

Descoberta de conhecimento a partir de uma base de indicadores de desenvolvimento social utilizando WEKA

Carlos M. Murasse^{1,2}, Denise F. Tsunoda^{1,3}

¹Programa de Pós-graduação em Ciência, Gestão e Tecnologia da Informação.
Universidade Federal do Paraná - UFPR. Curitiba, PR - Brasil

²Serviço Federal de Processamento de Dados - SERPRO. Curitiba, PR - Brasil

³Departamento de Ciência e Gestão da Informação
Universidade Federal do Paraná - UFPR
Av Prefeito Lothário Meissner, 632 - Jardim Botânico
80210-170 - Curitiba, PR - Brasil

{carlos.murasse, dtsunoda}@ufpr.br

Abstract. *Starting from a database with social development indicators, a knowledge database discovery (KDD) process was applied using data mining techniques and the open-source tool Weka. Three hypotheses of correlation between economic, demographic and health indicators with mortality indicators were tested. For one case, a reliable correlation with causal relationship was identified, which represents a concrete contribution in decision making and direction of future actions of ODM social development program.*

Resumo. *A partir de um banco de dados com indicadores de desenvolvimento social, foi aplicado um processo de descoberta de conhecimento (KDD) com uso de técnicas de mineração de dados e a ferramenta open source de data mining Weka. Três hipóteses de correlação entre indicadores de economia, demografia e saúde com indicadores de taxa de mortalidade foram testadas. Para um dos casos, foi identificada uma correlação confiável com relação causa-efeito, o que representa uma contribuição concreta para a tomada de decisão e direcionamento de ações futuras do programa de desenvolvimento social ODM.*

1. Introdução

Nos diferentes segmentos da sociedade, cresce a busca por tecnologias que agreguem valor aos negócios, seja agilizando operações ou viabilizando inovações [SILVA, 2004]. Com o mercado cada vez mais competitivo, já não basta organizar a produção, reduzir os custos e atender bem. É preciso adquirir conhecimento sobre os clientes e interpretar as suas expectativas [Braga, 2005].

Os sistemas de informação, impulsionados também pela versatilidade e alcance proporcionados pela Internet, disponibilizam em uma escala cada vez maior dados oriundos de atividades operacionais. A disponibilidade de imensas bases de dados proporciona a oportunidade de transformar tais dados em informação e conhecimento úteis para o suporte às decisões organizacionais. Uma sequência ideal nesse processo de transformação seria: Dados → Informação → Conhecimento → Decisão.

As ferramentas e técnicas empregadas na análise destes repositórios de dados são os objetos de estudo da “descoberta de conhecimento em bancos de dados”, uma área da inteligência artificial também conhecida como *Knowledge Discovery in Databases* (KDD).

A mineração de dados, ou *data mining*, é a etapa do KDD responsável pela seleção dos métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação, juntamente com a busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão. O *data mining* seria então o processo de identificar, nos dados disponíveis, padrões válidos, novos, úteis e compreensíveis [FAYYAD et al., 1996].

Este artigo trata da aplicação de KDD sobre um banco de dados com indicadores de desenvolvimento social, utilizando uma ferramenta *opensource* de *data mining*, com o intuito de demonstrar o potencial desta tecnologia para agregar valor também no campo social e das decisões governamentais, com a vantagem adicional de não exigir grandes investimentos.

2. Tarefas, técnicas e algoritmos de mineração

Existem diversas tarefas de KDD, ou problemas de descoberta de conhecimento, que extraem tipos diferentes de conhecimento do banco de dados [MARTINHAGO, 2005]. Neste estudo serão abordadas as tarefas de classificação e de associação.

A classificação é uma das tarefas mais estudadas pela comunidade científica de KDD. O princípio da tarefa de classificação é descobrir algum relacionamento entre os atributos preditivos e o atributo meta, de modo a descobrir um padrão que possa ser utilizado para previsões.

A literatura mostra que existem diversas técnicas para a tarefa de classificação. Uma das técnicas de classificação mais utilizadas é a árvore de decisão. Existem diversos algoritmos que retornam como resultado árvores de decisão, dentre os quais cabe destacar o ID3 e o C4.5, ambos desenvolvidos por Ross Quinlan, considerado por alguns como “pai das árvores de decisão” [QUINLAN, 1993].

O algoritmo ID3 (*inductive decision tree*) foi desenvolvido em 1983 e possibilita montar uma árvore de decisão com base na escolha sucessiva do atributo que melhor divide os exemplos, aplicando os conceitos de entropia e ganho da informação propostos por Shannon (1948). No anexo 1, apresentamos um exemplo de aplicação do algoritmo ID3.

O algoritmo C4.5 estende o algoritmo básico ID3 para tratar alguns problemas do mundo real, tais como valores ausentes e valores numéricos contínuos.

A tarefa de associação ou de afinidade de grupos, inicialmente proposto por AGRAWAL et al. (1993), visa combinar itens de forma que a presença de um item (ou mais itens) pressupõe a presença de outro item (ou outros itens). Uma área que utiliza com frequência a tarefa de associação é a de vendas, geralmente visando descobrir quais produtos costumam ser vendidos juntos numa mesma transação.

Um dos algoritmos mais referenciados para realizar a tarefa de associação é o Apriori, o qual avalia e retorna associações relevantes entre os itens, segundo critérios de suporte e confiança [ROMÃO, 2002]. O objetivo, então, é encontrar todas as regras de associação relevantes entre os itens do tipo: $X \text{ (antecedente)} \Rightarrow Y \text{ (conseqüente)}$.

AGRAWAL et al. (1993) propuseram um modelo matemático, onde as regras de associação geradas devem atender a um suporte e confiança mínimos especificados. O suporte corresponde à frequência com que ocorrem os padrões em toda a base, enquanto a confiança é uma medida da força das regras.

O suporte mínimo s é a fração das transações que satisfaz a união dos itens do grupo conseqüente com os itens do grupo antecedente, de forma que estejam presentes em pelo menos $s\%$ das transações. A confiança mínima c garante que ao menos $c\%$ das transações que satisfaçam o antecedente das regras também satisfaçam o conseqüente das regras.

3. Base de dados e ferramentas

Para este estudo sobre descoberta de conhecimento em banco de dados, foi selecionado um banco de dados com indicadores de desenvolvimento social e uma ferramenta *open source* de *data mining* denominada Weka.

O banco de dados selecionado suporta um programa intitulado Objetivos de Desenvolvimento do Milênio (ODM), que é formado por objetivos, metas e indicadores pactuados pelo Brasil e por outros 190 países membros das Nações Unidas para melhorar indicadores sociais, ambientais e econômicos [PNUD]. Um quadro-resumo dos objetivos, metas e indicadores ODM encontra-se no anexo 2. Para este trabalho, foi utilizada uma versão do banco de dados ODM referente ao Estado do Paraná, que foi obtida junto à Federação das Indústrias do Estado do Paraná (Fiep), umas das entidades que apóia o programa ODM.

A ferramenta de *data mining* selecionada foi desenvolvida na Universidade de Waikato na Nova Zelândia. O Weka pode ser definido como uma coleção de algoritmos do tipo *machine learning* para tarefas de *data mining*. Vários desses algoritmos incorporam conceitos de inteligência artificial. O desenvolvimento do Weka foi iniciado por iniciativa da Fundação de Pesquisa, Ciência e Tecnologia do governo da Nova Zelândia. O Weka vem sendo cada vez mais utilizado e algumas características interessantes ajudam a explicar seu sucesso [MARKOV e RUSSELL, 2006]:

- contém diversos algoritmos para *data mining*, *web mining* e *machine learning*;
- tem código aberto e está disponível na Web gratuitamente;
- é relativamente fácil de usar, inclusive por pessoas que não sejam especialistas;
- proporciona recursos flexíveis para experimentos;
- é mantido atualizado, pois novos algoritmos são adicionados logo que surgem na literatura.

Para este trabalho, foram utilizados alguns dos algoritmos de classificação e associação, bem como um algoritmo de seleção de atributos, disponíveis no Weka versão 3.6. Na preparação dos dados, planilhas eletrônicas foram utilizadas como ferramentas complementares para realizar manipulações simples tais como classificações e filtros.

4. O processo de mineração

Para Sérgio Navega (2002), os passos fundamentais de uma mineração bem sucedida seriam os seguintes: a partir de fontes de dados (bancos de dados, relatórios, logs de acesso, transações, etc) efetua-se uma limpeza (consistência, preenchimento de informações, remoção de ruído e redundâncias, etc), que dá origem aos repositórios

organizados (*Data Marts* e *Data Warehouses*). Seria a partir deles que faz-se a seleção de dados para atravessarem o processo de mineração. Geralmente de forma interativa e usando visualização gráfica, refina-se o processo até que padrões valiosos apareçam.

Por sua vez, Fayyad propõe um processo mais elaborado de descoberta de conhecimento em bancos de dados, também conhecido como KDD (*Knowledge Discovery in Databases*). Esse processo é interativo, iterativo, cognitivo e exploratório, envolvendo várias etapas, com muitas decisões sendo tomadas pelo analista (que é um especialista do domínio dos dados, ou um especialista de análise dos dados), conforme descrito a seguir [FAYYAD et al., 1996; SILVA, 2004, p. 3]:

1. Definição do tipo de conhecimento a descobrir, o que pressupõe uma compreensão do domínio da aplicação bem como do tipo de decisão que tal conhecimento pode contribuir para melhorar.
2. Criação de um conjunto de dados alvo (*Selection*): selecionar um conjunto de dados, ou focar num subconjunto, onde a descoberta deve ser realizada.
3. Limpeza de dados e pré-processamento (*Preprocessing*): operações básicas tais como remoção de ruídos quando necessário, coleta da informação necessária para modelar ou estimar ruído, escolha de estratégias para manipular campos de dados ausentes, formatação de dados de forma a adequá-los à ferramenta de mineração.
4. Redução de dados e projeção (*Transformation*): localização de características úteis para representar os dados dependendo do objetivo da tarefa, visando a redução do número de variáveis e/ou instâncias a serem consideradas para o conjunto de dados, bem como o enriquecimento semântico das informações.
5. Mineração de dados (*Data Mining*): selecionar os métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação ou conjunto de representações; busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão.
6. Interpretação dos padrões minerados (*Interpretation/Evaluation*), com um possível retorno aos passos 1-6 para posterior iteração.
7. Implantação do conhecimento descoberto (*Knowledge*): incorporar este conhecimento à performance do sistema, ou documentá-lo e reportá-lo às partes interessadas.

Considerando os objetivos deste artigo, optou-se por adotar o processo proposto por Fayyad, exceto a última etapa, de implantação do conhecimento descoberto. A restrição deve-se ao fato de, até este ponto da pesquisa, não ter sido possível apresentar os resultados obtidos, juntamente com uma proposta de implantação de ações, aos responsáveis pela gestão do programa ODM.

5. Aplicação do Processo

Na etapa de definição do tipo de conhecimento a descobrir, um estudo prévio do programa denominado Objetivos do Milênio (ODM), cujos objetivos e indicadores estão descritos no Anexo 2, e da base de dados correspondente aos indicadores do Estado do Paraná, indicou que haveria interesse potencial para pesquisar relações de causa-efeito entre os indicadores de economia, demografia e saúde e os indicadores de mortalidade geral, infantil e materna. Foram então definidas 3 hipóteses:

Hipótese 1: Os indicadores selecionados estão correlacionados com a mortalidade geral

Hipótese 2: Os indicadores selecionados estão correlacionados com a mortalidade infantil

Hipótese 3: Os indicadores selecionados estão correlacionados com a mortalidade materna

Na etapa de criação de um conjunto de dados alvo, foram selecionados os seguintes indicadores das áreas de economia, demografia e saúde:

- População apurada por censo do IBGE (2000 e 2007)
- População estimada (demais anos)
- Renda média do trabalhador formal - Salários mínimos
- Proporção de mulheres assalariadas no emprego formal
- Empregados no mercado formal
- AIDS novos casos
- Proporção de crianças menores de 1 ano com vacinação em dia
- Proporção de crianças menores de 2 anos com peso abaixo da média
- Proporção de gestantes quanto ao número de consultas pré-natal - de 1 a 3 consultas
- Proporção de gestantes quanto ao número de consultas pré-natal - de 4 a 6 consultas
- Proporção de gestantes quanto ao número de consultas pré-natal - 7 ou mais consultas
- Natalidade

Ainda na etapa de seleção de dados, na dimensão do tempo, foi selecionado inicialmente o período de 1997 a 2008, e na dimensão de área geográfica, foram inicialmente selecionados todos os 399 municípios do Paraná.

Na etapa de limpeza de dados e pré-processamento, alguns dados ausentes foram substituídos por valor médio. Os dados numéricos foram discretizados de acordo com a sua distribuição de frequência. Além disso, novos atributos foram construídos a partir de informações pré-existent:

Proporção de empregados no mercado formal, derivado a partir de:

- População
- Empregados no mercado formal

População, obtido pela junção de:

- População apurada por censo do IBGE (2000 e 2007)
- População estimada (demais anos)

Na etapa de redução de dados e projeção, visando a redução do número de variáveis e/ou instâncias a serem consideradas para o conjunto de dados, foi realizada uma carga preliminar e aplicado o algoritmo “attributeSelection” do Weka para cada um dos atributos-meta.

Ainda na etapa de redução de dados, na dimensão tempo, devido a falta de informações contínuas de todos os indicadores o período foi reduzido para 2000 a 2006, e na dimensão área geográfica, devido a falta de informações contínuas no nível de município foi utilizado um nível mais agregado de informação, constituído por 10 regiões geográficas.

Na etapa de mineração de dados, foram aplicados algoritmos de classificação Id3 e J48 e algoritmo de associação Apriori, com o objetivo localizar padrões nos dados. O algoritmo J48 é uma implementação Java no algoritmo C4.5 na ferramenta Weka. É importante ressaltar que as etapas de mineração de dados e etapas seguintes ficaram restritas à hipótese 2, uma vez que a etapa de redução de dados invalidou as hipóteses 1 e 3.

Na etapa de interpretação dos padrões minerados, foram analisados os resultados de processamento na ferramenta. Esta atividade é dependente da percepção pessoal do analista e a busca por padrões que agregassem valor com potencial de contribuição para tomadas de decisão foi baseada no conhecimento adquirido sobre a base de dados durante as etapas de preparação.

6. Resultados

Os resultados iniciais do estudo, ainda na etapa de redução de dados, com apoio do algoritmo de seleção de atributos (attributeSelection) do Weka, sugeriram que os seguintes indicadores teriam forte correlação com a taxa de mortalidade infantil:

- Proporção de empregados no mercado formal
- Proporção de gestantes quanto ao número de consultas pré-natal
- Natalidade

O algoritmo de seleção de atributos demonstrou também que os mesmos indicadores não apresentaram correlação significativa com a taxa de mortalidade geral e com a taxa de mortalidade materna, com uma taxa de acerto abaixo de 50% em ambos os casos.

Após a conclusão da preparação dos dados, estes foram carregados no Weka Explorer e então os algoritmos selecionados foram aplicados com o objetivo de identificar padrões.

A aplicação de métodos de classificação Id3 e J48 gerou uma árvore de decisão para a taxa de mortalidade infantil (anexo 3) que sugere que o indicador de maior relevância para determinar a taxa de mortalidade infantil é a proporção de gestantes quanto ao número de consultas pré-natal.

Quanto à confiabilidade da classificação, o algoritmo Id3 do Weka indicou que 87,5% das instâncias foram corretamente classificadas, e algoritmo J48 do Weka indicou que 88,75% das instâncias foram corretamente classificadas, confirmando que o conjunto de atributos selecionados e árvore de decisão proposta explica de maneira confiável o atributo-meta taxa de mortalidade infantil.

A aplicação do método de associação Apriori resultou em 47 regras com confiança de 100% , das quais 24 diretamente relacionadas com taxa de mortalidade infantil (anexo 4). Uma análise deste subconjunto de regras sugere que:

- Existe uma relação entre número de consultas pré-natal e taxa de mortalidade infantil, e que a relação é inversa, ou seja, quanto maior o número de consultas menor a taxa de mortalidade infantil;
- Existe uma relação entre taxa de natalidade e taxa de mortalidade infantil, e que a relação é direta, ou seja, quanto menor a taxa de natalidade menor a taxa de mortalidade infantil, entretanto esta relação é devida à proporcionalidade demográfica e não caracteriza padrão que

agregue valor.

7. Conclusões

A aplicação de um processo de descoberta de conhecimento (KDD) sobre a base de dados ODM permitiu verificar se indicadores de economia, demografia e saúde teriam correlação com taxas de mortalidade geral (hipótese 1), infantil (hipótese 2) e materna (hipótese 3).

O estudo demonstrou que os indicadores selecionados, com ênfase para proporção de empregados no mercado formal, proporção de gestantes quanto ao número de consultas pré-natal e natalidade, têm correlação com a taxa de mortalidade infantil, configurando uma relação de causa-efeito.

A identificação desta relação entre os indicadores pode direcionar de forma mais efetiva os esforços para atingimento dos objetivos do programa ODM. Por exemplo, pode-se inferir que investimentos para ampliar o serviço de consulta pré-natal para as mulheres gestantes teria um efeito positivo e significativo na redução da taxa de mortalidade infantil.

Como oportunidade para estudos futuros, sugere-se: aumentar a escala do experimento, utilizando base de dados ODM nacional ou mundial; estabelecer um canal de interação com as entidades gestoras do programa ODM para aprofundar compreensão do contexto e possibilitar a implementação dos resultados; buscar a complementação dos indicadores omissos no banco de dados ODM, de forma a possibilitar a mineração de dados no nível de município e englobando um período de tempo mais amplo; combinar outras bases de indicadores de desenvolvimento social para ampliar o escopo da análise.

Referências

- AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases, Proceedings of the 1993 ACM SIGMOD international conference on Management of data, p.207-216, Washington, D.C., United States. May 25-28, 1993.
- BRAGA, Luis Paulo Vieira. Introdução à Mineração de Dados. 2a edição. Rio de Janeiro: E-Papers, 2005.
- FAYYAD, U. PIATETSKY-SHAPIO, G. and SMYTH, P. From Data Mining to Knowledge Discovery in Databases. AI magazine. P 37-54, 1996.
- MARKOV, Z.; RUSSELL, I. An Introduction to the WEKA Data Mining System. Proceedings of the 11th annual SIGCSE conference on Innovation and technology in computer science education. P. 367 - 368. Bologna, Italy, 2006.
- MARTINHAGO, S. Descoberta de Conhecimento sobre o Processo Seletivo da UFPR. Dissertação de Mestrado. Curitiba, 2005.
- NAVEGA, Sergio. Princípios essenciais do Data Mining. Anais do Infoimagem 2002. Cenadem, 2002.
- PNUD. Objetivos de Desenvolvimento do Milênio. Disponível em: <http://www.pnud.org.br/odm/>. Acessado em: 16/12/2009.

- QUINLAN, J. R. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers, 1993.
- ROMÃO, W. Descoberta de conhecimento relevante em banco de dados sobre ciência e tecnologia. Tese (Doutorado) - Curso de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina, Florianópolis, 2002.
- SHANNON, C. E. A Mathematical Theory of Communication. Bell System Technical Journal. Vol. 27, pp. 379–423, 623–656, July, October, 1948.
- SILVA, M.P. Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka. Em Escola Regional de Informática RJ/ES, v. 1, pp. 19-21. Sociedade Brasileira de Computação, 2004.

Anexo 1 – Exemplo de aplicação do algoritmo ID3 para construção de árvore de decisão

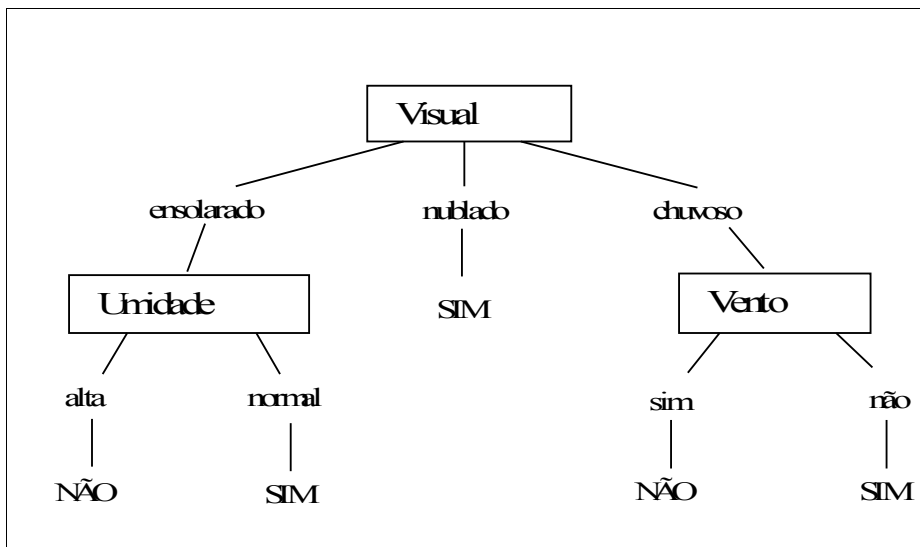
Descrição: Árvore de decisão para o problema “Jogar Golfe” considerando as condições climáticas

Fontes: Martinhago (2005) e arquivo-exemplo “weather.nominal” do Weka 3.6

Dados:

Visual	Temperatura	Umidade	Vento	Jogar
ensolarado	alta	alta	não	não
ensolarado	alta	alta	sim	não
nublado	alta	alta	não	sim
chuvoso	média	alta	não	sim
chuvoso	baixa	normal	não	sim
chuvoso	baixa	normal	sim	não
nublado	baixa	normal	sim	sim
ensolarado	média	alta	não	não
ensolarado	média	normal	não	sim
chuvoso	média	normal	não	sim
ensolarado	média	normal	sim	sim
nublado	média	alta	sim	sim
nublado	alta	normal	não	sim
chuvoso	média	alta	sim	não

Árvore de decisão:



Anexo 2 - Objetivos de Desenvolvimento do Milênio (ODM)

OBJETIVO	META	INDICADORES
ERRADICAR A POBREZA EXTREMA E A FOME	Reduzir à metade, entre 1990 e 2015, a proporção de pessoas com renda inferior a US\$1/dia	População em situação de pobreza extrema segundo linhas nacionais (custo da cesta básica de alimentos)
	Reduzir à metade, entre 1990 e 2015, a proporção de pessoas que sofrem de fome	Crianças menores de cinco anos com peso inferior ao normal População abaixo do nível mínimo de consumo de energia alimentar
ATINGIR O ENSINO BÁSICO UNIVERSAL	Garantir, para o ano 2015, que os meninos e meninas de todo o mundo possam concluir um ciclo completo do ensino fundamental	Taxa líquida de matrícula no ensino fundamental Jovens entre 15 e 19 anos de idade que concluíram o ensino fundamental
PROMOVER A IGUALDADE ENTRE OS SEXOS E A AUTONOMIA DAS MULHERES	Eliminar a disparidade entre os sexos no ensino fundamental e médio, se possível até 2005, e em todos os níveis do ensino até fins de 2015	Mulheres em relação aos homens no ensino fundamental, médio e superior
REDUZIR A MORTALIDADE INFANTIL	Reduzir em dois terços, entre 1990 e 2015, a mortalidade de crianças menores de cinco anos	Taxa de mortalidade de crianças menores de cinco anos Taxa de mortalidade infantil
MELHORAR A SAÚDE MATERNA	Reduzir em 75%, entre 1990 e 2015, as taxas de mortalidade materna	Taxa de mortalidade materna
COMBATER O HIV/AIDS, A MALÁRIA E OUTRAS DOENÇAS	Até 2015 ter detido e começado a reduzir a propagação do HIV/Aids	Homens e mulheres de 15 a 49 anos de idade que vivem com o HIV/Aids Uso de preservativos como percentual da taxa de uso de anticoncepcionais Relação entre a frequência escolar de crianças órfãs e a frequência escolar de crianças não órfãs de 10 a 14 anos de idade
	Até 2015 ter detido e começado a reduzir a incidência da malária e de outras doenças graves	Mortes associadas à malária População de zonas de risco de malária que aplica medidas eficazes de prevenção e tratamento desta doença Taxas de prevalência e mortalidade associadas à tuberculose Casos de tuberculose detectados e curados com tratamento intensivo sob observação direta

OBJETIVO	META	INDICADORES
GARANTIR A SUSTENTABILIDADE AMBIENTAL	Incorporar os princípios de desenvolvimento sustentável nas políticas públicas e programas nacionais e inverter a perda de recursos ambientais	Proporção da superfície coberta por matas Relação entre zonas protegidas para manter a diversidade biológica e a superfície total Uso de energia (equivalente em quilogramas de petróleo) por US\$1 do produto interno bruto (PPA) Emissões de dióxido de carbono per capita e consumo de clorofluorcarbonetos que esgotam a camada de ozônio (toneladas de PAO)
	Reduzir à metade, até 2015, a proporção da população sem acesso permanente e sustentável a água potável e a saneamento básico	Acesso sustentável a melhores fontes de abastecimento de água urbana e rural Acesso a melhores serviços de saneamento urbano e rural
	Até 2020, ter alcançado melhoria significativa na vida de pelo menos 100 milhões de habitantes de áreas degradadas	Proporção de famílias com acesso à posse segura da moradia
ESTABELECECER UMA PARCERIA MUNDIAL PARA O DESENVOLVIMENTO	Avançar no desenvolvimento de um sistema comercial e financeiro aberto, baseado em regras, previsível e não-discriminatório (inclui-se o compromisso de alcançar uma boa gestão dos assuntos públicos e a redução da pobreza em cada país e no plano internacional)	
	Atender às necessidades especiais dos países menos desenvolvidos, o que inclui o acesso livre de tarifas e cotas para as exportações de seus produtos, a melhoria do programa de alívio da dívida dos países pobres muito endividados e o cancelamento da dívida bilateral oficial, e a concessão de assistência oficial para o desenvolvimento (AOD) mais generosa aos países que mostraram determinação em reduzir a pobreza	
	Atender às necessidades especiais dos países sem acesso ao mar e dos pequenos Estados insulares em desenvolvimento (mediante o Programa de Ação para o Desenvolvimento Sustentável dos Pequenos Estados Insulares em Desenvolvimento e as decisões adotadas no vigésimo segundo período extraordinário de sessões da Assembleia Geral)	
	Tratar globalmente o problema da dívida dos países em desenvolvimento, mediante medidas nacionais e internacionais, de maneira a tornar a sua dívida sustentável a longo prazo	

Fonte: PNUD Brasil (<http://www.pnud.org.br/odm/>)

Anexo 3 – Árvores de Decisão produzidas pelos algoritmos de Classificação ID3 e J48 do Weka

Árvore ID3:

```

PGPN-M7 = 1: 2
PGPN-M7 = 2
| PEMF = 1
| | NAT = 1: null
| | NAT = 2: 1
| | NAT = 3
| | | PGPN-13 = 1: 1
| | | PGPN-13 = 2: 1
| | | PGPN-13 = 3: null
| | | PGPN-13 = 4: null
| | | PGPN-13 = 5: null
| | NAT = 4: 2
| | NAT = 5: null
| PEMF = 2: 4
| PEMF = 3: null
| PEMF = 4: null
| PEMF = 5: null
PGPN-M7 = 3
| PEMF = 1: 1
| PEMF = 2
| | NAT = 1: null
| | NAT = 2: 3
| | NAT = 3: 4
| | NAT = 4: null
| | NAT = 5: null
| PEMF = 3: null
| PEMF = 4: null
| PEMF = 5: null
PGPN-M7 = 4
| PEMF = 1: 1
| PEMF = 2: null
| PEMF = 3
| | NAT = 1: 1
| | NAT = 2: 1
| | NAT = 3: null
| | NAT = 4: null
| | NAT = 5: null
| PEMF = 4: 1
| PEMF = 5: 1
PGPN-M7 = 5: 1

```

Árvore J48:

```

PGPN-M7 = 1: 2 (3.0)
PGPN-M7 = 2
| NAT = 1: 1 (0.0)
| NAT = 2: 1 (2.0)
| NAT = 3: 1 (5.0/1.0)
| NAT = 4: 2 (5.0/2.0)
| NAT = 5: 1 (0.0)
PGPN-M7 = 3
| PEMF = 1: 1 (13.0)
| PEMF = 2: 3 (3.0/1.0)
| PEMF = 3: 1 (0.0)
| PEMF = 4: 1 (0.0)
| PEMF = 5: 1 (0.0)
PGPN-M7 = 4: 1 (27.0/1.0)
PGPN-M7 = 5: 1 (22.0)

```

Legenda:

PGPN-M7: Proporção de gestantes com mais de 7 consultas pré-natal
PGPN-13: Proporção de gestantes com 1 a 3 consultas pré-natal
PEMF: Proporção de empregados no mercado formal
NAT: Natalidade

Anexo 4 - Regras geradas pelo algoritmo Apriori do Weka

Parâmetros: Regras com confiança=1 e diretamente relacionadas com a taxa de mortalidade infantil

1. PEMF=1 PGPN-13=1 47 ==> TM-I=1 47
3. PGPN-M7=5 22 ==> TM-I=1 22
5. PGPN-13=1 PGPN-M7=5 22 ==> TM-I=1 22
8. NAT=1 21 ==> TM-I=1 21
10. PGPN-13=1 NAT=1 21 ==> TM-I=1 21
11. NAT=1 21 ==> PGPN-13=1 TM-I=1 21
12. PEMF=1 NAT=2 20 ==> TM-I=1 20
13. PEMF=1 PGPN-M7=4 19 ==> TM-I=1 19
14. PEMF=1 PGPN-13=1 PGPN-M7=4 18 ==> TM-I=1 18
15. PEMF=1 PGPN-13=1 NAT=2 18 ==> TM-I=1 18
17. PEMF=1 NAT=1 17 ==> TM-I=1 17
19. PEMF=1 PGPN-13=1 NAT=1 17 ==> TM-I=1 17
22. PEMF=1 PGPN-M7=5 16 ==> TM-I=1 16
24. PEMF=1 PGPN-13=1 PGPN-M7=5 16 ==> TM-I=1 16
28. PEMF=1 PGPN-M7=3 13 ==> TM-I=1 13
30. PGPN-M7=5 NAT=1 13 ==> TM-I=1 13
33. PGPN-13=1 PGPN-M7=5 NAT=1 13 ==> TM-I=1 13
34. PGPN-M7=5 NAT=1 13 ==> PGPN-13=1 TM-I=1 13
35. PEMF=1 PGPN-13=1 NAT=3 12 ==> TM-I=1 12
37. PEMF=1 PGPN-M7=5 NAT=1 11 ==> TM-I=1 11
39. PEMF=1 PGPN-13=1 PGPN-M7=5 NAT=1 11 ==> TM-I=1 11
42. PEMF=1 PGPN-M7=4 NAT=2 10 ==> TM-I=1 10
44. PEMF=1 PGPN-13=1 PGPN-M7=4 NAT=2 10 ==> TM-I=1 10
47. PEMF=1 PGPN-13=1 PGPN-M7=3 9 ==> TM-I=1 9