Aplicação do processo de KDD em uma gestora de planos de saúde

Lucas Carvalho de Paula
Graduado em Sistemas de Informação pela Libertas Faculdades Integradas
Ely Fernando do Prado
Mestrando em Sistemas de Informação e professor da Libertas Faculdades Integradas

RESUMO
O data mining, faz parte de um processo muito maior que é o KDD (Knowledge Discovery in Databases), o qual pode ser usado para diversas finalidades, entre uma das principais é a clusterização ou agrupamento. Foi aplicada essa técnica para separação dos usuários de uma gestora de plano de saúde, de acordo com a utilização de serviços de cada um em um determinado período. Após esse processo os clusters (grupos) descobertos foram analisados e extraídos deles o conhecimento. O qual foi usado para elaboração de medidas de medicina preventiva para economia de gastos da gestora e conservação da saúde dos usuários.

Palavras chave: Data mining, KDD, WEKA, planos de saúde, K-means, Oracle, clusterização, reconhecimento de padrões,

1 – INTRODUÇÃO

As gestoras de planos de saúde estão sempre buscando maneiras eficazes de prevenção de gastos. Isso pode ser feito de várias maneiras, por exemplo: cortes no orçamento, eliminar serviços da cobertura do usuário, medidas de medicina preventiva, entre outras. Sendo que aplicar medidas de medicina preventiva tem o melhor custo benefício, tanto do lado econômico quanto humanitário (ROSE, 2010).

Em uma pesquisa realizada pelo Instituto de Estudos de Saúde Suplementar (IESS), foi usado nessa pesquisa a Variação dos Custos Médico-hospitalares (VCMH), que engloba todos os gastos de um paciente gera. E este índice foi comparado com o Índice Nacional de Preços ao Consumidor Amplo (IPCA) que mede a inflação acumulada em um período de 12 meses, o resultado está apresentado no gráfico 1, conforme CECHIN; MARTINS & LEITE (2009).

É possível perceber que o VCMH sofreu reajustes maiores que o IPCA, conclui-se que o valor médio do VCMH está muito defasado em relação a inflação.

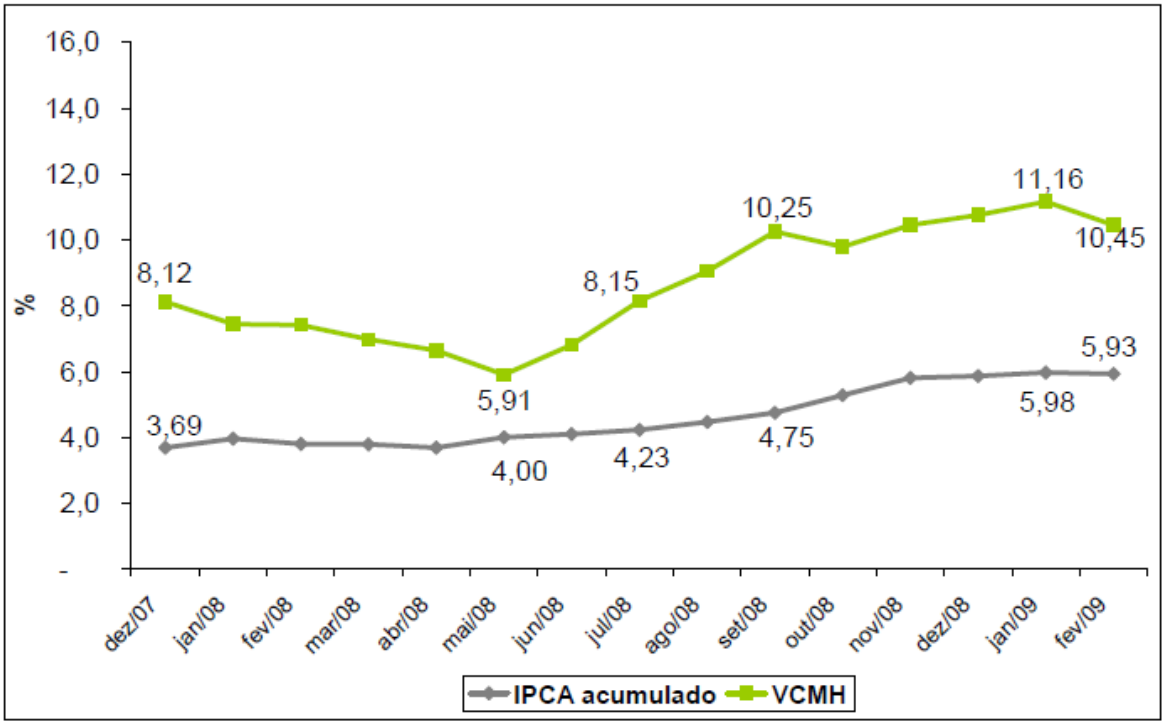

Gráfico 1 - Gráfico comparaivo IPCA X VCMH
Fonte: http://www.iess.org.br/html/TD00252009VCMH.pdf

2 – REFERENCIAL TEÓRICO

2.1 – Dados, banco de dados e SGBD

Segundo Carneiro (2004, p. 4): "(dados)... são fatos fornecidos que descrevem uma característica de um objeto ou evento de mundo real." Um dado é qualquer característica que descreva um fato ou objeto, uma informação surge quando um dado é inserido num determinado contexto. O banco de dados substitui o antigo sistema de armazenamento de informações: as fichas, fichários e arquivos. Um banco de dados é: "Um conjunto de dados relacionados entre si armazenados segundo uma determinada lógica de forma para que possam ser recuperados quando necessário." Carneiro (2004, p. 4). Com o passar do tempo foi desenvolvido o Sistema Gerenciador de banco de dados (SGBD) que auxiliam na gerência e manutenção dos dados.

2.2 – Oracle

O Oracle é um SGBD que foi desenvolvido no final da década de 70 pela Oracle Corporation. O Oracle foi um dos primeiros bancos de dados relacionais do mercado, antes disso havia os modelos hierárquico e em rede, nesses modelos o armazenamento dos dados era pensado de acordo com o armazenamento físico. De acordo com o manual *PL/SQL User's Guide and Reference* (2005, p. 19), a linguagem PL/SQL é utilizada para consulta no Oracle, que possui forte ligação a linguagem SQL.

2.3 – KDD – Descoberta de conhecimento em base de dados

Uma das melhores definições para esse processo foi feita por Fayyad (1996): "Knowledge Discovery in Databases é um processo não trivial de identificar padrões válidos, originais, potencialmente úteis e compreensíveis em determinados bancos de dados". Alguns estudos usam a nomenclatura KDD e data mining como sinônimos, mas o data mining é uma etapa do processo de KDD. O KDD pode ser divido em etapas, podendo ser três ou cinco dependendo do autor. Este trabalho adotou a divisão em cinco etapas que segundo Fayyad (1996) são: Seleção, Pré-processamento, Formatação, Mineração de Dados (Data Mining) e Interpretação/avaliação. Cada parte do KDD tem suas características próprias, descritas a seguir:

Seleção: Define os dados que serão utilizados no processo e é nesta etapa que os dados são separados de acordo com a necessidade e objetivo do projeto.

Pré-processamento: Depois de selecionados os dados é preciso corrigir possíveis erros utilizando algumas técnicas, conforme Figueira (1998) são: padronização, remoção de duplicidade, eliminação de ruídos, preenchimento ou exclusão de ausentes.

Formatação: Nesta etapa os dados são convertidos ou migrados para o formato que a ferramenta escolhida utilize.

Mineração de dados (data mining):O processamento de todos os dados selecionados ocorre nessa fase do processo, as técnicas (algoritmos) são aplicadas a eles de acordo com o objetivo desejado. Alguns exemplos de técnicas usadas para data mining, segundo Goldschmit & Passos(2005, cap. 4): descoberta de associações, descoberta de sequências, classificação, sumarização, clusterização, previsão de séries temporais, entre outras.

Interpretação/avaliação: Esta avaliação deve ser feita pelos envolvidos na extração dos dados e especialistas no resultado a que se quer chegar.

Exemplos de aplicação do KDD: segue abaixo alguns casos de sucesso na aplicação do data mining, conforme Loss & Rabelo (2004, p. 7-8):

> "A rede americana Wall-Mart, pioneira no uso de *Data Mining*, descobriu ao explorar seus números que 60% das mães que compram boneca Barbie, levam também uma barra de chocolate.
>
> O SERPRO no Brasil, implantando o seu Data Warehouse e *Data Mining*, já consegue hoje cruzar e analisar informações em cinco minutos, o que antes demandavam quinze dias de trabalho.
>
> E o clássico exemplo de uma grande rede varejista americana (Wall-Mart) que descobriu, através de seu *Data Mining*,que as vendas de fraldas estavam intimamente ligadas às vendas de cerveja. Explicação, os pais que saiam à noite para comprar fraldas, compravam cerveja também."

2.4 – WEKA

O Waikato Environment for Knowledge Analysis (WEKA) começou a ser escrito em 1993, usando a linguagem de programação Java, na Universidade de Waikato, Nova Zelândia. O Weka é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. O Weka utiliza o padrão Attribute Relation File Format (ARFF) e tem a licença GNU General Public License. Fonte: http://www.cs.waikato.ac.nz/ml/weka/

3 - METODOLOGIA

Para o desenvolvimento e aplicação deste trabalho foram usadas três ferramentas (aplicativos) diferentes: SGBD Oracle, PL/SQL Developer e WEKA para o data mining;

Seleção: Juntamente com esse *select*, foi feita a etapa de pré-processamento dos dados, os casos de duplicidade, ruídos, omissão de dados e falta padronização dos dados, foram tratados. Os campos selecionados foram: idade, sexo, nome do serviço e a quantidade realizada deste serviço no período de um ano.

Processamento: Foi realizado juntamente com a seleção dos dados, não havendo a necessidade da segunda etapa, para um tratamento mais detalhado de campos com dados incorretos ou inválidos e a padronização das informações.

Transformação: O formato do arquivo de saída do *select* é "csv". No próprio WEKA existe a funcionalidade para conversão de arquivos "csv" para "arff".

Data mining: A técnica de clusterização atende às necessidades do trabalho, pois o problema de pesquisa necessita de vários grupos como resposta e não há uma quantidade definida nem um padrão para cada grupo. O algoritmo usado para a clusterização foi o K-means, que segundo Berry e Linoff (2004, p. 354) e Goldschmit & Passos (2005), é um dos mais usados para essa técnica. No WEKA está disponível o algoritmo SimpleKmeans, baseado no K-means.

Interpretação: O primeiro resultado está apresentado no gráfico 2. A cor dos pontos significa o tipo do sexo de cada instância (vermelho "F", azul "M"), o eixo x é a idade, partindo de 0 até 101 anos, e o eixo y a quantidade de vezes que um determinado usuário realizou um determinado procedimento no período de um ano. Foram feitos dois grupos os quais serão analisados no próximo capítulo.

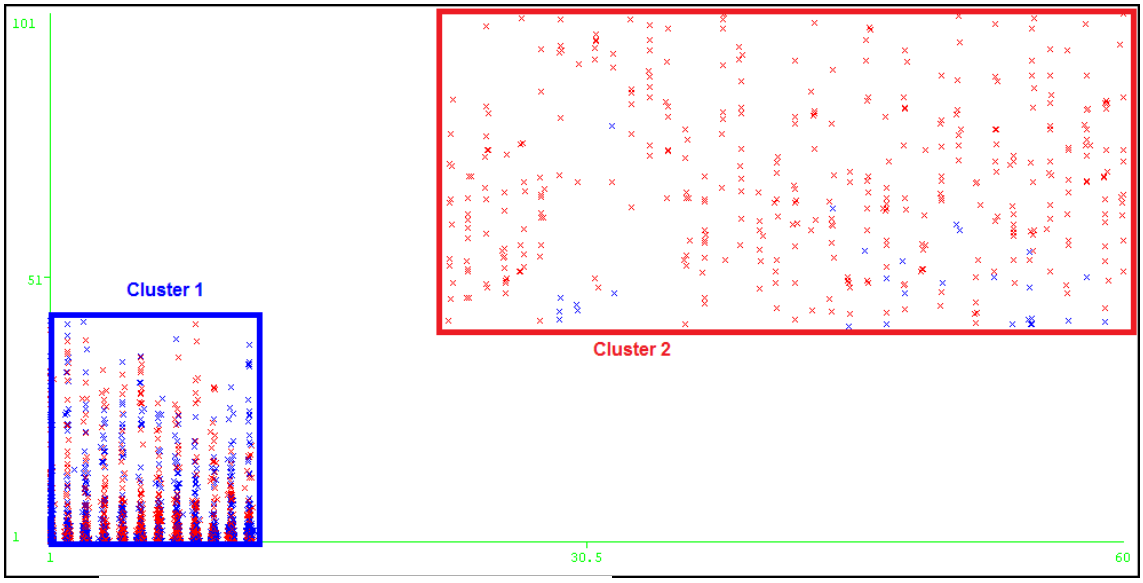

Gráfico 2 - Resultados pré-selecionados

4 – ANÁLISE DE RESULTADOS

Como mostrado no capítulo anterior nos gráfico 2, foram descobertos dois agrupamentos. Foi feita a análise juntamente com o departamento de marketing da gestora e chegou-se às seguintes conclusões, com base nesse primeiro resultado:

- No primeiro cluster estão usuários de 0 a 10 anos de idade e destacam-se os procedimentos de caráter diagnóstico (laboratoriais e radiológicos), porque nessa faixa etária são crianças que, por terem a imunidade mais baixa em relação aos adultos apresentam doenças que necessitam deste tipo de exames.
- No segundo cluster estão os usuários de 23 a 60 anos, a faixa etária em que se apresenta o período fértil das mulheres está neste grupo que abrange a faixa dos 15 aos 49 anos de idade. Os exames mais realizados dos 23 aos 38 anos estão diretamente relacionados à gravidez. Na faixa dos 38 aos 60 aparecem em maior número os exames hormonais, pois a mulher está no período da menopausa.

No segundo resultado foi utilizado o algoritmo SimpleKmeans, e encontrados vinte clusters distintos. Após a avaliação verificou-se apenas sete clusters apresentavam alguma informação útil, apresentados no gráfico 3. Segue abaixo uma breve descrição de cada cluster:

- Cluster um (amarelo): apresentam mulheres de 25 a 32 anos, principais procedimentos realizados: exames de laboratórios relacionados à gravidez e às doenças sexuais e procedimento de parto via vaginal;
- Cluster dois (rosa): formado por mulheres de 33 a 38 anos de idade, os procedimentos mais comuns neste cluster são: exames laboratoriais relacionados à gravidez e às doenças sexuais, procedimentos de parto e ultrassons diversos;
- Cluster três (azul): formado por mulheres de 39 a 44 anos, com utilização dos procedimentos: mamografia, exames hormonais, procedimentos de parto, punção mamária, alguns casos de exérese de lesões na pele;

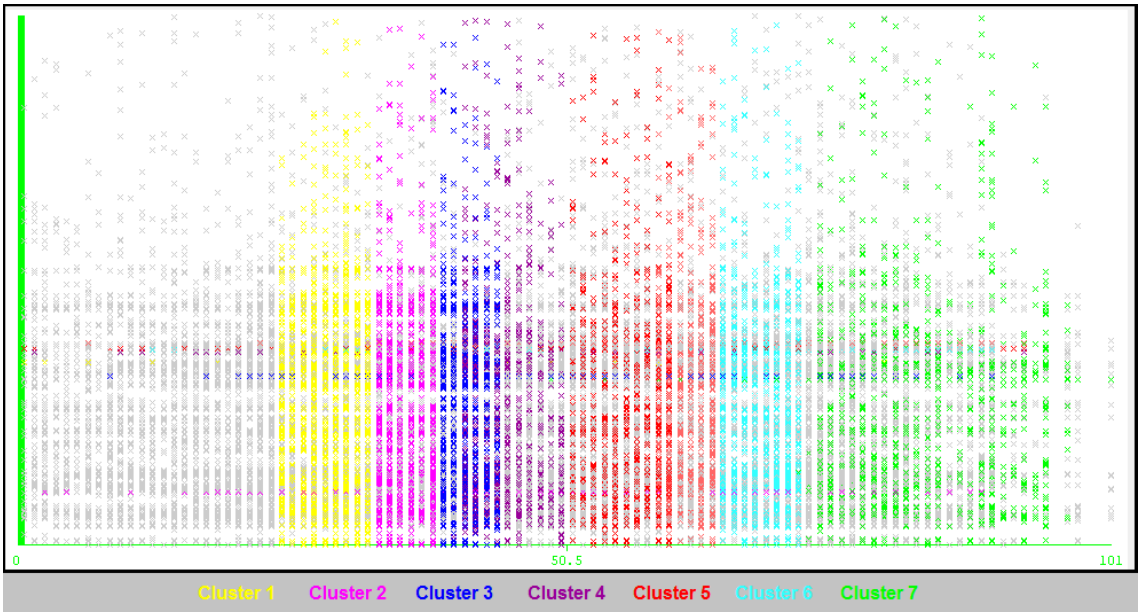

Gráfico 3 – Resultados úteis com o algoritmo SimpleKmeans

- Cluster quatro (roxo): formado por homens na faixa etária dos 41 aos 50 anos, onde os procedimentos mais comuns são: exames de próstata, exames de urina e exames cardíacos;
- Cluster cinco (vermelho): também formado por homens, da idade de 51 a 60 anos, os procedimentos mais comuns neste cluster são: uma frequência maior dos exames de próstata do que apresenta o cluster quatro, cirurgias de próstata, exames laboratoriais, cirurgias e exames nas articulações.;
- Cluster seis (turquesa): cluster formado por mulheres na faixa dos 66 a 72 anos de idade, com os procedimentos mais comuns sendo exames: oftalmológicos, auditivos, hormonais, e aparecem também exames mais complexos (ressonâncias magnéticas, cintilografias e tomografias);
- Cluster sete (verde): formado por homens da idade de 74 a 95 anos, que realizam com frequência os procedimentos: cirurgia de remoção de próstata, exérese na pele, cirurgias ortopédicas, exames oftalmológicos, exames auditivos, e tratamentos para esses dois últimos, e com uma frequência muito grande aparecem fisioterapias de todos os tipos;

5 – CONSIDERAÇÕES FINAIS

Após a análise de cada cluster feitas no capítulo 4 deste trabalho foram enviadas as seguintes sugestões de medidas preventivas para a gestora:

- Realizar um maior número de palestras e cursos focados em gestantes e futuras gestantes, evitando várias consultas e exames que podem ser resolvidos nestes eventos;
- Realizar cursos preparatórios para os pais, principalmente quando for o primeiro filho, evitando realizar consultas muito frequentes;
- Fazer eventos que envolva a terceira idade, pessoas acima de 60 anos, estimulando-as a praticas de exercícios e buscar uma vida mais saudável, pois melhorando a qualidade de vida evita-se gastos com tratamentos como: fisioterapias e até mesmo cirurgias;
- Incentivar o exame anual de próstata nos homens a partir dos 45 anos, através de cartilhas ou palestras mostrando que quanto mais cedo é detectado o câncer maior a chance de tratamento e menos invasivo é o procedimento;
- Incentivar a realização da mamografia nas mulheres a partir 40 ou 50 anos, dependendo se há fatores de riscos, como casos confirmados de câncer de mama na família;
- Com este estudo foi comprovado que é possível através da aplicação do processo de KDD, reconhecer e classificar os usuários pertencentes a uma gestora de planos de saúde de acordo com a utilização dos mesmos. E de acordo com esse grupos tomar as decisões de quais medidas preventivas deve ser tomadas para evitar gastos futuros e zelar pela saúde dos usuários.

REFERÊNCIAS

A HISTÓRIA DO ORACLE: Inovação, Liderança e Resultados. Disponível em: <http://www.oracle.com/br/corporate/press/story-346137-ptb.html>. Acessado em 07 de março de 2012.

ATKINSON, M.; BANCILHON, F.; DITTRICH, K.; MAIER, D.; ZDONIK, S.; The Object-Oriented Database System Manifesto. 1989. Disponível em: < cs.cmu.edu/afs/cs.cmu.edu/user/clamen/OODBMS/Manifesto/htManifesto/Manifesto.html>. Acessado em: 19 de março 2012.

ATTRIBUTE-RELATION FILE FORMAT (ARFF). Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/arff.html>. Acessado em: 27 de abril de 2012.

BERRY, Michael J. A.; LINOFF, Gordon, S.. Data Mining Techniques: For Marketing, Sales and Customer Relationship Management. 2nd ed.. Indianapolis: Wiley Publishing, Inc., 2004. 643p.

BOUCKAERT, Remco R. et al. WEKA: Manual for Version 3-7-5. 2011. Disponível em: <http://ufpr.dl.sourceforge.net/project/weka/documentation/3.7.x/WekaManual-3-7-5.pdf>. Acessado em: 25 de abril de 2012.

CARNEIRO, José Luís. Introdução a banco de dados. 2004. Salvador: s.n., 2004. 65p.

CECHIN, José; MARTINS, Carina Burri; LEITE, Francine. VCMH – Variação dos Custos Médico-Hospitalares, 2009. Disponível em: <http://www.iess.org.br/html/TD00252009VCMH.pdf>. Acessado em: 25 de Janeiro de 2012.

CHIARA, Ramon; Aplicação de Técnicas de Data Mining em Logs de Servidores Web. Universidade de São Paulo – São Carlos, 2003. 176 p.

COSTA, Rogério Luís de Carvalho. SQL: Guia Prático, 2nd ed. Rio de Janeiro: Brasport, 2006. 232 p. Disponível em: <http://books.google.com.br/books?id=3Lxv-q6-S3MC&pg=PA16&lpg=PA16&dq=A+LMD+trata+dos+comandos+ligados+%C3%A0+manipula%C3%A7%C3%A3o+de+dados,+definindo+os+comandos+para+a+sele%C3%A7%C3%A3o,+inclus%C3%A3o,+altera%C3%A7%C3%A3o+e+exclus%C3%A3o+de+dados+de+tabelas.+J%C3%A1+a+LDD+re%C3%BAne+os+comandos+para+a&source=bl&ots=IuvY6l8OxM&sig=ELJMrnQ2mxE_JoIFIiiXDoMvmY0&hl=pt-BR&sa=X&ei=_JrfT_nMLofm9ASs44nHCQ&ved=0CEsQ6AEwAA#v=onepage&q&f=true>. Acessado em: 18 de junho de 2012.

ELMARI, Rames; NAVATHE, Shamkant B.. Fundamentals of Database Systems. 4th ed.. Boston: Pearson Addison Wesley, 2003. 1030 p.

FANDERUFF, Damaris. Dominando o Oracle 9i: Modelagem e Desenvolvimento. São Paulo: Editora Pearson Education do Brasil, 2003. 372p.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD Process for extracting Useful Knowledge from Volumes of Data. Comunications of the ACM, v. 39, p. 27-34, nov. de 1996.

FIGUEIRA, Rafael. Mineração de dados e bancos de dados orientados a objetos. 1998. 96f, Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2004.

GENERAL PUBLIC LICENSE. Disponível em: <http://www.gnu.org/licenses/gpl.html>. Acessado em: 26 de abril de 2012.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. Data Mining: um guia prático. Rio de Janeiro: Elsevier, 2005. 261 p.

GOMES, Romeu et al. A prevenção do câncer de próstata: uma revisão da literatura. Ciência & Saúde Coletiva, v. 3, p. 235-246, 2008. Disponível em: <http://www.scielo.br/pdf/csc/v13n1/26.pdf>. Acessado em 24 de junho de 2012.

GONÇALVES, Lóren Pinto Ferreira, Avaliação de ferramentas de mineração de dados como fonte de dados relevantes para a tomada de decisão: aplicação na rede unidão de supermercados, São Leopoldo - RS. 2001. 92f. Dissertação (Mestrado em Administração) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2001.

HADDAD, Nagib; SILVA, Maria Barbosa. Mortalidade feminina em idade reprodutiva no Estado de São Paulo, Brasil, 1991-1995: causas básicas de óbito e mortalidade materna. Revista de Saúde Pública, v. 34, nº 1, p. 64-70, fev. 2000. Disponível em: <http://www.scielosp.org/pdf/rsp/v34n1/1383.pdf>. Acessado em: 24 de junho de 2012.

HAYES, Frank; The Story So Far. 2002. Disponível em: <http://www.computerworld.com/s/article/70102/The_Story_So_Far?taxonomyId=009> Acessado em: 16 de março de 2012.

LOSS, Leandro; RABELO, Ricardo José. Sistemas de Data Mining. Florianópolis: Universidade Federal de Santa Catarina, 2004. 12 p.

MARCHI, Ailton Augustinho; GURGEL, Maria Salete Costa; FONSECHI-CARVASAN, Gislaine Aparecida. Rastreamento mamográfico do câncer de mama em serviços de saúde públicos e privados. Revista Brasileira de Ginecologia e Obstetrícia, v. 28, p. 214-219, 2006. Disponível em: <http://www.scielo.br/pdf/rbgo/v28n4/a02v28n4.pdf>. Acessado em: 25 de junho de 2012.

MUNIZ, Eliane. Introdução a banco de dados. S. l.:s.n. 200-?.

ON TARGGET TREINAMENTO E CONSULTORIA. Introdução ao Oracle 8i: Volume I. S.l.: s.n., 2000. p. 15-16. Disponível em: <http://pt.scribd.com/doc/29296145/Introducao-ao-Oracle-8i>. Acessado em: 13 de março de 2012.

PL/SQL USER's GUIDE AND REFERENCE. Oracle, 2005. Disponível em: <http://docs.oracle.com/cd/B19306_01/appdev.102/b14261.pdf> Acessado em 07 de março de 2012.
RAMAKRISHNAN, Raghu, GEHRKE, Johannes. Database Management Systems. 3 ed.. New York: McGraw-Hill, 2003. 1065 p.

REZENDE, Ricardo. Conceitos Fundamentais de Banco de Dados – Parte 2.Disponível em: <http://www.sqlmagazine.com.br/Colunistas/RicardoRezende/03_ConceitosBD_P2.asp>. Acessado em: 07 de março 2012.

ROSE, Geoffrey. Estratégias de medicina preventiva. Porto Alegre: Artmed, 2010. 192 p.

SQL REFERENCE. Oracle, 2005. Disponível em: <http://docs.oracle.com/cd/B19306_01/server.102/b14200.pdf>. Acessado em 07 de março de 2012.

TAKAI, Osvaldo Koaro; ITALIANO, Isabel Cristina; FERREIRA, João Eduardo. Introdução a banco de dados. São Paulo: DCC-IME-USP, 2005. 124p. Capítulo 8.

WEKA 3: Data Mining Software in Java. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/index.html>. Acessado em: 26 de abril de 2012.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A.. Data Mining: Pratical Machine Learning Tool and Techniques. 3rd ed. Burlington: Morgan Kaufmann Publishes, 2011. 629 p.