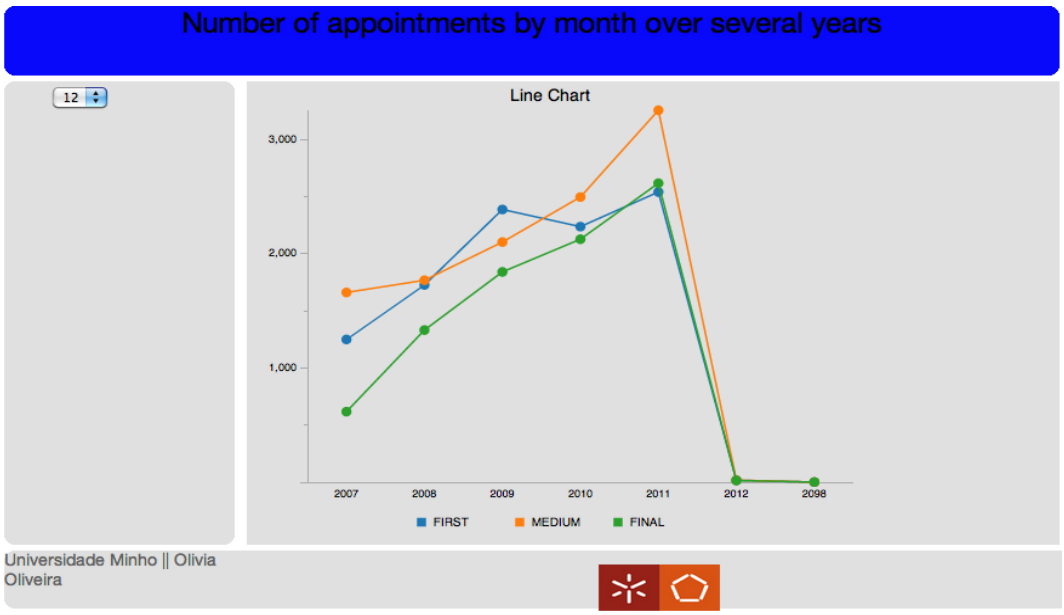


Figura A.40: Número de consultas em lista de espera durante vários anos no mês de Dezembro.

A.9 Lista de Espera para Consulta por Distritos

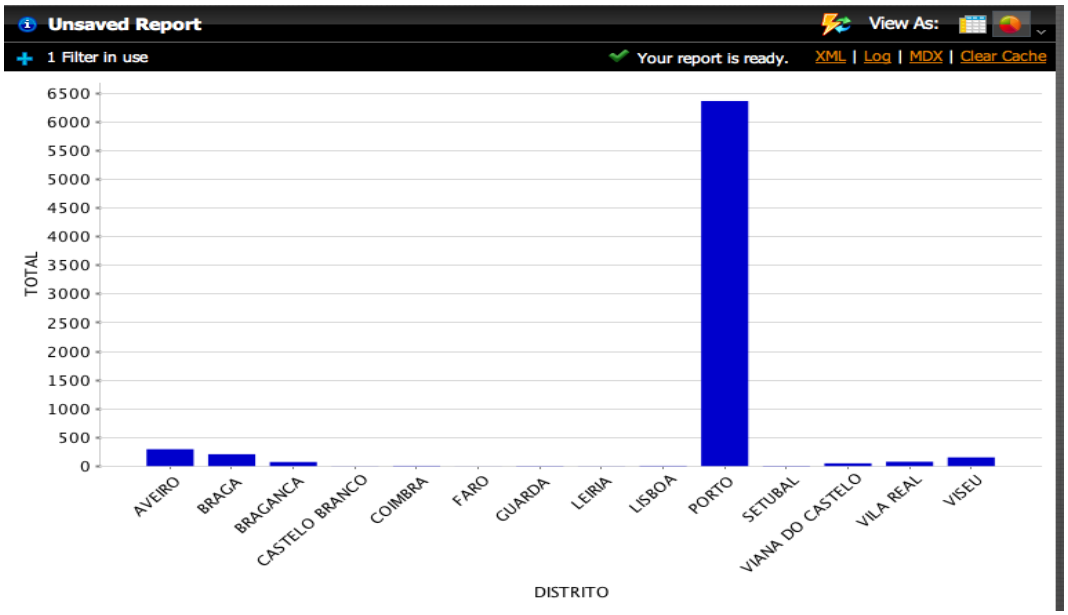


Figura A.41: Gráfico de barras verticais com o número de pacientes em lista de espera para consulta por distritos.

Listas de Espera para Consulta por Distritos

PAIS	DISTRITO	TOTAL
PORTUGAL	AVEIRO	294
PORTUGAL	BRAGA	207
PORTUGAL	BRAGANCA	71
PORTUGAL	CASTELO BRANCO	2
PORTUGAL	COIMBRA	6
PORTUGAL	FARO	1
PORTUGAL	GUARDA	4
PORTUGAL	ILHA DA MADEIRA	1
PORTUGAL	ILHA TERCEIRA	1
PORTUGAL	LEIRIA	2
PORTUGAL	LISBOA	6
PORTUGAL	PORTO	6371
PORTUGAL	SETUBAL	5
PORTUGAL	VIANA DO CASTELO	48
PORTUGAL	VILA REAL	76
PORTUGAL	VISEU	152

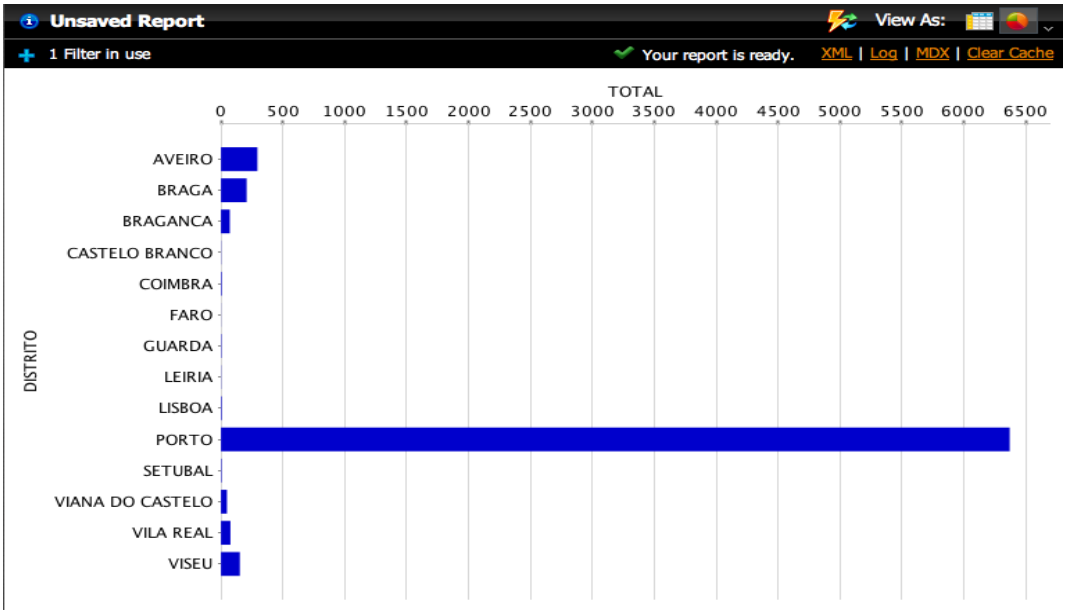


Figura A.42: Gráfico de barras horizontais com o número de pacientes em lista de espera para consulta por distritos.

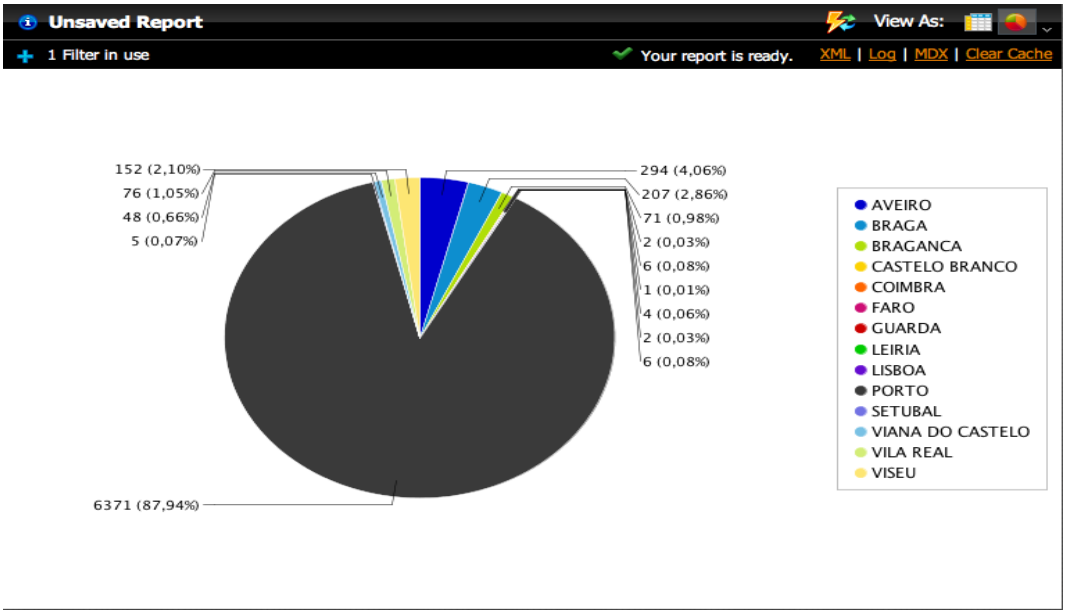


Figura A.43: Gráfico circular com o número de pacientes em lista de espera para consulta por distritos.

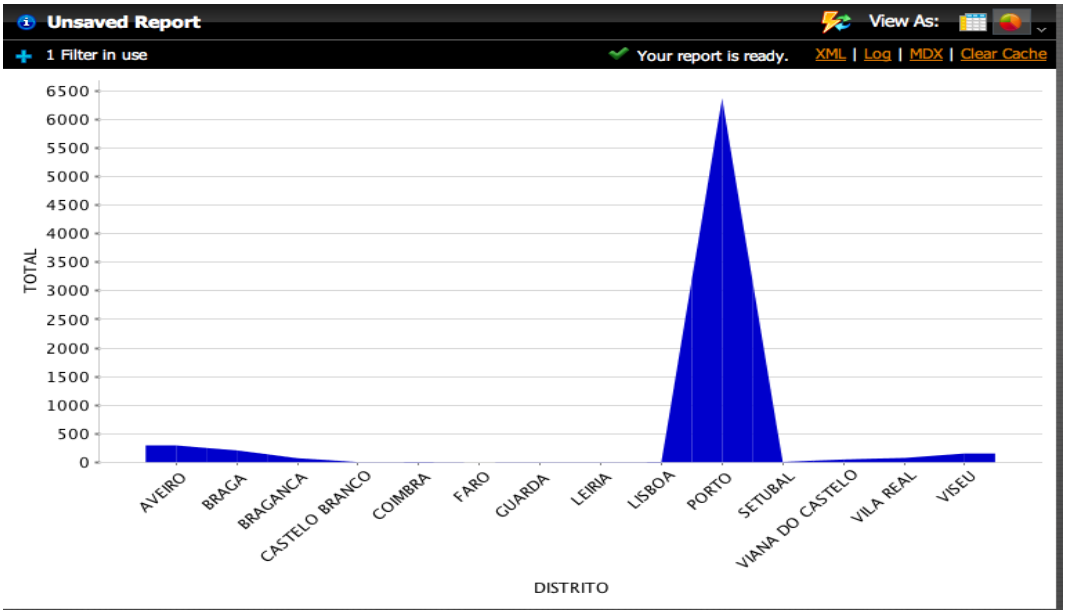


Figura A.44: Gráfico de área com o número de pacientes em lista de espera para consulta por distritos.

A.10 Lista de Espera para Consulta por concelhos do distrito do Porto

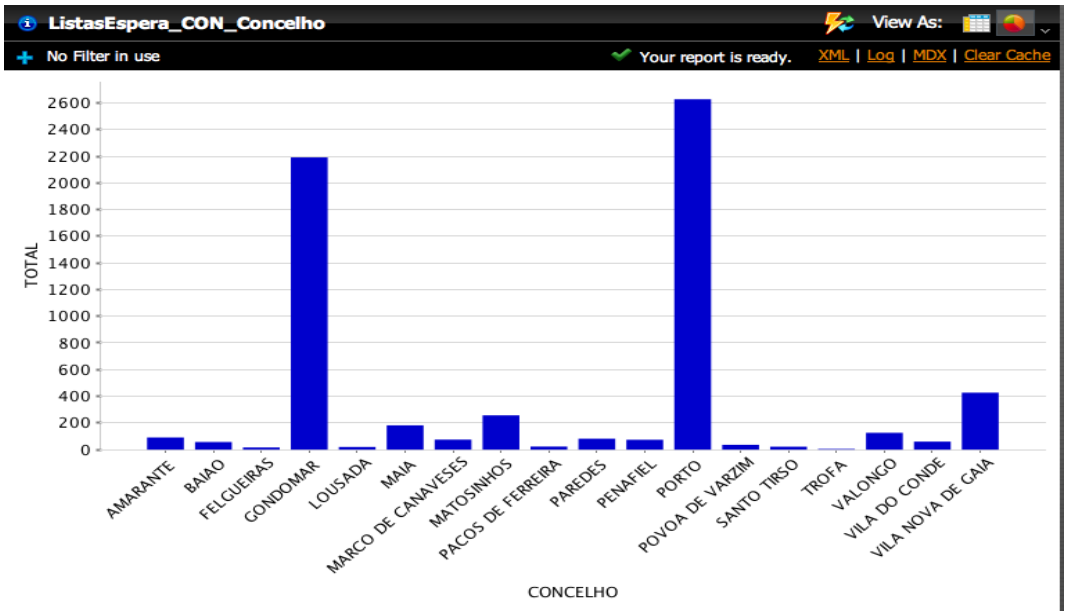


Figura A.45: Gráfico de barras verticais com o número de pacientes em lista de espera para consulta por concelhos do distrito do Porto.

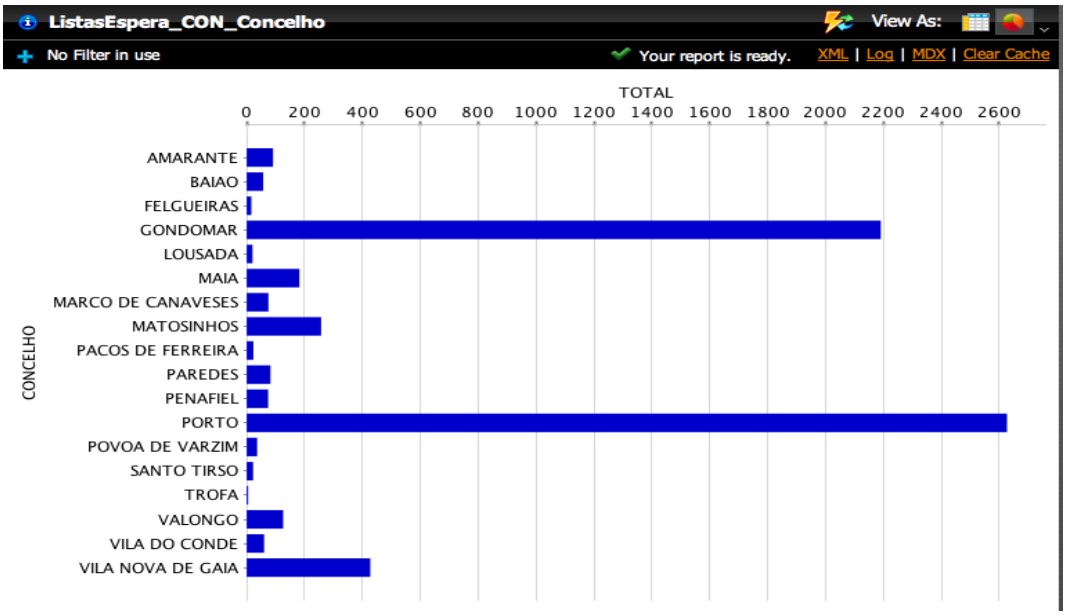


Figura A.46: Gráfico de barras horizontais com o número de pacientes em lista de espera para consulta por concelhos do distrito do Porto.

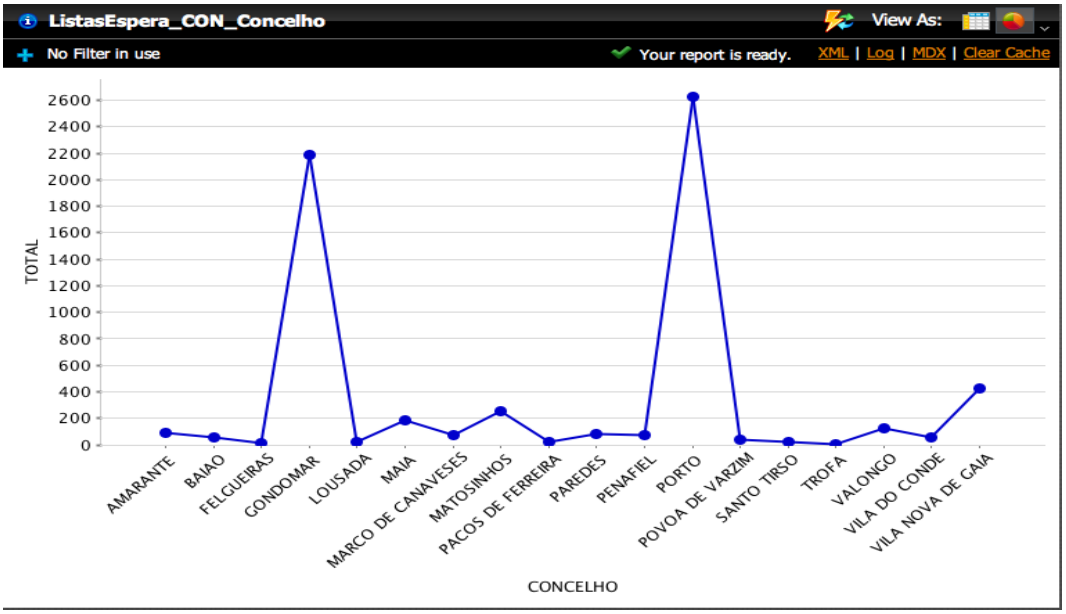


Figura A.47: Gráfico linear, representando o número de pacientes em lista de espera para consulta por distritos.

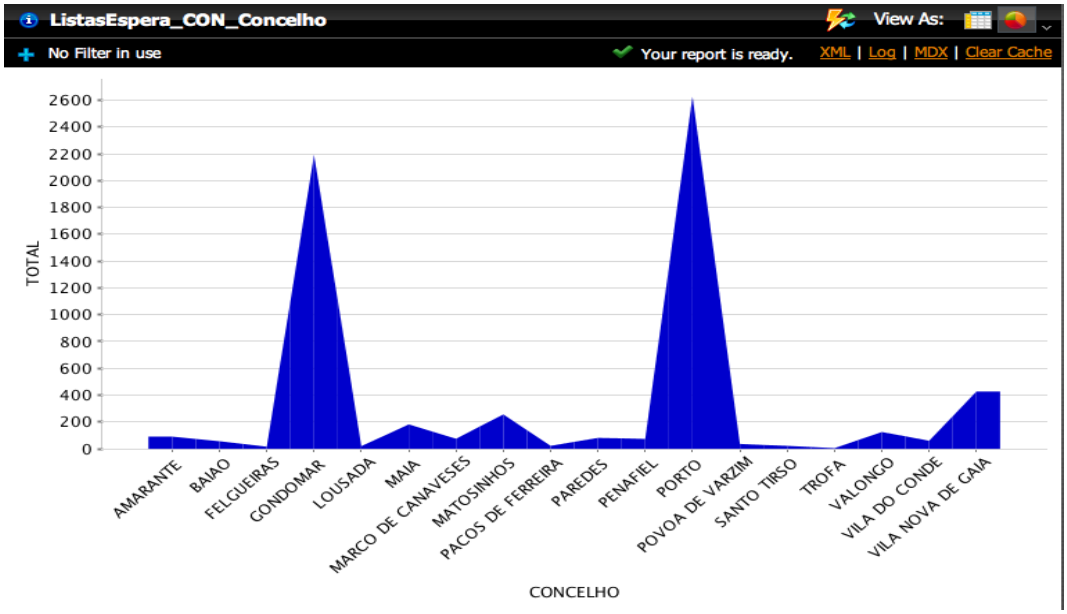


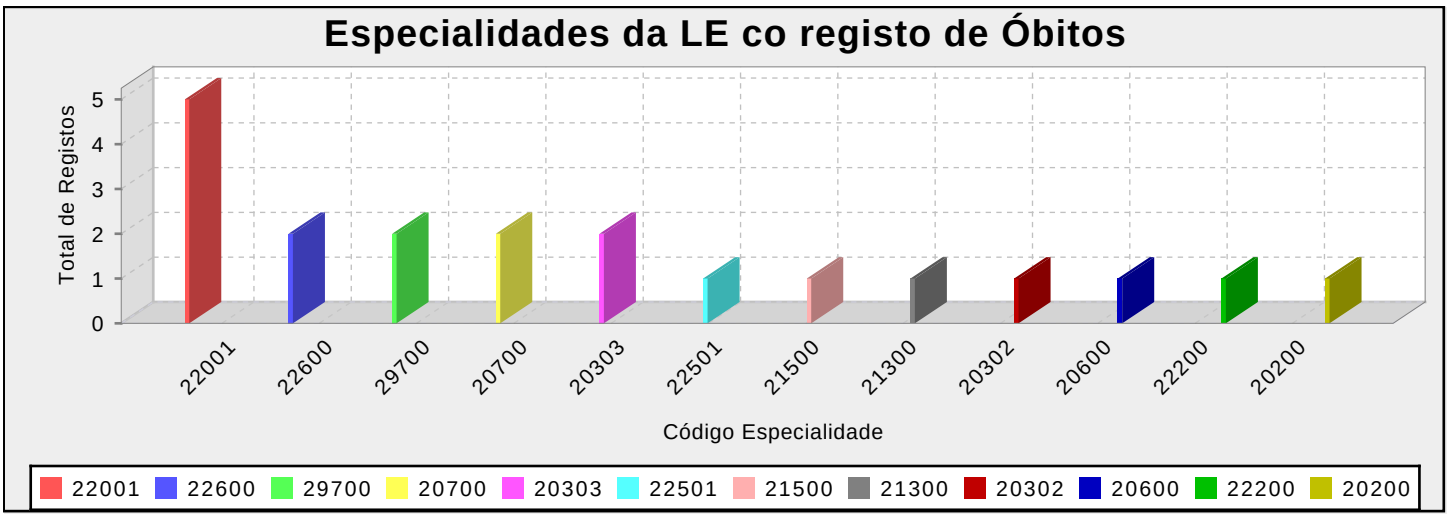
Figura A.48: Gráfico de área com o número de pacientes em lista de espera para consulta por concelhos do distrito do Porto.

A.11 Registo de Óbitos em Lista de Espera para Consulta

A.11.1 Relatório desenvolvido de raiz no PRD

Obitos em Lista de Espera para Consulta (por Especialidade)

CODIGO	ESPECIALIDADE	TOTAL
22.001	CE CENTRO OFTALMOLOGICO ..	5
22.600	CE UROLOGIA /HSA	2
29.700	CE CIRURGIA AMBULATORIO /HSA	2
20.700	CE ENDOCRINOLOGIA ..	2
20.303	CE CIRURGIA 2B /HSA	2
22.501	CE PSIQUIATRIA (ANT.P) /HSA	1
21.500	CE NEFROLOGIA /HSA	1
21.300	CE FISIATRIA - GERAL /HSA	1
20.302	CE CIRURGIA 2A /HSA	1
20.600	CE DERMATOLOGIA /HSA	1
22.200	CE O.R.L. /HSA	1
20.200	CE CARDIOLOGIA /HSA	1



Desenvolvido em Pentaho Report Designer

A.11.2 Relatório Wizard desenvolvido no PRD

Registo de Óbitos em Lista de Espera para Consulta

CODIGO	ESPECIALIDADE	TOTAL
22.001	CE CENTRO OFTALMOLOGICO /HSA	5
22.600	CE UROLOGIA /HSA	2
29.700	CE CIRURGIA AMBULATORIO /HSA	2
20.700	CE ENDOCRINOLOGIA /HSA	2
20.303	CE CIRURGIA 2B /HSA	2
22.501	CE PSIQUIATRIA (ANT.P) /HSA	1
21.500	CE NEFROLOGIA /HSA	1
21.300	CE FISIATRIA - GERAL /HSA	1
20.302	CE CIRURGIA 2A /HSA	1
20.600	CE DERMATOLOGIA /HSA	1
22.200	CE O.R.L. /HSA	1
20.200	CE CARDIOLOGIA /HSA	1

A.12 Óbitos em Lista de Espera para Consulta - Análise DM

=== Run information ===

Scheme: weka.associations.PredictiveApriori -N 100 -c -1
Relation: NewRelation-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-weka.filters.unsupervised.attribute.Remove-R1,3-11,13-37,39-47
Instances: 20
Attributes: 4
DES_ESPECIALIDADE
ESTADO
FALECEU_HOSP
DES_ESPECIALIDADE_1
=== Associator model (full training set) ===

PredictiveApriori
==========

Best rules found:

1. FALECEU_HOSP=S 12 ==> ESTADO=P 12 acc:(0.99479)
2. FALECEU_HOSP=N 8 ==> ESTADO=P 8 acc:(0.99412)
3. DES_ESPECIALIDADE=CE CENTRO OFTALMOLOGICO /HSA 5 ==> ESTADO=P 5 acc:(0.99108)
4. DES_ESPECIALIDADE=CE CIRURGIA 2B /HSA 2 ==> ESTADO=P FALECEU_HOSP=S 2 acc:(0.97141)
5. DES_ESPECIALIDADE=CE CIRURGIA 2B /HSA 2 ==> ESTADO=P DES_ESPECIALIDADE_1=INT GASTRENTEROLOGIA /HSA 2 acc:(0.97141)
…
26. ESTADO=P 20 ==> DES_ESPECIALIDADE=CE CENTRO OFTALMOLOGICO /HSA FALECEU_HOSP=S 3 acc:(0.13688)
27. ESTADO=P 20 ==> DES_ESPECIALIDADE=CE CIRURGIA AMBULATORIO /HSA FALECEU_HOSP=N 2 acc:(0.09637)
28. ESTADO=P 20 ==> DES_ESPECIALIDADE=CE CENTRO OFTALMOLOGICO /HSA FALECEU_HOSP=N 2 acc:(0.09637)

=== Run information ===

Scheme:weka.classifiers.trees.RandomTree -K 0 -M 1.0 -S 1
Relation: NewRelation-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-weka.filters.unsupervised.attribute.Remove-R1,3-11,13-17,19-37,39-47
Instances: 20
Attributes: 5
DES_ESPECIALIDADE

ESTADO
COD_PATOLOGIA
FALECEU_HOSP
DES_ESPECIALIDADE_1
Test mode:evaluate on training data

=== Classifier model (full training set) ===

RandomTree
==========

DES_ESPECIALIDADE = CE CARDIOLOGIA /HSA : S (1/0)
DES_ESPECIALIDADE = CE CENTRO OFTALMOLOGICO /HSA : S (5/2)
DES_ESPECIALIDADE = CE CIRURGIA 2A /HSA : N (1/0)
DES_ESPECIALIDADE = CE CIRURGIA 2B /HSA : S (2/0)
DES_ESPECIALIDADE = CE CIRURGIA AMBULATORIO /HSA : N (2/0)
DES_ESPECIALIDADE = CE DERMATOLOGIA /HSA : S (1/0)
DES_ESPECIALIDADE = CE ENDOCRINOLOGIA /HSA : S (2/0)
DES_ESPECIALIDADE = CE FISIATRIA - GERAL /HSA : S (1/0)
DES_ESPECIALIDADE = CE NEFROLOGIA /HSA : N (1/0)
DES_ESPECIALIDADE = CE O.R.L. /HSA : N (1/0)
DES_ESPECIALIDADE = CE PSIQUIATRIA (ANT.P) /HSA : S (1/0)
DES_ESPECIALIDADE = CE UROLOGIA /HSA : N (2/1)

Size of the tree : 13
Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances	17	85 %
Incorrectly Classified Instances	3	15 %
Kappa statistic	0.6809	
Mean absolute error	0.17	
Root mean squared error	0.2915	
Relative absolute error	35.283 %	
Root relative squared error	59.5017 %	
Total Number of Instances	20	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.75	0.083	0.857	0.75	0.8	0.943	N
	0.917	0.25	0.846	0.917	0.88	0.943	S
Weighted Avg.	0.85	0.183	0.851	0.85	0.848	0.943	

=== Confusion Matrix ===

```
  a  b   <-- classified as
  6  2 |  a = N
  1 11 |  b = S
```

=== Run information ===

Scheme:weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"
Relation: NewRelation-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-weka.filters.unsupervised.attribute.Remove-R1,3-11,13-17,19-37,39-47
Instances: 20
Attributes: 5
DES_ESPECIALIDADE
ESTADO
COD_PATOLOGIA
FALECEU_HOSP
DES_ESPECIALIDADE_1
Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

Decision Table:

Number of training instances: 20
Number of Rules : 10
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 11
Merit of best subset found: 100
Evaluation (for feature selection): CV (leave one out)
Feature set: 5,4

Time taken to build model: 0.01 seconds

=== Evaluation on test split ===
=== Summary ===

```
Correctly Classified Instances          6               85.7143 %
Incorrectly Classified Instances        1               14.2857 %
Kappa statistic                          0.5882
Mean absolute error                      0.2881
Root mean squared error                  0.3009
Relative absolute error                 48.0159 %
Root relative squared error             47.3078 %
Total Number of Instances                7

=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.5        0.833     1         0.909      1         N
                 0.5       0          1         0.5       0.667      1         S
Weighted Avg.    0.857     0.357      0.881     0.857     0.84       1

=== Confusion Matrix ===

 a b   <-- classified as
 5 0 | a = N
 1 1 | b = S
```

```
=== Run information ===

Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance
-R first-last\""
Relation:     NewRelation-weka.filters.unsupervised.attribute.Discretize-B10-M-
1.0-Rfirst-last-weka.filters.unsupervised.attribute.Remove-R1,3-11,13-17,19-
37,39-47
Instances:    20
Attributes:   5
              DES_ESPECIALIDADE
              ESTADO
              COD_PATOLOGIA
              FALECEU_HOSP
              DES_ESPECIALIDADE_1
Test mode:evaluate on training data

=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
```

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances	18	90 %
Incorrectly Classified Instances	2	10 %
Kappa statistic	0.7826	
Mean absolute error	0.1191	
Root mean squared error	0.223	
Relative absolute error	24.7212 %	
Root relative squared error	45.5084 %	
Total Number of Instances	20	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.75	0	1	0.75	0.857	1	N
	1	0.25	0.857	1	0.923	1	S
Weighted Avg.	0.9	0.15	0.914	0.9	0.897	1	

=== Confusion Matrix ===

a	b	<-- classified as
6	2	a = N
0	12	b = S

A.13 Número de Consultas em Lista de Espera Anual - Deteção de Falhas (Análise DM)

=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation: NewRelation
Instances: 7669
Attributes: 1
TO_CHAR(DTA_RETORNO,'YYYY')
Test mode:evaluate on training data

=== Model and evaluation on training set ===

kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 1051.0
Missing values globally replaced with mean/mode

Cluster centroids:

	Cluster#		
Attribute	Full Data	0	1
	(7669)	(1592)	(6077)
TO_CHAR(DTA_RETORNO,'YYYY')	2011	2007	2011

Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 1592 (21%)
1 6077 (79%)

----------Utilizando a distância de Manhattan e definindo 5 clusters----------
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 5 -A "weka.core.ManhattanDistance -R first-last" -I 500 -S 10
Relation: NewRelation
Instances: 7669
Attributes: 1
TO_CHAR(DTA_RETORNO,'YYYY')
Test mode:evaluate on training data

=== Model and evaluation on training set ===

kMeans
======

Number of iterations: 2
Sum of within cluster distances: 464.0
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (7669)	Cluster# 0 (1005)	1 (6077)	2 (108)	3 (215)	4 (264)
TO_CHAR(DTA_RETORNO,'YYYY')	2011	2007	2011	2012	2008	2009

Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===
Clustered Instances

0 1005 (13%)
1 6077 (79%)
2 108 (1%)
3 215 (3%)
4 264 (3%)

=== Run information ===

Scheme:weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100
Relation: NewRelation
Instances: 7669
Attributes: 1
TO_CHAR(DTA_RETORNO,'YYYY')
Test mode:evaluate on training data

=== Model and evaluation on training set ===

EM
==

Number of clusters selected by cross validation: 1

Attribute	Cluster 0 (1)
TO_CHAR(DTA_RETORNO,'YYYY')	

2007	542
2008	216
2009	265
2010	322
2011	6078
2012	109
2013	4
2014	4
2015	8
2016	2
2017	3
2018	2
2019	2
2020	3
2021	4
2022	3
2023	2
2024	3
2025	4
2026	2
2027	3
2028	3
2029	2
2032	2
2033	2
2036	2
2040	2
2041	3
2044	4
2045	2
2047	7
2052	5
2054	2
2055	4
2056	2
2058	2
2063	5
2065	6
2066	2
2069	8
2070	4
2074	2
2077	2
2078	3
2079	2
2080	2
2085	3
2087	4
2088	2

2089	8
2090	5
2092	2
2093	2
2096	3
2098	4
2099	4
2100	2
2101	2
2111	2
2112	3
2201	3
2202	2
2203	7
2204	2
2205	2
2206	2
2207	2
2208	3
2209	2
2300	2
2993	2
3001	2
3003	2
3012	3
[total]	7743

Time taken to build model (full training data) : 4.72 seconds

=== Model and evaluation on training set ===
Clustered Instances
0 7669 (100%)

Log likelihood: -0.93106

=== Run information ===

Scheme:weka.clusterers.FarthestFirst -N 2 -S 1
Relation: NewRelation
Instances: 7669
Attributes: 1
TO_CHAR(DTA_RETORNO,'YYYY')
Test mode:evaluate on training data

=== Model and evaluation on training set ===

FarthestFirst
==========

Cluster centroids:

Cluster 0
        2011
Cluster 1
        2055

Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===
Clustered Instances

0   7666 (100%)
1   3 (0%)

=== Run information ===

Scheme:weka.clusterers.DBScan -E 0.9 -M 6 -I
weka.clusterers.forOPTICSAndDBScan.Databases.SequentialDatabase -D
weka.clusterers.forOPTICSAndDBScan.DataObjects.EuclidianDataObject
Relation: NewRelation
Instances: 7669
Attributes: 1
        TO_CHAR(DTA_RETORNO,'YYYY')
Test mode:evaluate on training data

=== Model and evaluation on training set ===

DBScan clustering results
==========
Clustered DataObjects: 7669
Number of attributes: 1
Epsilon: 0.9; minPoints: 6
Index: weka.clusterers.forOPTICSAndDBScan.Databases.SequentialDatabase
Distance-type:
weka.clusterers.forOPTICSAndDBScan.DataObjects.EuclidianDataObject
Number of generated clusters: 11
Elapsed time: ,46

(0.) 2011   --> 0
(1.) 2011   --> 0
(2.) 2011   --> 0
(3.) 2011   --> 0
(4.) 2055   --> NOISE
(5.) 2011   --> 0
(6.) 2011   --> 0
(7.) 2011   --> 0

(8.) 2011 --> 0
(9.) 2011 --> 0
(10.) 2011 --> 0

...

(7663.) 2012 --> 2
(7664.) 2012 --> 2
(7665.) 2012 --> 2
(7666.) 2012 --> 2
(7667.) 2012 --> 2
(7668.) 2012 --> 2

Time taken to build model (full training data) : 0.46 seconds

=== Model and evaluation on training set ===
Clustered Instances

0 6077 (80%)
1 6 (0%)
2 108 (1%)
3 321 (4%)
4 215 (3%)
5 541 (7%)
6 7 (0%)
7 7 (0%)
8 6 (0%)
9 7 (0%)
10 264 (3%)

Unclustered instances : 110

A.14 Utilização do Bloco Operatório - Análise DM

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: NewRelation
Instances: 995
Attributes: 5
MES
DIA
SALAOPERATORIA
TEMPO5
INI5
Test mode:evaluate on training data

=== Classifier model (full training set) ===
J48 pruned tree

DIA = 01
| MES = 01
| | INI5 <= 20.4: SALA-F BLOCO CENTRAL (5.0/2.0)
| | INI5 > 20.4: SALA 1 OBST MJD (2.0/1.0)
| MES = 02
| | TEMPO5 <= 1.2
| | | INI5 <= 12.9
| | | | INI5 <= 12.2
| | | | | TEMPO5 <= 0.4: SALA 2 OBST MJD (2.0/1.0)
| | | | | TEMPO5 > 0.4: ESTOMAT - AMB (3.0/1.0)
| | | | INI5 > 12.2: SALA B MPIA (2.0/1.0)
| | | INI5 > 12.9
| | | | TEMPO5 <= 0.7
| | | | | INI5 <= 18: SALA 1 GINEC MJD (3.0)
| | | | | INI5 > 18: SALA-A B.ED.CLASSICO (2.0)
| | | | TEMPO5 > 0.7: S. BLOCO S.FRANCISCO (3.0/2.0)
| | TEMPO5 > 1.2
| | | TEMPO5 <= 2
| | | | TEMPO5 <= 1.7
| | | | | TEMPO5 <= 1.3: SALA 2 OBST MJD (2.0/1.0)
| | | | | TEMPO5 > 1.3: SALA-B B.ED.CLASSICO (3.0/1.0)
| | | | TEMPO5 > 1.7
| | | | | TEMPO5 <= 1.8: SALA-F BLOCO CENTRAL (2.0/1.0)
| | | | | TEMPO5 > 1.8: SALA-C BLOCO CENTRAL (2.0)
| | | TEMPO5 > 2
| | | | INI5 <= 14.1: SALA-H BLOCO CENTRAL (5.0/3.0)
| | | | INI5 > 14.1: SALA-B BLOCO CENTRAL (3.0/2.0)
DIA = 02
| MES = 01
| | TEMPO5 <= 1.6
| | | TEMPO5 <= 0.7: SALA B MPIA (3.0/1.0)
| | | TEMPO5 > 0.7: SALA-G BLOCO CENTRAL (6.0/3.0)
| | TEMPO5 > 1.6

```
|   |   |   TEMPO5 <= 2.1: SALA 1 GINEC MJD (3.0/1.0)
|   |   |   TEMPO5 > 2.1: SALA 2 OBST MJD (6.0/4.0)
|   MES = 02
|   |   INI5 <= 10.8: SALA-B B.ED.CLASSICO (10.0/8.0)
|   |   INI5 > 10.8
|   |   |   TEMPO5 <= 1.7: SALA 2 OBST MJD (4.0)
|   |   |   TEMPO5 > 1.7
|   |   |   |   TEMPO5 <= 2.3: SALA-B B.ED.CLASSICO (2.0/1.0)
|   |   |   |   TEMPO5 > 2.3: SALA-B BLOCO CENTRAL (3.0/1.0)
|   |   |   |
...

Number of Leaves  :  308
Size of the tree :  586

Time taken to build model: 0.05 seconds

=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         554               55.6784 %
Incorrectly Classified Instances       441               44.3216 %
Kappa statistic                          0.5299
Mean absolute error                      0.04
Root mean squared error                  0.1414
Relative absolute error                 54.9568 %
Root relative squared error             74.1488 %
Total Number of Instances              995
```

```
=== Run information ===

Scheme:weka.clusterers.FarthestFirst -N 2 -S 1
Relation:     NewRelation
Instances:    995
Attributes:   5
              MES
              DIA
              SALAOPERATORIA
              TEMPO5
              INI5
Test mode:evaluate on training data

=== Model and evaluation on training set ===
 FarthestFirst
==============
Cluster centroids:

Cluster 0
              01 10 SALA-C BLOCO CENTRAL 1.3 9.9
```

Cluster 1
02 29 SALA-N1 BL NEUROCIRU 10.8 8.5

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances
0 594 (60%)
1 401 (40%)

=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation: NewRelation
Instances: 995
Attributes: 5
MES
DIA
SALAOPERATORIA
TEMPO5
INI5
Test mode:evaluate on training data

=== Model and evaluation on training set ===
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 1874.1872117051516
Missing values globally replaced with mean/mode

Cluster centroids:

	Cluster#		
Attribute	Full Data	0	1
	(995)	(563)	(432)
MES	01	01	02
DIA	16	16	01
SALAOPERATORIA	SALA 2 OBST MJD	SALA 2 OBST MJD	SALA 2 OBST MJD
TEMPO5	1.7638	1.7179	1.8236
INI5	12.6515	12.3568	13.0354

Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances
0 563 (57%)
1 432 (43%)

=== Run information ===

Scheme:weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100
Relation: NewRelation
Instances: 995
Attributes: 5
MES
DIA
SALAOPERATORIA
TEMPO5
INI5
Test mode:evaluate on training data

=== Model and evaluation on training set ===
EM
==
Number of clusters selected by cross validation: 8

Attribute	Cluster 0 (0.06)	1 (0.1)	2 (0.15)	3 (0.06)	4 (0.23)	5 (0.04)	6 (0.18)	7 (0.19)
MES								
01	52.6478	56.7012	4.9036	2.5696	224.6212	23.4486	6.6183	192.4896
02	12.5768	40.2427	142.9527	56.5248	4.766	14	173.4217	2.5153
[total]	65.2246	96.9439	147.8563	59.0944	229.3872	37.4486	180.0401	195.0049
DIA								
01	1.3362	5.2497	11.167	9.4912	4.3986	2	12.3295	1.0279
02	3.7787	2.9934	7.4647	3.2885	9.813	3	7.9258	6.7358
03	2.4649	1.2584	9.8041	2.3126	15.7099	4	5.6211	12.829
04	6.1646	5.3476	1.7733	1.2051	9.0944	1	3.898	8.5169
05	3.0803	6.5708	1.0696	1.0495	16.8657	2	1.3432	9.0208
06	1.4614	3.3722	9.8199	2.4696	7.293	1	8.1201	7.4637
07	1.2308	2.9824	5.2052	4.8579	6.4764	2	7.1657	6.0816
08	1.2735	4.2361	9.4836	6.9938	3.0046	2	10.984	1.0244
09	3.2685	3.4497	5.7262	3.7567	9.6402	2	10.4578	8.7008
10	8.4007	3.461	6.0003	1.4435	16.7132	3	2.9501	9.0312
11	4.3086	2.0526	3.4653	1.5211	19.6053	2	5.0822	9.9649
12	4.7861	5.8887	1.0467	1.0648	10.0243	2	1.126	14.0633
13	1.2538	2.3675	10.6852	1.3619	9.9487	4	14.6595	4.7234
14	2.3326	5.869	5.8138	2.2741	3.9378	1	12.7423	1.0304

15 1.1165 5.8267 5.6117 2.9901 3.0938 2 4.1222 1.239
16 3.1084 8.6367 2.4864 3.7031 14.8833 2 17.9075
14.2744
17 2.9522 3.9038 10.5431 1.0393 7.2844 5 3.0304 8.2468
18 1.1961 4.6034 2.0995 2.0163 6.745 2 2.0399 16.2998
19 4.582 6.5577 1.2688 1.0626 8.8974 3 1.054 10.5774
20 2.7656 2.4614 4.8762 3.3584 8.5506 1 6.1824 1.8053
21 3.422 2.4459 8.3055 2.7727 2.9249 1 10.4532 3.6757
22 3.3862 3.6726 13.2133 5.1024 1.0309 3 4.5674 1.0272
23 3.6123 2.6708 8.697 8.1429 12.3708 1 11.8653 5.641
24 3.3066 3.6294 3.5117 2.116 6.8449 3 9.5442 8.0471
25 2.6229 5.3822 2.8118 1.0186 3.0744 2.4486 1.1779
10.4636
26 4.014 1.9561 4.2798 1.067 5.5154 1 2.6035 13.5641
27 2.9911 2.9553 8.3147 1.5784 9.3696 2 15.9788 2.8122
28 1.0898 3.2723 7.0778 2.8855 1.4571 3 2.8889 6.3285
29 3.1355 7.4947 3.1597 4.1047 1.9572 1 9.1338 1.0144
30 3.4374 3.5761 1.0396 1.0275 9.5289 1 1.0352 7.3553
31 2.3453 1.7996 1.0349 1.0184 12.3332 2 1.0498
11.4188
[total] 94.2246 125.9439 176.8563 88.0944 258.3872 66.4486
209.0401 224.0049
SALAOPERATORIA
ESTOMAT - AMB 1.0306 1.0004 1.0928 13.3109 1.1343 2 2.733
21.698
S. BLOCO S.FRANCISCO 1.044 1.0303 1.0348 7.0949 1.1357 2
1.8923 8.768
SALA 1 GINEC MJD 1.1906 1.0569 6.1983 14.977 6.3087 3 4.9798
21.2886
SALA 1 OBST MJD 1.0567 11.8205 1.1279 1.5126 1.7296 2 7.7717
5.981
SALA 2 OBST MJD 1.4228 5.5667 4.0447 4.676 19.5775 8 33.1653
32.547
SALA A MPIA 1.1235 1.3505 2.2466 3.3128 3.7783 3 11.6275
11.5607
SALA B MPIA 3.0303 1.4471 4.8305 1.7079 8.1488 1 9.4889
16.3465
SALA C MPIA 1.1544 1.2992 1.069 1.6676 2.0903 1 5.9722
5.7474
SALA D B.ED.CLASSICO 1.0311 1.0011 1.027 1.9092 1.2478 1
1.0329 1.7509
SALA PMA - MJD 1.0383 1.0005 1.093 8.6025 1.1593 1 3.7081
7.3984
SALA X MPIA 1.0471 1.5117 1.5638 1.0001 1.9529 1 1.4856
2.4388
SALA-A B.ED.CLASSICO 1.1977 4.6532 1.1224 2.4768 6.8137 1
9.7551 5.9811
SALA-A BLOCO CENTRAL 7.08 11.7747 14.3587 2.6181 24.9607 1
2.1767 5.031

SALA-B B.ED.CLASSICO 6.8449 1.5583 7.4504 1.1841 20.5657 2
23.5784 8.8182
SALA-B BLOCO CENTRAL 1.4426 1.3487 25.1136 1.5644 20.9828 2.4486
3.3021 4.7972
SALA-C B.ED.CLASSICO 1.2094 1.1509 2.4408 1.0243 4.6815 1
2.4771 6.0159
SALA-C BLOCO CENTRAL 2.9876 1.2918 14.7107 3.5369 20.9334 3
13.683 14.8566
SALA-D BLOCO CENTRAL 5.0735 2.2241 15.3907 1.0068 18.2406 3
2.8597 7.2047
SALA-E BLOCO CENTRAL 4.3126 23.2603 9.7744 1.1383 16.4901 4
15.1739 3.8504
SALA-F BLOCO CENTRAL 3.9914 35.1051 6.3819 1.9703 24.3361 3
15.5717 5.6435
SALA-G BLOCO CENTRAL 12.9728 4.1751 9.6423 1.154 10.2426 5
7.4106 4.4025
[total] 89.2246 120.9439 171.8563 83.0944 253.3872 61.4486
204.0401 219.0049
TEMPO5
mean 4.6845 1.6228 2.9098 0.4788 2.0757 0 1.2771 0.796
std. dev. 2.0529 0.8098 1.0146 0.205 0.8006 1.3664 0.376
0.3317
INI5
mean 11.2883 20.0623 11.7866 13.1067 11.6285 0.0025 13.1692
13.0222
std. dev. 3.2871 2.3894 3.3131 3.4222 3.8026 0.0224 3.5378
3.0279

Time taken to build model (full training data) : 38.77 seconds

=== Model and evaluation on training set ===

Clustered Instances
0 49 (5%)
1 90 (9%)
2 142 (14%)
3 57 (6%)
4 227 (23%)
5 35 (4%)
6 187 (19%)
7 208 (21%)

Log likelihood: -10.87106