

Avaliação de Modelos Para a Classificação De Beneficiários Com Indicativos para o Diabetes Mellitus Tipo 2.

Marcelo Rosano Dallagassa¹, Sandra Honorato da Silva¹, Deborah Ribeiro Carvalho².

¹ Pontifícia Universidade Católica do Paraná, PUCPR, Brasil

² Instituto Paranaense de Desenvolvimento Econômico e Social. IPARDES, Brasil

Resumo - O grande volume de informações em bases de dados e a complexidade de interpretar e avaliar informações, demanda a utilização de ferramentas e técnicas que permitam uma análise mais automática e inteligente para a geração de novos conhecimentos. Este trabalho direcionou-se para a utilização de Descoberta do Conhecimento e Base de Dados (KDD) e Mineração de Dados (DM) para seleção de um método de mineração de dados eficiente para a classificação de uma carteira de beneficiários referente ao diabetes mellitus tipo 2. Para o processo de mineração foram selecionadas 13 variáveis, aplicadas a 60000 beneficiários, obtendo-se como resultado uma taxa de acerto de 93% com a descoberta de 1425 regras, mostrando-se uma ferramenta eficaz para o processo de classificação para subsidiar ações relacionadas a promoção e prevenção a saúde.

Palavras-chave: Diabetes Mellitus, Mineração de Dados - Descoberta do Conhecimento em Base de Dados

Abstract - The large volume of information in databases and the complexity to interpret and evaluate information, demanded to use tools and techniques that allow an automatic and intelligent analysis for the generation of new knowledge. In this work introduce the concept of KDD - Knowledge Discovery in Databases, data mining (DM) and identify a method of Data Mining efficient of the classification of clients relation to the diabetes illness mellitus type 2, and in the end, presents the results of the mining task of a sample of 60000 beneficiaries, getting up as a result of a hit rate of 93% in 1425 with the discovery rules and were an effective tool for the process of classification to subsidize activities related to health promotion and prevention.

Keywords: Diabetes Mellitus, Data Mining, KDD - Knowledge Discovery in Databases

Introdução

Este artigo tem como objetivo apresentar alguns conceitos e técnicas de mineração de dados para a classificação dos beneficiários com indicativos de propensão ao diabetes mellitus tipo 2, baseada em dados históricos de utilização de procedimentos e exames, para a formação de indicadores de promoção e atenção à saúde.

Com criação da ANS em novembro de 1999, pela medida provisória n. 1928, convertida na lei n. 9.961, como órgão de regulação, normatização, controle e fiscalização das atividades que garantem a assistência suplementar à saúde ocorreram mudanças na estrutura do setor de saúde suplementar no Brasil, conforme figura [1].

<u>Antes Regulamentação</u>	<u>Depois Regulamentação</u>
Modelo Centrado na doença	Modelo de atenção com ênfase nas ações de promoção e prevenção de doenças.
Ausência de sistemas de Informações	Sistemas de informações como insumo estratégico.

Figura 1: Mudanças estruturais do setor de saúde suplementar

Neste processo de mudanças para o setor da saúde suplementar, percebe-se a importância do

modelo de atenção focado nas ações de promoção à saúde e prevenção de doenças e o imperativo da utilização de sistemas de informações gerenciais, como insumo estratégico para apoiar a tomada de decisões.

Desenvolver modelos de intervenção, reunir e tratar informações e monitorar resultados são focos prioritários das operadoras de planos de saúde. Conhecer os ciclos de atendimentos: diagnóstico, gerenciamento e prevenção de doenças, como estratégias para melhorar a qualidade do atendimento e reduzir os custos, bem como a relevância da identificação e domínio para abordagens de gerenciamento de doenças. Os planos de saúde devem assumir a responsabilidade de ajudar os clientes a compreender os fatores que afetam sua saúde e a melhor forma de promover o autocuidado, atuando como “conselheiros e defensores da saúde” [2].

Neste contexto, os órgãos e instituições de saúde vêm direcionando suas ações para a prevenção das Doenças Crônicas Não Transmissíveis (DCNT), como as cardiovasculares, o câncer, o diabetes, a cirrose hepática, as pulmonares obstrutivas crônicas e os transtornos mentais, pois segundo estimativas da OMS [3], a cada ano, pelos menos: 4,9 milhões de pessoas

morrem em decorrência do consumo do tabaco; 2,6 milhões de pessoas morrem como consequência do sobrepeso ou da obesidade; 4,4 milhões de pessoas morrem em decorrência de níveis de colesterol elevados; 7,1 milhões de pessoas morrem em decorrência de pressão sanguínea elevada.

As Doenças Crônicas Não Transmissíveis (DCNT), se caracterizam por apresentar, de forma geral, um longo período de latência, levando os indivíduos a se tornarem progressivamente enfermos, especialmente se não tiverem um tratamento adequado. Este fato reforça a preocupação das instituições com os custos da assistência a saúde, fortalecendo a motivação pela prevenção e o controle das DCNTs [4].

Neste estudo, desenvolveu-se uma avaliação sobre o Diabetes Mellitus Tipo 2, também conhecido como o diabetes do adulto. O Diabetes caracteriza-se como um distúrbio do metabolismo dos açúcares (glicoses e outros), gorduras (lipídios) e proteínas e quando identificada precocemente e controlada suas complicações crônicas podem ser evitadas [4].

Uma das dificuldades das operadoras de plano de saúde é a inexistência de informações clínicas em suas bases de dados, dificultando a identificação da sua carteira de beneficiário, para o desenvolvimento de ações de prevenção e promoção da saúde.

Devido ao grande volume de informações nas bases de dados geradas pelas solicitações de atendimentos de uma operadora de planos de saúde e a complexidade de interpretar e avaliar as informações para gerar novos conhecimentos, faz-se necessária a utilização de ferramentas e técnicas que permitam uma análise mais automática e inteligente como DW – *Data Warehouse*, KDD – *Knowledge Discovery in Databases – Data Mining*, Redes Neurais e Inteligência Artificial e Sistemas Especialistas.

A diversidade de base de dados em ambiente transacional demanda um processo que concentre os dados, incluindo históricos e dados externos, que atendam as necessidades de consultas estruturadas, relatórios analíticos e de suporte à decisão, conforme a definição de DW - *Data Warehouse* [5].

KDD é o processo não trivial de identificar padrões nos dados, que sejam válidos e previamente desconhecidos, potencialmente úteis e compreensíveis, visando melhorar o entendimento do problema, subsidiando o processo de tomada de decisões [6].

KDD caracteriza o processo, as fases da descoberta do conhecimento em banco de dados, determinado em múltiplos passos executados em

seqüência. O primeiro KDD foi apresentado em 1996 [6], com descrição das etapas deste processo.

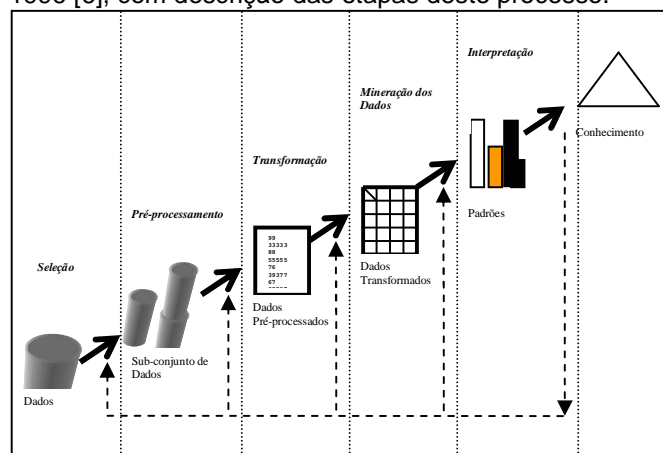


Figura 2 - Fases do KDD. Fayyad et al, 1996

As fases compreendem:

- **Análise inicial:** Levantamento de informações e aprendizagem sobre o conhecimento que se pretende descobrir, o que compreende um domínio da aplicação e o tipo de decisão no qual o conhecimento pode contribuir para o processo.
- **Seleção:** Identificação de um subconjunto de dados (atributos) para a realização da tarefa da descoberta de conhecimento.
- **Limpeza dos dados e pré-processamento.** Envolve operações como a unificação e transformação de formato dos dados, limpeza dos dados, correção de erros no conjunto de dados, detecção e remoção de dados distorcidos e inserção e correção de valores ausentes.
- **Redução de dados e projeção:** identificação de características para a identificação dos dados em função do objetivo da tarefa, visando a redução do número de variáveis.
- **Mineração de dados:** compreende selecionar os métodos a serem utilizados para a tarefa de mineração de dados, a escolha de um método e um algoritmo mais apropriado e o próprio processo da geração dos padrões nos dados.

No método de classificação, a entrada de dados para um processo é uma coleção de registros. Cada registro, também conhecido como instância ou exemplo, é caracterizado por um tupla (x, y) , onde x é um atributo do conjunto e y é um atributo especial, definido como a classe identificadora da tupla [7].

A classificação é um método que consiste, na identificação de um modelo (padrões de comportamento), por meio de uma função de aprendizado f que mapeie e represente a melhor precisão, baseada nos valores dos atributos identificadores da classe y .

Os classificadores são compostos de dois conjuntos de dados: O primeiro é um conjunto de

atributos, sendo um deles definido como o atributo classe (identificador do registro) e o segundo, um conjunto de teste utilizado para determinar a precisão do modelo. Desta maneira, aos registros previamente desconhecidos é assinalada uma classe tão precisa quanto possível [7].

Os algoritmos de árvore de decisão possuem uma estrutura, onde a divisão de cada nó interno é definida utilizando um conjunto de dados de treinamento, que possui uma classe de caracterização do registro, que por meio de procedimentos estatísticos divide os nós, utilizando a estratégia “*dividir-para-conquistar*”. Esta técnica é chamada de algoritmo de aprendizado (“Learning Algorithms”) [8].

A estrutura de uma árvore de decisão indutiva é composta de um nó raiz o topo da árvore (“root node”) e representado como sendo a divisão mais relevante, um conjunto de nós internos (“internal node”) que representam uma avaliação sobre um determinado atributo e de nós terminais chamados de folhas (“leaves”) que representam os identificadores das classes [8].

Existem vários algoritmos de árvores de decisão, entre eles o pioneiro CART (*Classification and Regression Trees*), sugerido por Breiman et al (1984). Sua principal característica é ser binária, contendo apenas, duas divisões para cada nó de decisão e suas divisões são baseadas na medida de impureza. O ID3 e o C4.5, proposto por Quinlan (1993), utilizam o conceito de “ganho da informação” como critério de ramificação, para a construção da árvore de decisão. O ganho da informação é baseado no índice de entropia para medição da homogeneidade de cada nó [9].

- **Interpretação e avaliação dos padrões:** nesta etapa avalia-se e identificam-se os padrões extraídos e seus modelos. A vantagem da utilização dos algoritmos de árvores de decisão é que a derivação de regras representadas como conjuntos do tipo “Se – Então”, podem ser aplicadas diretamente em uma determinada função, tornando-o um modelo claro e demonstrando os atributos que estão discriminando os padrões [10].

Para a análise e a avaliação de algoritmos de classificação de acordo com os modelos de indução, utilizam-se algumas métricas como a acurácia (1), que é a quantidade de registros classificados corretamente e a taxa de erro (2) que constituem os registros classificados incorretamente, e indicam o quão o modelo é confiável.

Para a representação destas métricas, utiliza-se uma matriz, chamada de matriz de confusão.

Classe Real	Classe Prevista	
	Classe =1	Classe = 0
Classe = 1	f_{11}	f_{10}
Classe = 0	f_{01}	f_{00}
Classe Real		

Quadro 1 - Matriz de Confusão para 2 Classes

Onde: f_{11} – Verdadeiro Positivo, f_{10} – Verdadeiro Negativo, f_{01} – Falso Negativo e f_{00} – Falso Positivo

$$\text{Acuracia ou Precisão} = \frac{f_{11} + f_{00}}{f_{10} + f_{01} + f_{00}} \quad (1)$$

$$\text{Taxa de Erro} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}} \quad (2)$$

Metodologia

O estudo foi desenvolvido utilizando-se as técnicas de KDD, sobre a base de dados do Data Warehouse da Unimed Federação do Estado do Paraná, , estrutura que possibilita gerar as informações de atendimentos de beneficiários ativos, do Estado do Paraná, conforme autorização de acesso aos dados pela instituição e aprovação do projeto no Comitê de Ética em Pesquisa da PUC-PR, nº 2383.

O processo contemplou como etapas:

- ➔ *Análise Inicial*
- ➔ *Pré-processamento*
- ➔ *Mineração de Dados*
- ➔ *Avaliação e Interpretação dos Resultados*

Análise Inicial

Para esta etapa aplicou-se um desenho individuado – observacional – longitudinal - retrospectivo, chamado também, de estudo de caso-controle, concebido especialmente para investigar associações etiológicas em doenças de baixa incidência e/ou condições com período de latência prolongado. Neste tipo de estudo, após a identificação dos elementos diagnosticadores, é realizada uma retrospectiva na sua historia, para investigação dos possíveis fatores, que possam ser considerados como importantes para a identificação e classificação de novos elementos [11].

Para esta investigação histórica, foram selecionados beneficiários que tiveram internamentos pelo Código Internacional de Doenças, 10ª revisão (CID-10) de diabetes mellitus, no terceiro trimestre de 2007, com idade igual ou superior a 25 anos, sendo avaliados os seus atendimentos de forma retrospectiva nos 5 anos anteriores, ou seja o período compreendido entre 01 de janeiro de 2003 à 31 de dezembro de 2007.

Pré-Processamento

Utilizou-se para esta etapa, a mesma base do ambiente Data Warehouse da Federação das Unimed's do Estado do Paraná, no período histórico de 6 anos, de 01 de janeiro de 2002 a 31 de dezembro de 2007, sendo selecionados aleatoriamente 60000 instâncias para treinamento e validação.

Os dados foram agrupados e sumarizados por beneficiário, sem a sua identificação. Para a fase de treinamento e validação, determinou-se a classe identificadora, com a seguinte regra:

Se existência atendimento com CID de diabetes e solicitação de hemoglobina glicosilada ≤ 1 então "sem indicativo para";

Se existência atendimento com CID de diabetes e solicitação de hemoglobina glicosilada ≤ 4 então "com indicativo para";

Se existência atendimento com CID de diabetes e solicitação de hemoglobina glicosilada > 4 então "diabético".

Mineração de Dados

Definiu-se pelo algoritmo com aprendizagem supervisionada e a representação do conhecimento, utilizando a técnica de árvore de decisão, considerando que estes métodos são preditivos, pois desempenham inferência nos dados com o intuito de fornecer previsões ou tendências.

Avaliação e interpretação dos resultados

Apresentam-se e avaliam-se os resultados, por meio das regras e padrões descobertos e a matriz de confusão, derivando a acurácia e taxa de erro.

Resultados

Análise Inicial

No conjunto selecionado para o estudo retrospectivo, foram identificadas 59 instâncias (beneficiários) e 20.525 eventos realizados no período de 5 anos.

Neste conjunto de dados foi realizada uma avaliação da idade dos beneficiários, apresentada na figura 3.

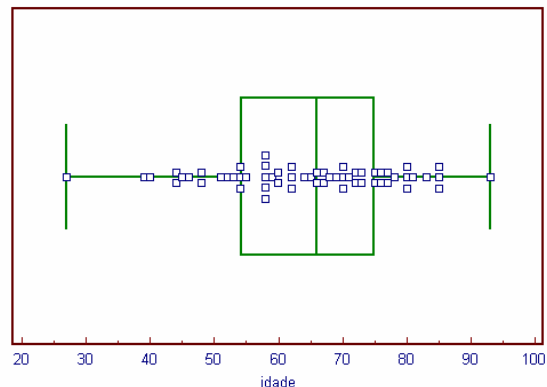


Figura 3 – Gráfico Box Plot - Valor Mínimo, 1º.Quartile, Média, 3º.Quartile e Valor Máximo

Para a avaliação dos eventos que caracterizaram o atendimento, de cada um dos 59 beneficiários neste período foram realizadas as pesquisas; consultas por especialidades e exames solicitados para o levantamento dos elementos importantes para o estudo.

Inicialmente desenvolveu-se a pesquisa relacionando os procedimentos de consultas, para o levantamento do quantitativo por especialidade médica, obtendo-se as informações apresentadas no gráfico 4.

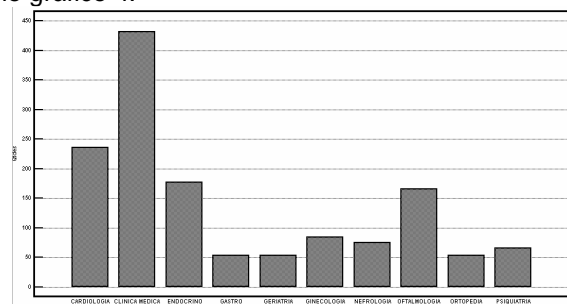


Gráfico 4 - Consultas por especialidade. Federação das Unimed's do Estado do Paraná – Período 01/01/2003 a 31/12/2007

Desta forma ficaram evidenciadas as especialidades cardiologia, endocrinologia e oftalmologia pela alta frequência de ocorrências, 236, 177 e 166 respectivamente. Foi desprezada a clinica médica com 431 casos, por não favorecer uma associação direta ao diabetes. Desta maneira selecionaram-se as especialidades cardiologia, endocrinologia e oftalmologia, como variáveis importantes para o estudo. Foi realizada ainda, a inclusão da especialidade Nefrologia por delinear-se como uma especialidade associada diretamente às complicações do diabetes tipo 2.

Da mesma maneira, realizou-se a pesquisa sobre a quantidade de exames solicitados aos beneficiários, classificando-os em ordem decrescente de quantidade e selecionando os mais solicitados, sendo os resultados apresentados na Tabela 1.

Tabela 1 – Exames solicitados. Federação das Unimeds do Estado do Paraná – período jan./2003 a dez.2007

Exame	Quantidade solicitada
GLICOSE	1512
CREATININA	887
HEMOGRAMA COMPLETO	840
POTASSIO	613
SODIO	496
UREIA	459
COLESTEROL TOTAL	447
HEMOGLOBINA GLICOSILADA	444
ROTINA DE URINA	398
TRIGLICERIDIOS	379

Considerando os resultados apresentados, foram selecionados os exames; glicose, creatinina e hemoglobina glicosilada. Foram adicionalmente acrescidos os exames: microalbuminúria, curva glicêmica e mapeamento de retina, como exames adicionais e relevantes para o estudo.

O quadro 2 apresenta os atributos relevantes, identificados na análise inicial e selecionados para o estudo.

Exames Laboratoriais	Exames Especiais	Consultas Especialidades	Outras
Glicose	Curva	Oftalmologia	Estado
Creatinina	Glicêmica	Endocrinologia	Civil
Microalbuminúria	Mapeamento de Retina	Nefrologia	Idade
		Cardiologia	CID
			Obesidade

Quadro 2 – Atributos selecionados para o estudo

Mineração de Dados

Dentre os algoritmos disponíveis no software WEKA [12].– *Waikato Environment for Knowledge Analysis – Version 3.4.11*, software de livre utilização, produzido pela Universidade de Waikato – Nova Zelândia, optou-se pelos métodos de classificação C4.5 (J48). A metodologia para o teste consistiu em aplicar o método de referência cruzada, repetido 10 vezes (10 simulações).

Interpretação e avaliação dos resultados.

Nesta etapa estão apresentados e avaliados os resultados por meio do algoritmo de árvore de decisão C4.5. e da matriz de confusão, calculando a acurácia e taxa de erro.

Após a aplicação do método de classificação C4.5 (J48) em referência cruzada, obteve-se um resultado com taxa de acurácia de aproximadamente 92%, ou seja, dos 60.000

registros utilizados para a etapa de treinamento e validação, 55.341 foram classificados corretamente.

Tabela 2 – Matriz de Confusão

REAL	PREVISTO		
	Sem indicativo	Com indicativo	Com diabete
sem indicativo	53067	470	183
com indicativo	2899	627	356
com diabete	474	277	1647

Observa-se ainda, a principal característica das técnicas de algoritmos de árvore de decisão, relacionada a compreensibilidade do resultado final, que pode ser visualizado por meio da estrutura (SENTAÇÃO), evidenciando em cada resultado da classificação (folhas), um indicativo dos registros classificados corretamente em relação aos classificados incorretamente. A figura 5 mostra, apenas 10 das 1425 estruturas encontradas pelo algoritmo e 3 resultados de classificação, dentro dos 733 obtidos.

```

Se Qt_Exa_Cur_Gli > 6
  Se Qt_Exa_Cur_Gli <= 25
    Se Qt_Exa_Micro <= 0
      Se Qt_Exa_Glicose <= 5
        Se Qt_Exa_Cur_Gli <= 14
          Se Qt_Exa_Glicose <= 2
            Se Qt_Exa_Creat <= 1
              Sem Indicativo (116/7)
            Se Qt_Exa_Creat > 1
              Se Qt_Oftamo <= 3
                Sem Indicativo (24/4)
              Se Qt_Oftamo > 3
                Com Indicativo (4/0)
          
```

Figura 5 – Estrutura Árvore de Decisão

Pode-se ainda, por meio da ferramenta WEKA [12], visualizá-las graficamente, conforme a figura 6.

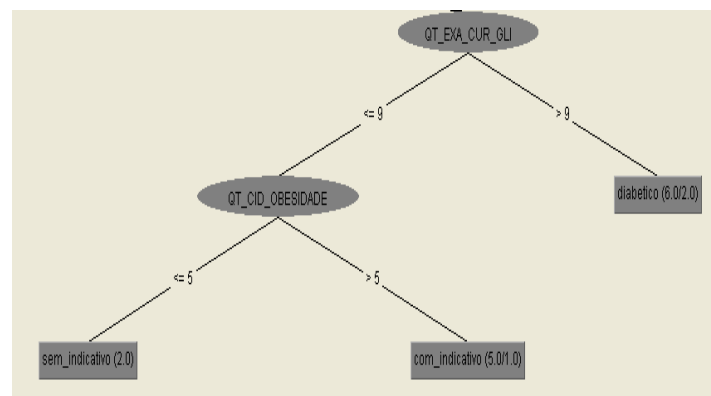


Figura 6 – Visualização Gráfica da Árvore de Decisão

Discussão e Conclusões

Portanto, comprova-se a eficiência na aplicação dos algoritmos de árvores de decisão C4.5, permitindo a identificação e seleção de beneficiários com indicativos ao diabetes, por meio de seu ciclo de atendimento na operadora, para a formação de indicadores para a prevenção e monitoramento da saúde da carteira de beneficiários.

Vislumbra-se a partir dos resultados encontrados, a possibilidade de estabelecimento de indicadores que monitorem todo o processo de prestação de serviços, considerando que combinar a visão de indicadores focado na estratégia das operadoras, com ênfase nas ações de promoção à saúde e prevenção, reforça o conceito em transformar um indicador intangível em um resultado direto para o beneficiário e,

conseqüentemente, reduzir os custos das operadoras de planos de saúde [13].

Desta forma, os indicadores deverão servir de elementos para a execução de ações de prevenção primária, que conforme a definição de Leavell e Clark [14], “é a prevenção realizada no período de pré-patogênese e que tal prevenção exige uma ação antecipada, baseada no conhecimento, a fim de tornar improvável o progresso posterior da doença”. Prática medicina preventiva aqueles que utilizam o conhecimento moderno, na medida de sua capacidade, para desenvolver a saúde, evitar a doença e a invalidez e prolongar a vida.

Pretende-se em trabalhos futuros, estender esta técnica, para outras doenças crônicas não transmissíveis, numa perspectiva de adoção de um novo paradigma assistencial e melhoria de qualidade de vida dos beneficiários.

Referências

1. BRASIL. Conselho Nacional de Secretários de Saúde. Ciência e Tecnologia em Saúde / Conselho Nacional de Secretários de Saúde. Brasília: CONASS, 2007.
2. Porter EM, Teisberg EO. Repensando a Saúde: estratégia para melhorar a qualidade e reduzir os custos, Porto Alegre: Bookman, 2007.
3. Organização Mundial Da Saúde (OMS). Prevenção de Doenças Crônicas um investimento vital, 2005.
4. Agência Nacional de Saúde Suplementar (Brasil). Manual Técnico de Promoção da saúde e prevenção de riscos e doenças na saúde suplementar. Rio de Janeiro: ANS, 2006.
5. Harjinder GS, Prakash RC. The Official Guide to Data Warehousing, Indianapolis: QUE Corporation, 1996,382p
6. Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery: An Overview, Cambridge: AAAI Press,1996.
7. Tan PN, Steinbach M, Kumar V. Introduction To Data Mining, Boston: Pearson Addison-Wesley Longman Publishing Co., 2005.
8. Han J, Kamber M. Data Mining: concepts and Techniques, Califórnia: Morgan Kaufmann Publishers, 2001
9. Larose, DT. Discovering Knowledge In Data: An Introduction To Data Mining, New Jersey: John Wiley & Sons, Inc., 2005.
10. Steiner M, Soma YN, Shimizu T, Nievola JC, NETO PJS. Abordagem de um Problema Médico por Meio do Processo de KDD com Ênfase à Análise Exploratória dos Dados. In: Revista G&P Gestão & Produção, v.13,n.2,p.325-337,2006.
11. Almeida NF, Rouquayrol MZ. Introdução a Epidemiologia, Rio de Janeiro: Guanabara Koogan, 2006.
12. Waikato Weka 3 – Machine Learning Software in Java – disponível em <http://www.cs.waikato.ac.nz/ml/weka>
13. Kaplan RS, Norton DP. Organização orientada para a estratégia: como as empresas que adotam o balanced scorecard prosperam no novo ambiente de negócios. Rio de Janeiro: Elsevier, 2000.
14. Leavell H, Clark EG. Medicina Preventiva. São Paulo: McGraw-Hill do Brasil,1976

Contato

Marcelo Rosano Dallagassa, Rua Luiz Barreto Murat, 842 – Sb.7. Curitiba – PR – email:mrdallagassa@gmail.com.br
fone:(41)3219-1428 / (41)9602-8509