JHI JOURNAL OF HEALTH INFORMATICS

Pré-processamento para mineração de dados de pacientes com HIV

Data mining preprocessing for HIV positive

Procesamiento para la minerá de datos de pacientes con VIH

Amanda Rocha Chaves[1], Tatiane Aparecida Barroso Silvério[2], Valdene Alves Barroso[2]

Descritores: Mineração de Dados; Pré-processamento; Incidência de HIV/Aids

RESUMO

Objetivo: Descrever o processo de descoberta de conhecimento utilizando-se de técnicas de mineração de dados em uma base de dados sobre a incidência de HIV em microrregiões mineiras, trazendo consigo uma descrição das características de localidade e perfil de pacientes registrados no Centro de Testagem e Aconselhamento da cidade de Diamantina, com dados representativos de duas microrregiões do Vale do Jequitinhonha. **Método:** Aplicação de técnicas de pré-processamento de dados como limpeza e transformação utilizando a ferramenta weka. **Resultado:** Do conjunto de dados foram extraídos atributos relevantes e irrelevantes para descrição da incidência de HIV/Aids e contabilizadas estatísticas que descreveram de forma precisa as relações existentes entre os atributos selecionados. **Conclusão:** Conseguiu-se um pré-processamento de dados relacionados à incidência de HIV/Aids que resultou no entendimento da área de estudo, dos dados selecionados e gerou uma base que servirá para uma aplicação futura de mineração via algoritmos de aprendizado de máquina.

Keywords: Data Mining; Incidence; HIV

ABSTRACT

Objective: To describe knowlege discorery process using data mining techniques derived from HIV reports about microregion in Minas Gerais state, bringing with it a description of the characteristics of the locality and the profile of patients registered on Counseling and Testing Center of Diamantina city, with representative data of two microregion in Jequitinhonha Valley. **Method:** Application techniques for pre-processing data as cleaning and processing using weka tool. **Results:** From dataset relevant and irrelevant attributes were extracted to describe the incidence of HIV/AIDS and recorded statistics that accurately described the relations between the selected attributes.**Conclusion:** It was shown a suitable pre-processing data from HIV and produced files in order to aply for mining algorithms across machine learning altorithms. Across of pre-processing data about incidence of HIV/AIDS gave an understanding of the field study, the selected data and data generated to serve as a basis for future mining application via machine learning algorithms.

Descriptores: Minéria de Datos; Incidencia; VIH

RESUMEN

Objetivo: Describir el proceso de descubrimiento de conocimiento utilizando las técnicas de minería de datos en una base de datos sobre la incidencia del VIH en micro regiones de Minas Gerais, trayendo consigo una descripción de las características de la localidad y el perfil de los pacientes registrados en el Centro de Pruebas y consejo de la ciudad de Diamantina, con datos representativos de dos micro regiones de Jequitinhonha Valley. **Método:** Aplicación de técnicas de pre-procesamiento como la limpieza y el procesamiento de datos utilizando la herramienta Weka. **Resultado:** Los datos que figuran atributos relevantes e irrelevantes se extrajeron para describir la incidencia del VIH / SIDA y registra las estadísticas que describen con precisión las relaciones existentes entre los atributos seleccionados. **Conclusión:** Se logró un pre-procesamiento relacionado con la incidencia del VIH / SIDA ha dado lugar a la comprensión de la zona de estudio, los datos y los datos generados seleccionados para servir como base para la aplicación de minería de futuro a través de algoritmos de aprendizaje automático.

[1] *Mestre em Ciência da Computação, Universidade Federal dos Vales do Jequitinhonha e Mucuri – UFVJM, Diamantina (MG), Brasil.*
[2] *Graduando em Ciência e Tecnologia; Universidade Federal dos Vales do Jequitinhonha e Mucuri – UFVJM, Diamantina (MG), Brasil.*

Autor Correspondente: **Amanda Rocha Chaves**
e-mail: **amanda.chaves@ict.ufvjm.edu.br**

Artigo recebido: **18/02/2014**
Aprovado: **21/08/14**

INTRODUÇÃO

Os primeiros casos registrados de HIV/Aids (*Acquired Immunodeficiency Syndrome*) datam de 1977 e foram registrados nos EUA, Haiti e África Central. No Brasil, o primeiro caso data de 1982, o que nos traz quase quatro décadas da doença e um aumento significativo dos registros de sua ocorrência, totalizando até 2012 aproximadamente 660.000 casos dos quais cerca de 240 mil resultaram em óbito diretamente relacionados à doença[1].

Nos primórdios da doença, a mesma estava associada a grupos considerados de risco como homossexuais, profissionais do sexo e usuários de drogas o que levou a uma atmosfera do preconceito sobre tais indivíduos[2]. O mapeamento da doença vem sendo exaustivamente realizado e registrado e a partir desta documentação tem se conseguido executar políticas públicas para controle de sua disseminação e manutenção da saúde dos infectados. Neste contexto, ferramentas que auxiliem a tomada de decisão de órgãos de gestão em saúde podem ser utilizadas para melhoria destas ações públicas.

Técnicas de Aprendizado de Máquina (AM) têm sido utilizadas com sucesso para a solução de vários problemas reais em áreas distintas de conhecimento como bioinformática, energia, telecomunicação, saúde, dentre outras. A área de saúde, em particular, tem aceitado utilizar ferramentas computacionais baseadas em AM para auxílio a exames clínicos, monitoramento do estado do paciente, etc[3].

Considerando que um mapeamento da ocorrência de qualquer doença seja fator importante para criar políticas públicas de controle e tratamento, o presente trabalho tem por objetivo descrever o processo de descoberta de conhecimento utilizando-se de técnicas de mineração de dados em uma base de dados sobre a incidência de HIV em microrregiões mineiras, trazendo consigo uma descrição das características de localidade e perfil de pacientes registrados no Centro de Testagem e Aconselhamento(CTA) da cidade de Diamantina, com dados representativos de duas microrregiões do Vale do Jequitinhonha.

Segundo o modelo de descoberta de conhecimento denominado CRISP-DM (*Cross Industry Standard Process for Data Mining*), a mineração de dados é definida como um processo não linear subdividido em seis fases que se inter-relacionam, a saber: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e utilização, sendo que a execução de uma das fases pode gerar um ciclo entre as mesmas com intuito de melhoramento[4]. Este trabalho se resume às três primeiras fases deste processo, que podem ser consideradas também como a fase de pré-processamento segundo Faceli, et al.[3], a qual inclui tudo que é realizado antes de se minerar os dados através de algoritmos de aprendizado de máquina.

MÉTODOS

Considerando as três primeiras fases da descoberta de conhecimento descritas anteriormente, realizou-se um estudo observacional, transversal e exploratório da base de dados extraída de prontuários registrados no Centro de Testagem e Aconselhamento do município de Diamantina, no ano de 2013[1]. Os dados cedidos pelo CTA foram utilizados por Lima[1] em seu trabalho de dissertação de mestrado, sob a aprovação do comitê de Ética da UFVJM com nº de protocolo (059/12). Vale ressaltar que os mesmos foram utilizados neste trabalho exclusivamente para pesquisa e nenhuma informação de identidade foi extraída dos prontuários afim de se preservar a integridade dos pacientes.

Os dados analisados foram disponibilizados em formato textual e continham informações de 112 registros de pacientes que foram notificados com incidência de HIV. Os registros passaram por um processo de limpeza e adequação ao processo de mineração utilizando-se do ambiente WEKA (*Waikato Environment for Knowledge Analysis*) na sua versão 3.6. O WEKA foi escolhido para auxiliar na extração de conhecimento da base supracitada por três motivos principais: é uma ferramenta de referência na área de mineração e, principalmente, é desenvolvida em uma linguagem de programação que prevê portabilidade (no caso do Java) e é de fácil acesso por possuir seu código aberto e ser de livre uso. O mesmo se encontra disponível em (http://

Tabela 1 - Atributos considerados após pré-processamento

Atributo	Valores possíveis
Idade	Acima de 12 anos
Sexo	1-Feminino, 2-Masculino
Estado Civil	1-Casado, 2-Solteiro, 3-Amasiado, 4-Viúvo
Raça/Cor	1-Branca, 2-Preta, 3-Parda
Profissão	1-lavrador, 2-professor, 3-serviços gerais, 4-garimpeiro, 5-doméstica, 6-estudante, 7-pedreiro, 8-conferente, 9-vendedor, 10-auxiliar administrativo, 11-funcionário público, 12-ajudante geral, 13-matemático, 14-escriturário, 15-motorista, 16-do lar, 17-técnica em enfermagem, 18-operador de sistema, 19-ajudante de pedreiro
Escolaridade	1-analfabeto, 2-educação infantil, 3-fundamental incompleto, 4-fundamental completo, 5-superior incompleto, 6-superior completo, 7-ensino médio incompleto, 8-ensino médio completo
Município de origem	1-Diamantina, 2-Felício dos Santos, 3-Itamarandiba, 4-Minas Novas, 5-Capelinha, 6-São Gonçalo do Rio Preto, 7-Turmalina, 8-Presidente Kubitschek, 9-Datas, 10-Coluna, 11-Congonhas do Norte, 12Gouveia, 13-Serro, 14-Senador Modestino Gonçalves, 15-Chapada do Norte, 16-Couto de Magalhães de Minas, 17-Aricanduva
Microrregião	1-Diamantina, 2-Minas Novas, Capelinha e Turmalina
Categoria de exposição	Homossexual, heterossexual, bissexual

www.cs.waikato.ac.nz/ml/weka).

O pré-processamento inicial resultou na seleção de nove atributos, conforme ilustra a Tabela 1.

Após a escolha dos atributos descritos na Tabela 1 foi gerado um arquivo com texto puro no formato ARFF(*Attribute Relation-File Format*), que é requisito para utilização dos recursos contidos no WEKA[5]. Este arquivo é formado por três partes: relação, atributos e dados e pode ser precedido por uma sessão de comentários cujo formato de escrita linear inicia com o caractere % e tem por objetivo descrever de forma linguística o conteúdo do arquivo, segue ao comentário o conteúdo propriamente utilizado dentro do WEKA. A primeira linha útil do arquivo inicia-se com o texto @relation seguido de um nome que representa a relação ou situação a ser analisada. As linhas seguintes iniciadas pela expressão @attribute são seguidas do nome do atributo e de seu tipo que pode ser numérico (exemplo: idade) ou nominal (exemplo: sexo: feminino ou masculino). Após a seção de atributos do arquivo ARFF segue uma linha delimitada pela expressão @data e posteriormente todas as linhas seguintes representam individualmente um exemplo de ocorrência da relação representada, contendo os valores possíveis para os atributos na ordem em que foram apresentados anteriormente. Finalmente, o arquivo gerado foi denominado incidenciaHiv.arff.

Normalmente a realização da mineração de dados através do uso de algoritmos de aprendizado de máquina se resumem a dois tipos de tarefas: predição e descrição. Segundo Faceli, et al.[3], a primeira tarefa se propõe a encontrar, a partir de dados de treinamento, no caso, atributos de entrada, um rótulo ou valor que caracterize um novo exemplo. Neste caso dado um arquivo ARFF como o descrito acima, o último atributo é considerado atributo alvo e a tarefa segue o padrão de aprendizado denominado supervisionado. Já tarefas de descrição têm como objetivo relatar ou analisar um conjunto de dados e não utilizam um atributo de saída, seguindo, portanto o paradigma de aprendizado não supervisionado. Exemplo dessa tarefa é identificar regras de associação que criem uma relação entre conjuntos de atributos.

Foi utilizado o módulo *explorer* (Figura 1) do WEKA, o qual permite visualizar de formas diversas os dados registrados no arquivo supracitado e extrair dentre outras medidas a frequência absoluta de cada atributo de forma discretizada.

RESULTADOS E DISCUSSÃO

Como o objetivo deste trabalho foi utilizar técnicas de pré-processamento de dados para descoberta de conhecimento, fez-se necessário obter uma base de dados e prepará-la para um futura tarefa de mineração de dados. A metodologia descrita conseguiu selecionar alguns atributos e eliminar aqueles considerados irrelevantes ou ambíguos. Além disso, foram extraídas medidas estatísticas que descrevessem de forma precisa as relações existentes entre os atributos selecionados, de forma a pontuar a

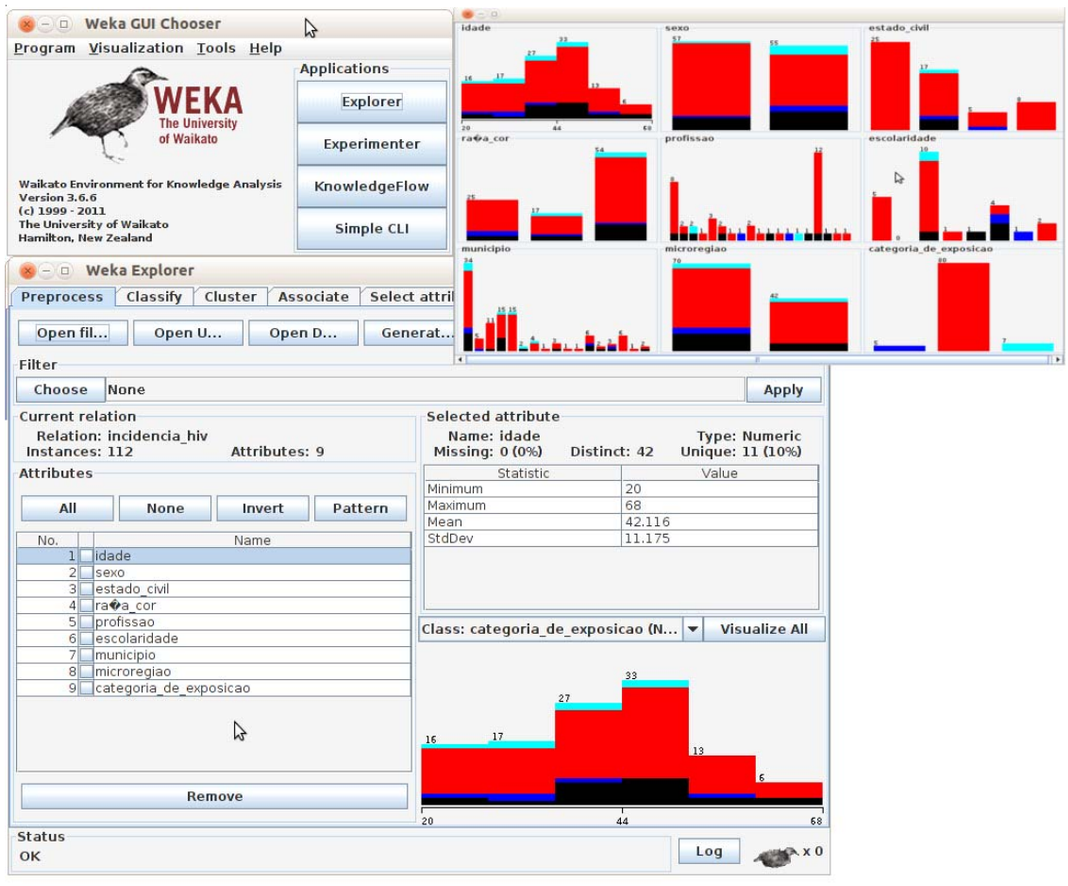

Figura 1 - Tela de apresentação e utilização do recurso *explorer* do ambiente WEKA

incidência do HIV/Aids nas duas microrregiões mineiras supracitadas. Dentre estas medidas temos frequência, localização de determinado atributo ou tendência central, dispersão ou espalhamento e distribuição.

Segundo Facelli, et al. [3] a medida de frequência mede a proporção de vezes que um atributo assume um dado valor em um determinado conjunto de dados. Ela pode ser aplicada a valores tanto numéricos quanto simbólicos, sendo muito utilizada nestes últimos. As medidas restantes são úteis para dados numéricos, o que nos levou a fazer uma adaptação dos dados para esse tipo.

Percebeu-se que muitos dos dados estavam incompletos ou foram inseridos em um atributo não adequado, como por exemplo, para o atributo escolaridade foram inseridos dados relacionados à profissão do paciente. O que implicou em uma decisão de projeto: subdividir atributos ambíguos por atributos simples em função dos conteúdos encontrados, como dividir escolaridade em dois atributos: escolaridade e profissão. Isso acabou gerando dois atributos incompletos. O processamento de dados incompletos pode prejudicar o desempenho de alguns algoritmos de aprendizado e os mesmos podem ser derivados de vários fatores adversos como distração na hora de preencher o formulário, falta de interesse em preencher determinados atributos por considerá-los vexativos ou até mesmo erro no processo de coleta e transcrição dos dados. Para esses casos se aconselha retirar instâncias com atributos ausentes, preencher manualmente valores para os atributos com valores ausentes ou utilizar alguma heurística para definir valores automaticamente ou empregar algoritmos que lidem bem com valores ausentes, como por exemplos algoritmos de árvore de decisão.

Neste trabalho, foi atribuído, conforme recomenda Hall, et al[6] , um valor para o atributo ausente representado pelo símbolo '?'. Logo, para cada instância da base de dados representada no arquivo ARFF cujo atributo estivesse sem conteúdo foi adicionada uma interrogação na posição relativa a tal atributo. Um fato importante que se deve considerar é que as informações obtidas na exploração de dados são de fundamental importância para selecionar a técnica mais apropriada no pré-processamento e aprendizado gerado pela mineração de dados. Dependendo do conteúdo presente nos atributos haverá ou não um algoritmo de aprendizado mais adequado ao seu processamento. Pensando nisso e tendo como proposta futura a mineração destes dados a partir da aplicação de um algoritmo de agrupamento que pudesse identificar conjuntos de atributos que juntos melhor identifiquem a localização (temporal, social, profissional, etc) da incidência de HIV, decidiu-se pela realização de uma conversão de conteúdos simbólicos

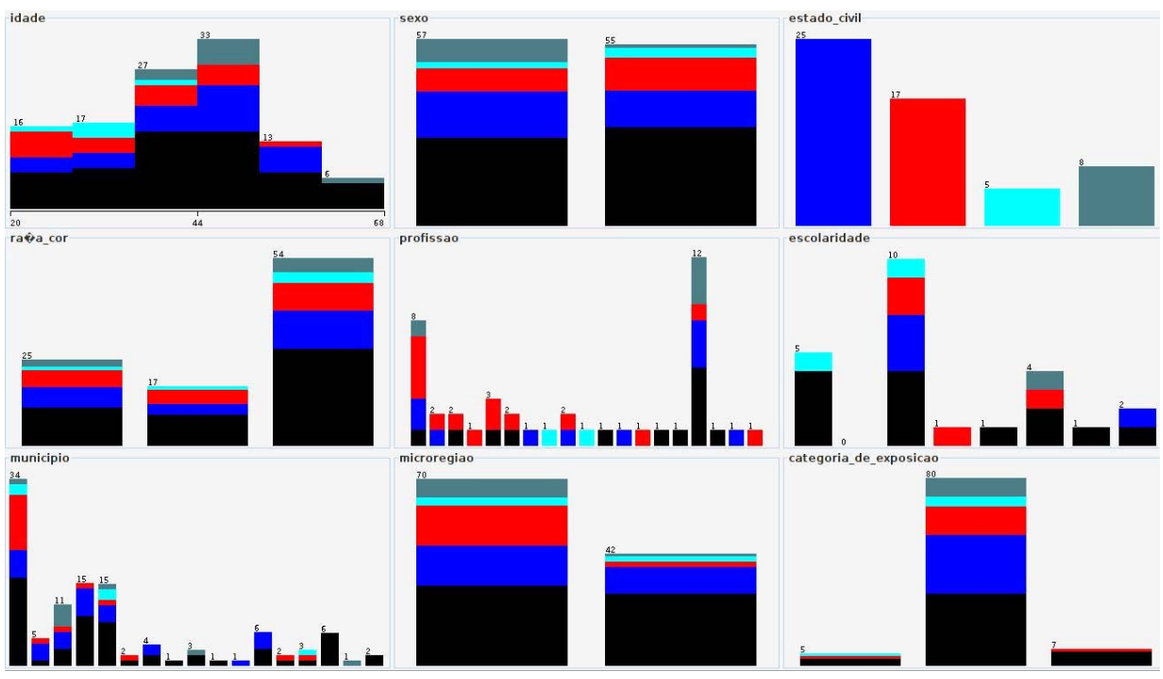

Figura 2 - Cruzamento de atributos

Tabela 2 - Faixa etária de incidência de HIV/Aids

Faixa etária (anos)	Quantidade (%)
20 à 28 anos	14
28 à 36 anos	15
36 à 44 anos	24
44 à 52 anos	30
52 à 60 anos	12
60 à 68 anos	5

para numéricos visto que técnicas de aprendizagem como redes neurais e máquinas de vetor de suporte, lidam apenas com dados numéricos[3] e estes algoritmos poderiam ser aplicados para o agrupamento.

Após a limpeza e transformação dos dados determinou-se a distribuição de frequência dos mesmos bem como a determinação de algumas relações, conforme ilustrado na Figura 2, a saber: a faixa etária inter-relacionada com os atributos estado civil, sexo e categoria de exposição; escolaridade e profissão; localidade de

residência; raça, etc..

Partindo da análise do atributo idade, conseguimos identificar seis faixas etárias, conforme exibido na Tabela 2, identificando portanto a menor idade como 20 anos e a maior 68; a maior incidência da doença se deu entre adultos de 44 a 52 anos. A média de idade da incidência da doença se deu aos 42 anos e o desvio padrão foi de 11 anos.

Um fato que se destacou nos dados foi constatar a maior incidência de HIV/Aids no grupo de heterossexuais (cerca de 86%), o que corrobora com a derrubada do mito de que os homossexuais são predominantes na incidência desta doença. Isso indica para os gestores públicos que os critérios de classificação de grupos de risco devem ser reavaliados. Fato esse que pode ser usado ainda nos trabalhos de prevenção à disseminação do vírus e no tratamento dos infectados.

Do total de indivíduos, 49% são homens e 51% mulheres, sendo que daqueles que declararam seu estado civil, 45% dos pacientes eram solteiros e 30% casados. Lembramos que as incidências registradas para estado civil totalizaram 50% de abstenção, ou seja, de 112 registros 57 não incluíram este dado no prontuário. Pode-se verificar que o número de viúvas é maior, o que nos leva a supor que possivelmente o cônjuge tenha trazido a enfermidade para o seio familiar e que seu falecimento tenha ocorrido em função da doença.

Sobre a categoria de exposição, 82% se descreveram em uma das categorias: homossexual, heterossexual ou bissexual, enquanto 18% se abstiveram. A incidência de HIV/Aids ocorreu em todas as faixas etárias entre os heterossexuais, já entre os homossexuais a incidência não foi registrada nas faixas etárias de 44 a 52 anos e de 60 a 68. E nesta última, não houve incidência também entre os bissexuais.

A incidência do HIV/Aids se acentuou entre pessoas com pouca renda e escolaridade, mas considerando que apenas 38% dos pacientes relataram sua profissão e 21% sua escolaridade, esses valores são pouco representativos. Apesar disso, vale ressaltar que dentre os registrados a maioria se encontra entre domésticas, do lar e lavradores, com ensino fundamental incompleto.

Em relação à localidade de residência, a maior incidência se deu nesta ordem nos municípios de Diamantina (30%), Minas Novas (13%) e Capelinha (13%) e a menor ocorrência foi registrada igualmente nas cidades de Coluna, Couto Magalhães de Minas, Presidente Kubitschek e Congonhas do Norte.

Por fim, dentre as pessoas declaradas como brancas a maior incidência de HIV/Aids ocorreu entre mulheres, enquanto entre as pardas ou negras, se deu entre homens. Em todas as faixas etárias, houve registro de brancos e pardos acometidos pela doença, enquanto na faixa entre 52 e 60 anos não registrou-se nenhum negro.

CONCLUSÃO

Neste artigo foi apresentado um trabalho de pré-processamento de dados relacionados à incidência de HIV/Aids em duas microrregiões mineiras que resultou em três etapas: entendimento da área de estudo, entendimento dos dados selecionados e preparação destes para aplicação de mineração. Esta tarefa foi auxiliada pelo ambiente WEKA que poderá ser utilizado para aplicação futura de algoritmos de aprendizado.

A descrição dos dados aponta para a utilização do uso de técnicas de AM não supervisionado já que, em contraposição ao treinamento supervisionado e ao treinamento por reforço, não há aqui nenhuma saída desejada explícita ou avaliação externa da saída produzida para cada dado de entrada.

Destacou-se uma desmistificação da ocorrência de HIV/Aids entre grupos outrora considerados de risco, como, por exemplo, os homossexuais, e o aparecimento de um novo grupo: heterossexuais casados.

Além disso, percebeu-se que muitos cruzamentos de dados deverão ainda ser analisados como por exemplo determinar o relacionamento entre o estado civil e profissão na incidência de HIV/Aids, dentre outros.ki0j;ki

REFERÊNCIAS

1. Lima WA. Vigilância em saúde em tempos de HIV/AIDS: sistemas de informação no serviço de atenção especializada [dissertação]. Diamantina: Universidade Federal dos Vales do Jequitinhonha e Mucuri; 2013.
2. Lelis RT, Garbin CAS, Garbin AJI, Soares GB. Vivendo com HIV/AIDS: estudo da ocorrência de discriminação nos serviços de saúde. Rev. Bras. Pesq. Saúde. 2012;14(4):22-8.
3. Faceli K, Lorena AC, Gama J, Carvalho ACPLF. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC; 2011.
4. Wirth R, Hipp J. CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining (PADD'00); 2000 Aug 27-29; New York; 2000. p. 29-39.
5. Santos R. Weka na munheca: um guia para uso do weka em scripts e integração com aplicações em Java. 2005 Abr [citado 2014 fev]. Disponível em: http://www.lac.inpe.br/~rafael.santos/index.jsp?tag=datamining
6. Hall M, Witten I, Frank E. Data mining: practical machine learning tools and techniques. 3a ed. San Francisco: Morgan Kaufmann; 2011.