FUNDAÇÃO CENTRO DE ANÁLISE, PESQUISA E INOVAÇÃO TECNOLÓGICA
INSTITUTO DE ENSINO SUPERIOR FUCAPI
COORDENAÇÃO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Uso de técnicas de *Data Mining* na classificação de sorotipos da dengue

Jeferson Barros Alves

Manaus - AM
Novembro de 2015

Jeferson Barros Alves

Uso de técnicas de *Data Mining* na classificação de sorotipos da dengue

Monografia apresentada ao Curso de Graduação em Ciência da Computação do Instituto de Ensino Superior FUCAPI - CESF como requisito parcial para a obtenção do Título de Bacharel em Ciência da Computação. Área de concentração: Banco de Dados.

Orientador

Márcio Palheta Piedade, M.Sc.

Manaus - AM

Novembro de 2015

Lista de figuras

Figura 1 – Etapas do processo KDD. Adaptado de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) 14
Figura 2 – Abordagem geral para a construção de um modelo de classificação. Adaptado de (TAN; STEINBACH; KUMAR, 2009) 17

Lista de tabelas

Tabela 1 – Descrição dos atributos do conjunto de dados 19

Lista de abreviaturas e siglas

KDD	Knowledge Discovery in Databases
WEKA	Waikato Environment for Knowledge Analysis
SVM	Support Vector Machine
SMO	Sequential Minimal Optimization

Sumário

1 INTRODUÇÃO 7

1.1 Especificação do problema 7

1.2 Objetivos 8

1.3 Justificativa 9

1.4 Trabalhos Relacionados 10

1.5 Metodologia de Desenvolvimento 11

1.6 Estruturação da Monografia 12

2 REFERENCIAL TEÓRICO 13

2.1 Descoberta de Conhecimento em Base de Dados 13

2.2 Data Mining 14

2.3 Aprendizagem Não Supervisionada 15

2.4 Técnicas de Classificação 16

2.4.1 Árvore de Decisão 17

2.4.2 Bayes 17

2.4.3 Máquinas de Vetores de Suporte 18

2.5 WEKA 18

3 BASE DE DADOS 19

3.1 Dados Abertos 19

3.2 SINAN 19

4 METODOLOGIA 21

5 RESULTADOS 22

5.1 Medida de Desempenho dos Algoritmos 22

5.1.1 J48 22

5.1.2 RandomTree 22

5.1.3 REPTree 22

5.1.4 NaiveBayes 22
5.1.5 SMO 22
5.2 Resultados Comparativos Entre os Modelos 22

6 CONCLUSÃO 23

Referências 24

ANEXO A – DICIONÁRIO DE DADOS - SINAN ONLINE 26

1 Introdução

Com a grande quantidade de informações geradas e armazenadas por meio do avanço tecnológico das últimas décadas, houve um acúmulo gigantesco de informações. Com isso surgiu uma abordagem de técnicas e ferramentas que buscam transformar dados, denominada Descoberta de Conhecimento em Base de Dados (knowledge Discovery in Databases - KDD), que foi proposto em 1989 com o objetivo de analisar os dados de uma base para que de alguma forma possa ser extraído conhecimento útil (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A computação tem apoiado o desenvolvimento da medicina em diversas áreas: em sistemas de apoio a coleta de dados clínicos e exames por imagens, na organização das informações obtidas, entre outras. Desta maneira, esse grande volume de dados é uma valiosa fonte de conhecimento que pode ser utilizada para o auxílio ao diagnóstico médico. Assim, é grande a importância do desenvolvimento de técnicas que permitam a descoberta de conhecimento em base de dados médicos para apoiar o médico em sua tarefa diária de tomada de decisões, aumentando a precisão, a confiabilidade e eficiência dos diagnósticos elaborados pelo especialista (COSTA, 2012).

1.1 Especificação do problema

(COSTA, 2012) mostrou que o uso de suporte computacional no atendimento médico é comum nos dias de hoje e gera grandes informações digitalizadas dos pacientes. Tais informações armazenadas em base de dados podem representar boas fontes de conhecimento, que pdoerão ser utilizadas no auxílio de diagnóstico médico. Segundo (BEZERRA, 2009), no passado, o médico detinha conhecimento suficiente para dispensar quase todo o cuidado necessário ao paciente.

Ao longo dos últimos anos a mineração de dados tem sido cada vez mais utilizada na literatura médica. No entanto, a sua aplicação à analise de dados médicos tem sido relativamente limitada, tendo em vista que a aplicação prática pode explorar o conhecimento disponível no contexto clínico e explicar decisões propostas, uma vez que os modelos são utilizados para apoiar decisões (BELLAZZI; ZUPAN, 2008).

É grande a importância do desenvolvimento de técnicas que permitam a descoberta de conhecimento em bases de dados médicos para apoiar o médico em sua tarefa diária de tomada de decisões, aumentando a precisão, a confiabilidade e eficiência dos diagnósticos elaborados pelo especialista (COSTA, 2012).

A dengue é hoje uma das doenças com maior incidência no Brasil, atingindo a população de todos os estados, independente de classe social. Nesse cenário, torna-se imperioso que um conjunto de ações para a prevenção da doença seja intensificado, permitindo assim a identificação precoce dos casos de dengue, a tomada de decisões e a implementação de medidas de maneira oportuna a fim de principalmente evitar óbitos. Preservar a vida humana é obrigação de todos. "COLOCAR FONTE DO MINISTÉRIO DA SAÚDE"

1.2 Objetivos

Construir um classificador que utilize dentre as diversas técnicas de classificação em Data Mining a que mais se adequa ao modelo de dados utilizados na ficha de investigação de dengue do SINAN.

Objetivos Específicos:

- Realizar experimentos com classificadores mais utilizados na literatura.
- Identificar qual classificador é mais adequado para o trabalho.
- Utilizar atributos mais relevantes para aplicação da técnica selecionada
- Tratar o conjunto de dados para aplicação do classificador

- Construir o modelo classificador
- Validar e interpretar os resultados obtidos no processo de classificação

Como os objetivos indicam ação, recomenda-se que eles sejam definidos por meio de verbos, tais como analisar, avaliar, caracterizar, discutir, diagnosticar, investigar, implantar, pesquisar, realizar, determinar, etc.

1.3 Justificativa

Na busca constante pela melhoria do atendimento médico hospitalar, as técnicas de mineração de dados estão sendo aplicadas nas bases de dados dos pacientes para descoberta de padrões, seja como ferramenta para auxílio ao diagnóstico ou descobertas que ajudem a alta direção nos hospitais a tomar atitudes com relação aos resultados obtidos (ROBERTINSON, 2012) "DOCUMENTO NÃO ENCONTRADO".

O desenvolvimento de técnicas que permitam a descoberta de conhecimento em bases médicas possui grande importância para apoiar o médico em sua tarefa diária de tomada de decisões, aumentando a precisão, a confiabilidade e a eficiência dos diagnósticos elaborados pelo especialista. Esse apoio computacional pode atuar como uma junta médica virtual, ao trazer para o especialista o conhecimento armazenado em exames e diagnósticos relacionados (COSTA, 2012).

Com o objetivo de estudo de métodos de suporte ao diagnóstico médico a partir de informações referentes aos sintomas dos pacientes conforme proposto por (JUNIOR; PIEDADE, 2013), observamos a necessidade de continuar o trabalho utilizando a técnica de regras de associação em outro conjunto de dados da mesma base.

1.4 Trabalhos Relacionados

No trabalho de (SHAKIL; ANIS; ALAM, 2015), observamos que o foco principal do trabalho foi a classificação de dengue, onde inicialmente aplicou-se a classificação nos datasets iniciais para identificar qual algoritmo obteve melhor resultado. O experimentos releram que os algoritmos Naive Bayes e J48 obtiveram melhores resultados, onde as maiores contribuições do trabalho foram:

- A extração da acurácia de classificação para predição do diagnóstico de dengue.
- Comparação de diferentes algoritmos de mineração no conjunto de dados de dengue.
- Identificar os algoritmos com melhor desempenho para previsão dos diagnósticos.

No trabalho de (THITIPRAYOONWONGSE; SURIYAPHOL; SOONTHORNPHISAJ, 2012), foi utilizado a técnica de classificação conhecida como Árvore de Decisão, aplicado a um conjunto de dados temporais contendo dados clínicos e laboratoriais. Utilizaram dois conjuntos de dados com mais de 400 atributos. Nos experimentos os dados foram divididos em: conjunto de teste e treinamento. Onde o objetivo principal deste trabalho foi identificar o dia zero da dengue, pois esta previsão torna-se crítica, tendo em vista que o dia zero tem forte relação com o melhor tratamento do paciente.

No trabalho (SANTOS; NETO,), foi desenvolvido um aplicativo para interação dos profissionais da saúde com os dados do SINAN. Onde resultou uma aplicação em java que implementa algoritmos de classificação e regras de associação da base de dados do SINAN utilizando a API WEKA. Como resultado esta ferramente possibilitou o auxílio no diagnóstico de pacientes.

1.5 Metodologia de Desenvolvimento

Com base na metodologia utilizada no processo de Descoberta de Conhecimento em Base de Dados (Knowledge Dicovery in Databases), a realização desta pesquisa será composta das seguintes etapas:

- A primeira etapa deste trabalho baseada na revisão bibliográfica da literatura relacionada a mineração de dados, para determinar outras pesquisas similares a esta, onde são aplicadas as técnicas de mineração de dados em base de dados médicos para descoberta de conhecimento novo e relevante, que de alguma forma possa auxiliar em diagnósticos do médico especialista.

- Na segunda etapa, ocorrerá a extração, manutenção e pré-processamento dos dados, onde realizaremos a análise de base de dados para extração de informações de atendimentos de pacientes necessários à pesquisa. Em todos os dados ocorrerá um processo de normalização, onde os sintomas relacionados serão dispostos em uma única linha e nos dados ausentes o preenchimento será realizado de forma padrão utilizando o sinal de "?".

- Na quarta etapa, pode-se executar algoritmos para descoberta do que mais se aproxima do esperado e na extração de informações propriamente dita. O objetivo desta etapa será encontrar padrões na base de dados que foi pré-processada, utilizando a ferramenta de mineração de dados WEKA, um software open-source amplamente utilizado por implementar diversos algoritmos de mineração de dados (HALL et al., 2009).

- A quinta e última etapa do processo ocorrerá a análise dos resultados obtidos com a a aplicação dos algoritmos de mineração de dados. Encontrar padrões de classificação que sejam relevantes para a pesquisa em questão.

1.6 Estruturação da Monografia

Além deste capítulo introdutório, este trabalho está organizado da forma descrita à seguir. No Capítulo 2, descrevemos conceitos básicos relacionados a compreensão deste trabalho, bem como referencial teórico. No Capítulo 3, apresentaremos os experimentos juntamente com os resultados obtidos. Finalmente no Capítulo 4, apresentaremos as conclusões observadas e sugestões de trabalhos futuros da pesquisa.

2 Referencial Teórico

Neste capítulo apresenta-se uma visão geral dos conceitos e procedimentos envolvidos em KDD e consequentemente em Data Mining. Apresentam-se também os trabalhos encontrados na literatura sobre assunto que são relevantes para o desenvolvimento deste trabalho.

2.1 Descoberta de Conhecimento em Base de Dados

Segundo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), a Descoberta de Conhecimento em Base de Dados é um processo não trivial de identificações de novos padrões, válidos e potencialmente úteis.

Em contrapartida, no trabalho de (THOMÉ, 2002), define-se KDD como sendo a busca de extração de conhecimento de bases de dados utilizando-se de técnicas e algoritmos que realizam a mineração dos dados para trabalhar e descobrir relações.

O processo de descoberta de conhecimento ocorre quando ocorre um conjunto de padrões que são semelhantes, e que podem levar a construção de um modelo. Este processo é formado por 5 etapas: seleção, pré-processamento, transformação, a mineração de dados e a interpretação dos dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Além do processo, o conhecimento que se deseja buscar deve estar de acordo com três características: deve ser correto, deve ser compreensível para o usuário e deve ter de alguma forma utilidade para o usuário (FREITAS, 2000).

A seguir serão detalhadas as etapas do processo de KDD de acordo com o apresentado na figura 1.

A etapa de seleção dos dados inicia com a definição do objetivo e mapeamento dos grupos ou conjuntos de informações que serão utilizados.

O pré-processamento é responsável pelo tratamento de ruídos e dados

incompletos.

A transformação tem como objetivo selecionar as principais características que serão utilizadas para representar os dados, ou seja, os dados devem ser selecionados de modo que sejam os mais úteis para o modelo proposto.

A etapa de mineração de dados é o momento em que serão escolhidos os algoritmos que mais se ajustam ao objetivo que se quer extrair da base de dados. Além disso, nesta fase são escolhidos os melhores parâmetros para que, no momento do processamento, os resultados sejam os mais rápidos e precisos possíveis.

Ao final do processo teremos a etapa de interpretação e avaliação dos resultados, onde o conhecimento extraído da base de dados é representado por padrões.

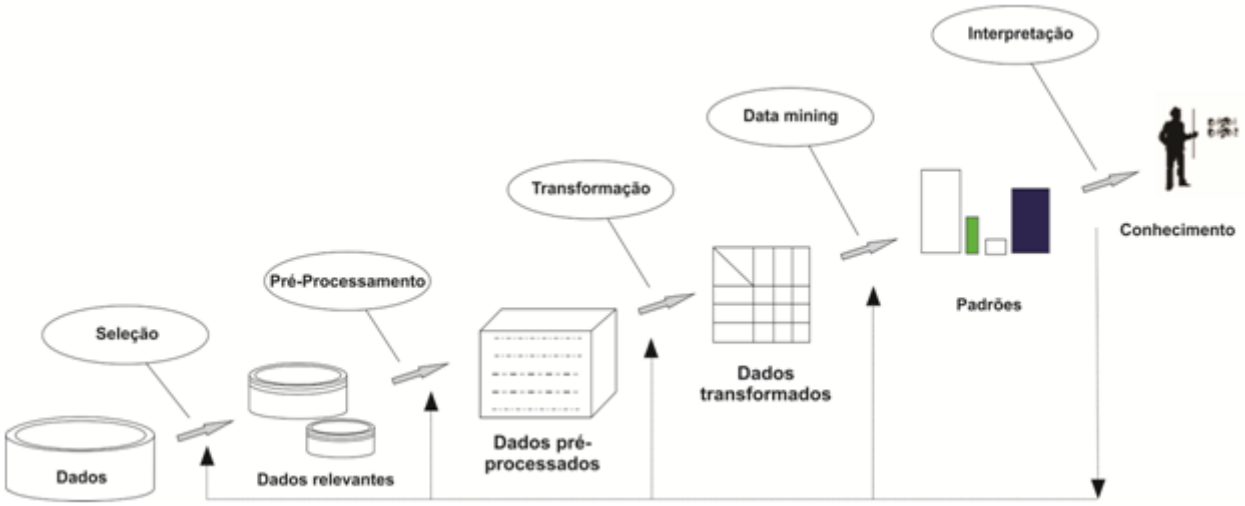

Figura 1 – Etapas do processo KDD. Adaptado de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

Inicialmente, é necessário definir que tipo de conhecimento se deseja extrair da base de dados, pois a técnica que será utilizada para a mineração de dados depende do objetivo a que se quer chegar (DAMASCENO, 2005).

2.2 Data Mining

De acordo com (ADRIAANS; ZANTINGE,), existe uma confusão entre os termos *Data Mining* e KDD, podendo ser usadas até como sinônimos em algumas situações. Em contrapartida, (BERRY; LINOFF, 1997), definiu como um processo de exploração e análise, de uma grande quantidade de dados, por meio automático ou semiautomático, com o propósito de descobrir regras e padrões significativos.

É uma técnica que faz parte de uma das etapas da descoberta de conhecimento em banco de dados. É uma área de pesquisa multidisciplinar, incluindo principalmente as tecnologias de banco de dados, inteligência artificial, estatística, reconhecimento de padrões, sistemas baseados em conhecimento, recuperação da informação, computação de alto desempenho e visualização de dados.

Em termos gerais, segundo ELMASRI (2002) "PROCURAR FONTE", a técnica de Data Mining compreende os seguintes propósitos:

- Previsão – pode mostrar como certos atributos dentro dos dados irão comportar-se no futuro;
- Identificação – padrões de dados podem ser utilizados para identificar a existência de um item, um evento ou uma atividade;
- Classificação – pode repartir os dados de modo que diferentes classes ou categorias possam ser identificadas com base em combinações de parâmetros;
- Otimização do uso de recursos limitados, como tempo, espaço, dinheiro ou matéria-prima e maximizar variáveis de resultado como vendas ou lucros sob um determinado conjunto de restrições.

2.3 Aprendizagem Não Supervisionada

Como mostrado por (DAMASCENO, 2005) aprendizagem não supervisionada é aquela que utiliza instâncias sem a determinação do atributo classe. Este tipo de aprendizado é utilizado geralmente para análise exploratória dos dados, utilizando técnicas de agrupamento ou regras de associação. Onde agrupamentos têm como objetivo relacionar instâncias com características em comuns. Lembrando que o agrupamento é uma técnica que utiliza o aprendizado não supervisionado, ou seja, não utiliza no processamento o atributo classe.

A partir da definição de uma métrica de similaridade, os dados são agrupados, dando a possibilidade de encontrar relações interessantes entre as instâncias. Assim, o cliente do conhecimento gerado pode aplicar uma determinada ação em um subconjunto de instâncias presente nos dados. A suíte Weka possui os algoritmos Cobweb e SimpleKMeans e EM para tarefas de agrupamento.

2.4 Técnicas de Classificação

Uma técnica de classificação é uma abordagem sistemática para construção de modelos de classificação a partir de im conjunto de dados de entrada. Exemplos incluem classificadores de árvores de decisão, classificadores baseados em regras, redes neurais, máquinas de vetores de suporte e classificadores Bayes simples. Cada técnica emprega um algoritmo de aprendizagem para identificar um modelo que seja mais apropriado para o relacionamento entre o conjunto de atributos e o rótulo da classe dos dados de entrada. O modelos gerado pelo algoritmo de aprendizagem deve se adaptar bem aos dados de entrada e prever corretamente os rótulos de classes de registros que ele nunca viu antes. Portanto, o objetivo chave do algoritmo de aprendizagem é construir modelos com boa capacidade de generalização (TAN; STEINBACH; KUMAR, 2009).

A Figura 2 mostra uma abordagem geral para resolver problemas de classificação. Primeiro, um conjunto de treinamento consistindo de registros cujos

rótulos sejam conhecidos e devem ser fornecidos. O conjunto de treinamento é usado para construir um modelo de classificação, que é subsequentemente aplicado ao conjunto de teste, que consiste de registros com rótulos de classes desconhecidas (TAN; STEINBACH; KUMAR, 2009).

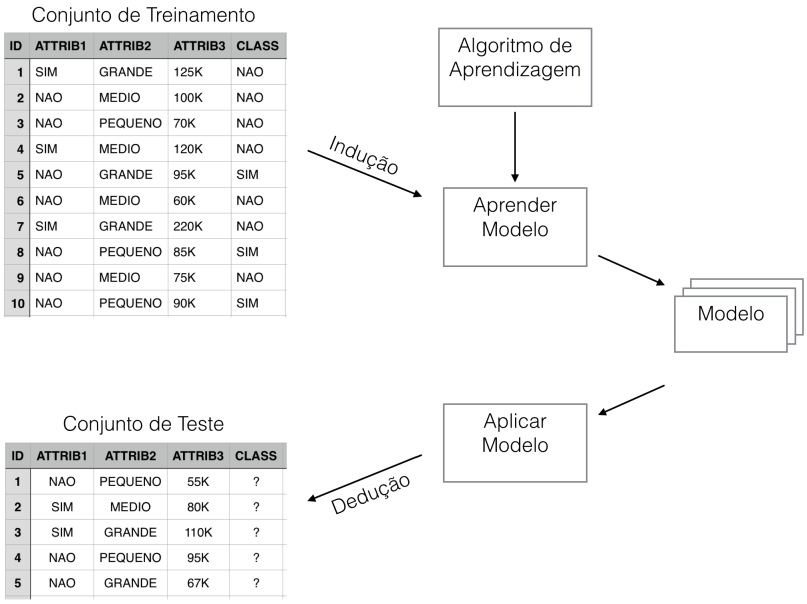

Figura 2 – Abordagem geral para a construção de um modelo de classificação. Adaptado de (TAN; STEINBACH; KUMAR, 2009)

2.4.1 Árvore de Decisão

Em uma árvore de decisão, cada nodo folha recebe um rótulo de classe. os nodos não terminais, que incluem o nodo raiz e outros nodos internos, contêm condições de testes de atributos para separar registros que possuem características diferentes.

Classificar um registro de testes é direto, assim que uma árvore de decisão tenha sido construída. Começando do nodo raiz, aplicamos a condição de teste ao registro e seguimos a ramificação apropriada baseados no resultado do teste. Isto nos levará a um outro nodo interno, para o qual uma nova condição de teste é aplicada, ou a um nodo folha (TAN; STEINBACH; KUMAR, 2009).

Utilizamos os algoritmos: J48, RandomTree e REPTree.

2.4.2 Bayes

Em muitas aplicações, o relacionamento entre o conjunto de atributos e a variável classe é não determinístico. O rótulo da classe de um registro de teste não pode ser previsto com certeza embora seu conjunto de atributos seja idêntico a alguns dos exemplos de treinamento. Esta situação pode surgir por causa de dados com ruídos oud a presença de determinados fatores de confusão que afetam a classificação mas que não são incluídos na análise.

O teorema de Bayes pode ser usado para resolver o problema de previsão.....

Utilizamos o algoritmo NaiveBayes

2.4.3 Máquinas de Vetores de Suporte

Utilizamos o algoritmo SMO.

2.5 WEKA

3 Base de Dados

A base de dados utilizada foi a do Município de Fortaleza

Neste capítulo, avaliamos os modelos gerados com os algoritmos que implementam a técnica de classificação através de um conjunto de experimentos.

- Metodologia de teste Cross-validation: Usa validação cruzada do tipo k-fold

3.1 Dados Abertos

Utilizamos a base de dados disponibilizada pelo portal de dados abertos da prefeitura do Rio de Janeiro de casos notificados de dengue.

3.2 SINAN

Nossos experimentos foram realizados a partir do conjunto de dados disponibilizado pela prefeitura do Rio de Janeiro do ano de 2015. Contendo inicialmente 26.568 instâncias e 97 atributos. Aplicamos os filtros: NumericToNominal -R 64, RemoveUseless -M 99.0. E observamos que a quantidade de atributos ficou igual a 66. Removemos os atributos: DT$_N ASC(ID12), NU_D DD_T EL(ID29), DDD_H OSP(ID64)$ $seaclassertuloAplicamosofiltroRemoveRange-R5001-last, pararealizarmosoexperimentocom5$

Aplicamos o filtro RemoveRange -R 10001-last, para realizarmos o experimento com 10.000 instâncias e o atributo CLASSI$_F IN com19\% dasinstnciasvazias.$

Aplicamos o filtro RemoveRange -R 15001-last, para realizarmos o experimento com 15.000 instâncias e o atributo CLASSI$_F IN com19\% dasinstnciasvazias.$

Tabela 1 – Descrição dos atributos do conjunto de dados

Atributos	Descrição
Data da investigação	Informar a data de investigação
Ocupação	Informar ramo da ocupação
Classificação	Informar a classificarão do caso

4 Metodologia

5 Resultados

5.1 Medida de Desempenho dos Algoritmos

As medidas de desepenho resultantes da classificação foram:

- Correctly Classified Instances
- Incorrectly Classified Instances
- TP Rate
- FP Rate
- Precision
- Recall
- F-Measure
- ROC Area

5.1.1 J48

5.1.2 RandomTree

5.1.3 REPTree

5.1.4 NaiveBayes

5.1.5 SMO

5.2 Resultados Comparativos Entre os Modelos

6 Conclusão

As considerações finais formam a parte final (fechamento) do texto, sendo dito de forma resumida (1) o que foi desenvolvido no presente trabalho e quais os resultados do mesmo, (2) o que se pôde concluir após o desenvolvimento bem como as principais contribuições do trabalho, e (3) perspectivas para o desenvolvimento de trabalhos futuros. O texto referente às considerações finais do autor deve salientar a extensão e os resultados da contribuição do trabalho e os argumentos utilizados estar baseados em dados comprovados e fundamentados nos resultados e na discussão do texto, contendo deduções lógicas correspondentes aos objetivos do trabalho, propostos inicialmente.

Referências

ADRIAANS, P.; ZANTINGE, D. Data mining, 1996. *Addision-Wesley, Harlow*. páginas 14

BELLAZZI, R.; ZUPAN, B. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, Elsevier, v. 77, n. 2, p. 81–97, 2008. páginas 8

BERRY, M. J.; LINOFF, G. *Data mining techniques: for marketing, sales, and customer support*. [S.l.]: John Wiley & Sons, Inc., 1997. páginas 14

BEZERRA, S. M. Prontuário eletrônico do paciente: uma ferramenta para aprimorar a qualidade dos serviços de saúde. *Revista Meta: Avaliação*, v. 1, n. 1, p. 73–82, 2009. páginas 7

COSTA, A. F. *Mineração de imagens médicas utilizando características de forma*. Tese (Doutorado) — Universidade de São Paulo, 2012. páginas 7, 8, 9

DAMASCENO, M. Introdução a mineração de dados utilizando o weka. *Instituto AAAFederal de Educação, Ciência e Tecnologia do Rio Grande do Norte*, 2005. páginas 14, 15

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996. páginas 2, 7, 13, 14

FREITAS, A. Uma introdução a data mining. *Informática Brasileira em Análise, Centro de Estudos e Sistemas Avançados do Recife (CESAR), Recife, Pe, ano II*, n. 32, 2000. páginas 13

HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009. páginas 11

JUNIOR, A. R. de M.; PIEDADE, M. P. Data mining na descoberta de padrões de sintomas com foco no auxílio ao diagnóstico médico. *Encontro Regional de Computação e Sistemas de Informação*, 2013. páginas 9

SANTOS, M. S.; NETO, J. C. da C. Amagodis: Algoritmos de mineração para apoio à gerência de ocorrências de dengue a partir de informações presentes na base dados do sinan. páginas 10

SHAKIL, K. A.; ANIS, S.; ALAM, M. Dengue disease prediction using weka data mining tool. *arXiv preprint arXiv:1502.05167*, 2015. páginas 10

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. [S.l.]: Ciencia Moderna, 2009. páginas 2, 16, 17

THITIPRAYOONWONGSE, D.; SURIYAPHOL, P.; SOONTHORNPHISAJ, N. Data mining of dengue infection using decision tree. *Entropy*, v. 2, p. 2, 2012. páginas 10

THOMÉ, A. C. G. Redes neurais: uma ferramenta para kdd e data mining. *Material Didático http://equipe. nce. ufrj. br/thome/grad/nn/mat_didatico/apostila_kdd_mbi. pdf, Outubro*, 2002. páginas 13

ANEXO A – Dicionário de dados - SINAN online

Este anexo descreve todos os atrubutos presentes no SINAN