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

Uso de técnicas de *Data Mining* na classificação de tipos de dengue

Jeferson Barros Alves

Manaus - AM
2015

Jeferson Barros Alves

Uso de técnicas de *Data Mining* na classificação de tipos de dengue

Monografia apresentada ao Curso de Graduação em Ciência da Computação da Faculdade Fucapi (Instituto de Ensino Superior Fucapi), 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

2015

Jeferson Barros Alves

Uso de técnicas de *Data Mining* na classificação de tipos de dengue

Monografia apresentada ao Curso de Graduação em Ciência da Computação da Faculdade Fucapi (Instituto de Ensino Superior Fucapi), como requisito parcial para a obtenção do Título de Bacharel em Ciência da Computação.
Aprovada em: / / , por:

Prof. Márcio Palheta Piedade, M.Sc.
Faculdade Fucapi
Orientador

Prof. Sergio Cleger Tamayo, Dr.
Faculdade Fucapi
Examinador

Profa. Marcela Sávia Picanço Pessoa, M.Sc.
Faculdade Fucapi
Examinadora

Manaus - AM

2015

Agradecimentos

Agradecimentos dirigidos àqueles que contribuíram de maneira relevante à elaboração do trabalho, sejam eles pessoas ou mesmo organizações.

Resumo

A tarefa de classificação em *Data Mining* consiste na predição de classe dado um determinado conjunto de atributos de uma instância. Esta técnica pode ser útil no diagnóstico médico, como ferramenta de apoio ao profissional da saúde. Na literatura, as melhores técnicas para classificação utilizam máquinas de vetores de suporte e árvore de decisão. Tais técnicas utilizam o conceito de hiperplano para separar de uma melhor forma as classes existentes em conjuntos de dados, também pode-se verificar o uso da divisão e conquista na identificação dos melhores atributos, sendo capazes de produzir resultados mais eficientes no processo de predição. Como resultado, obtivemos modelos classificadores baseados em SMO e J48, que alcançaram resultados de 89,61% e 86,63% na classificação correta de instâncias, respectivamente nos conjuntos de dados utilizados neste trabalho, mostrando resultados satisfatórios em relação ao classificador ZeroR utilizando como *baseline*. Foi possível observar também que os modelos gerados obtiveram melhores resultados à medida que os experimentos no tamanho do conjunto de dados aumentava.

Palavras-chave: Mineração de Dados, Classificação, Árvore de Decisão, Bayes, Máquina de Vetor de Suporte.

Abstract

The classification task in textit Data Mining is the class prediction given a certain set of attributes for an instance. This technique can be useful in medical diagnosis, as a support tool for the health professional. In the literature, the best techniques for classification using support vector machines and decision tree. Such techniques use the concept of hyperplane to separate in a better way the existing classes in data sets, you can also verify the use of divide and conquer to identify the best attributes, being able to produce more efficient results in the prediction process. As a result, we obtained classifiers based models SMO and J48, reaching results of 89.61 % and 86.63 % in the correct classification of instances respectively in the data sets used in this study, showing satisfactory results in relation to the classifier Zeror using as textit baseline. It was also observed that the generated models performed better as the experiments in the data set size increased.

Keywords: Data mining, Classifier, Decision Tree, Bayes, Support Vector Machine.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

KDD	*Knowledge Discovery in Databases*
WEKA	*Waikato Environment for Knowledge Analysis*
SVM	*Support Vector Machine*
SMO	*Sequential Minimal Optimization*
SINAN	Sistema de Informação de Agravos de Notificação
SUS	Sistema Único de Saúde

Sumário

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, surgiram diversas abordagens, técnicas e ferramentas que buscam transformar dados, também conhecida por Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases - KDD*). Tal abordagem foi proposta em 1989, com o objetivo de analisar os dados de uma base afim de extrair conhecimento útil (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A computação tem apoiado o desenvolvimento da medicina em diversas áreas como em sistemas de apoio a coleta de dados clínicos e exames por imagens, na organização das informações obtidas, entre outras (COSTA, 2012). Desta maneira, esse grande volume de dados é uma fonte valiosa de conhecimentos que pode ser utilizada como auxílio ao diagnóstico médico. Com base nisso, é necessário o desenvolvimento de técnicas que permitam a descoberta de conhecimentos 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

Ao longo dos últimos anos a mineração de dados tem sido cada vez mais utilizada na literatura médica. No entanto, a sua aplicação na análise 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 (BELLAZZI; ZUPAN, 2008). Faz-se importante o desenvolvimento de técnicas que permitam a descoberta de conhecimento em bases de dados médicos, que possam apoiar o profissional em sua tarefa diária de tomada de decisões, aumentando a precisão, a confiabilidade e eficiência dos diagnósticos elaborados pelo especialista (COSTA, 2012).

Segundo (SAÚDE, 2013), a dengue é uma das doenças com maior incidência no Brasil, atingindo a população de todos os estados, independente de classe social. Diante disto, faz-se necessário que um conjunto de ações com intuito de combater os casos de dengue sejam realizadas. Dentre elas temos a previsão de forma antecipada de possíveis novos casos. A classificação epidemiológica dos casos de dengue, que é feita habitualmente após desfecho clínico, na maioria das vezes é retrospectiva, dependendo de informações clínicas e laboratoriais, disponíveis no final do acompanhamento médico. Esses critérios não permitem o reconhecimento precoce de formas potencialmente graves, para as quais é crucial para a instituição de tratamento imediato (SAÚDE, 2013).

1.2 Objetivos

1.2.1 Objetivo Geral

Construir modelos utilizando as técnicas de classificação em *Data Mining* conhecidas como Árvore de Decisão, *Bayes* e *Support Vector Machine*. Após a criação dos modelos de classificação, avaliar o que obteve melhor resultado em relação ao conjunto de dados utilizados no SINAN (Sistema de Informação de Agravos de Notificação) Dengue do município de Fortaleza.

1.2.2 Objetivos Específicos

- Dividir o conjunto de dados na seguinte quantidade de instâncias: 5.000, 10.000, 15.000, 20.000 e 25.000;
- Utilizar atributos mais relevantes na tarefa de Classificação;
- Construir modelos clasificadores com os algoritmos de classificação: J48, REPTree, RandomTree, NaiveBayes e SMO;
- Identificar quais modelos gerados possuem melhor desempenho;
- Validar e interpretar os resultados obtidos no processo de classificação.

1.3 Justificativa

Transmitida pelo mosquito *Aedes aegypti*, a dengue é uma doença viral que se espalha rapidamente pelo mundo. Segundo a Organização Mundial da Saúde (OMS) por volta de 100 milhões de pessoas são infectadas por dengue anualmente, em mais de 100 países, sendo que 2,5 bilhões de pessoas estão sob o risco de infecção, tornando-se assim um dos maiores problemas de saúde pública no mundo (BHATT et al., 2013). A dengue é, hoje, uma das doenças mais frequentes no Brasil, atingindo a população em todos os estados, independente de classe social (SAÚDE, 2008).

Na região das Américas, a doença tem se disseminado com surtos cíclicos ocorrendo a cada 3/5 anos. No Brasil, a transmissão vem ocorrendo de forma continuada desde 1986, intercalando-se com a ocorrência de epidemias, geralmente associadas com a introdução de novos sorotipos em áreas anteriormente não se tinha registro ou alteração do sorotipo predominante. O maior surto no Brasil ocorreu em 2013, com aproximadamente 2 milhões de casos notificados. Atualmente, circulam no país os quatro sorotipos da doença (SAÚDE, 2015).

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. Os experimentos mostraram 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.
- Identificação dos algoritmos com melhor desempenho para previsão dos diagnósticos.

No trabalho de (THITIPRAYOONWONGSE; SURIYAPHOL; SOONTHORNPHISAJ, 2012), foi utilizada a técnica de classificação conhecida como Árvore de Decisão (TAN; STEINBACH; KUMAR, 2009), aplicado a um conjunto de dados temporais contendo dados clínicos e laboratoriais com mais de 400 atributos. Nos experimentos, os dados foram divididos em: conjunto de teste e treinamento, de modo que, o seu objetivo principal 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, 2011), foi desenvolvido um aplicativo para interação dos profissionais da saúde com os dados do SINAN (Sistema de Informação de Agravos de Notificação). Tal aplicativo, escrito em *Java*(ORACLE, 2015), implementa algoritmos de classificação e regras de associação da base de dados do SINAN, utilizando o *software* WEKA (*Waikato Environment for Knowledge Analysis*) (WAIKATO, 2015). Como resultado, esta ferramenta 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*) definida em (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), a realização desta pesquisa será composta das seguintes etapas:

- A primeira etapa deste trabalho baseado na revisão bibliográfica da literatura relacionada à mineração de dados, para determinar trabalhos relacionados, onde são aplicadas as técnicas de mineração em base de dados médicos para descoberta de conhecimento novo e relevante, que possa auxiliar em diagnósticos do médico especialista;

- Na segunda etapa, ocorrerá a extração, manutenção e pré-processamento dos dados, onde será realizada a análise de base de dados para extração de informações de atendimentos de pacientes;

- Na terceira etapa, será executado algoritmos para descoberta e extração de informações. O objetivo desta etapa será encontrar padrões na base de dados que foi pré-processada, utilizando a ferramenta de mineração WEKA(*Waikato Environment for Knowledge Analysis*), um software *open-source* (INITIATIVE, 2015) amplamente utilizado por implementar diversos algoritmos de mineração de dados (HALL et al., 2009).

- Por fim, na quarta 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, serão descritos conceitos básicos relacionados

à compreensão deste trabalho, bem como o referencial teórico. No Capítulo 3, serão apresentados os experimentos juntamente com os resultados obtidos. Finalmente no Capítulo 4, serão apresentadas as conclusões observadas e sugestões de trabalhos futuros da pesquisa.

2 Referencial Teórico

Este Capítulo apresenta uma visão geral dos conceitos envolvidos nas várias técnicas de classificação utilizadas ao longo deste trabalho.

Neste capítulo, apresentamos 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 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 mapea-

mento dos grupos ou conjuntos de informações que serão utilizados.

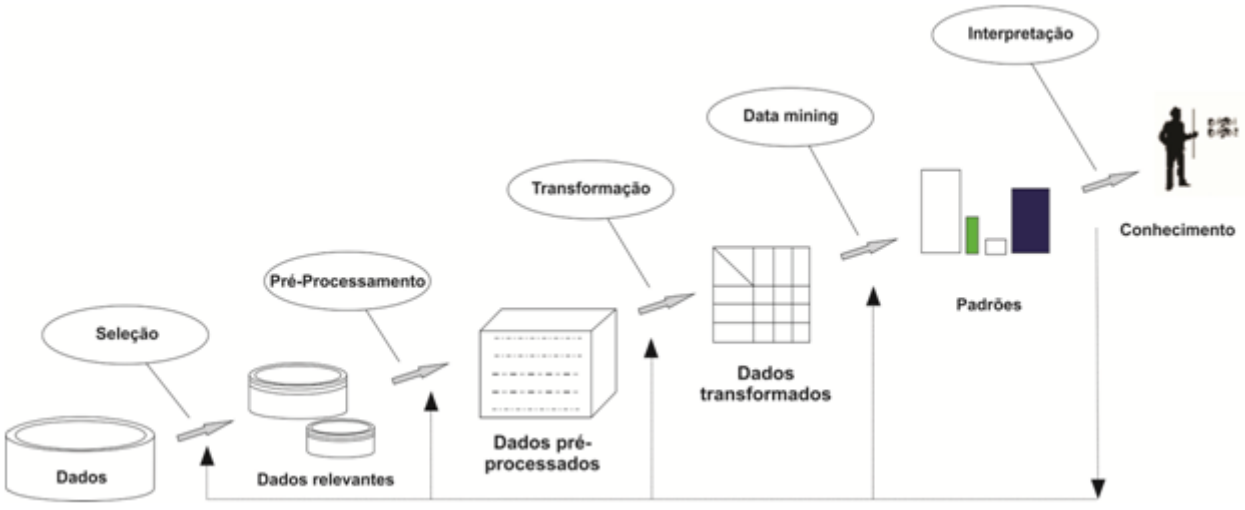
O pré-processamento é responsável pelo tratamento de ruídos e dados incompletos.

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.



Fonte: 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, 1996), 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, 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 Supervisionada (REVISAR)

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.

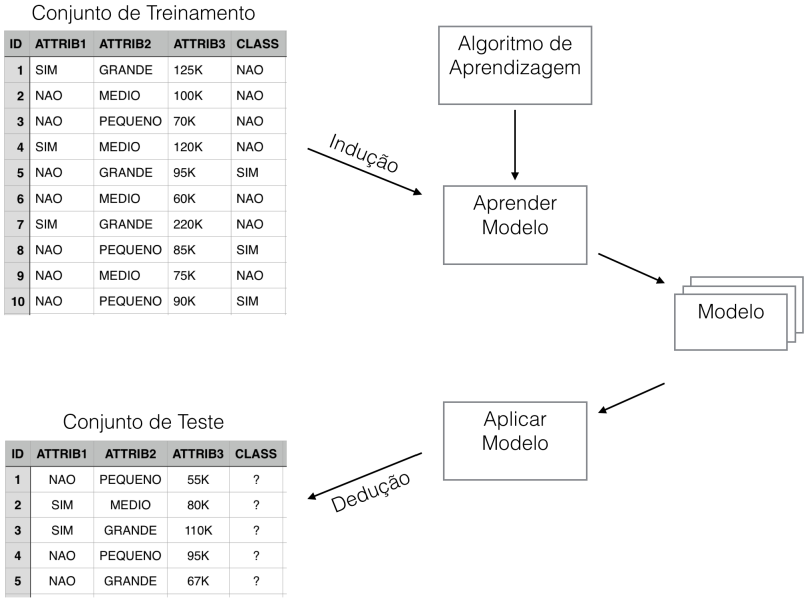
A partir da definição de uma métrica de similaridade, os dados são agrupados, dando a possibilidade de encontrar relações interessantes entre as instâncias. Assim, o cliente do conhecimento gerado pode aplicar uma determinada ação em um subconjunto de instâncias presente nos dados.

2.4 Técnicas de Classificação

Uma técnica de classificação é uma abordagem sistemática para construção de modelos de classificação a partir de um 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



Fonte: 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).

2.4.1.1 Algoritmo C4.5

O algoritmo J48 (WAIKATO, 2015) é uma implementação em Java (ORACLE, 2015) do algoritmo C4.5 (SALZBERG, 1994), que visa gerar árvore de decisão com tratamento de atributos contínuos e discretos, construindo uma árvore com um número de partições variável e com folhas sendo indicadas pelos valores do atributo categórico.

Segundo (HALMENSCHLAGER, 2002), para evitar a geração de todas as árvores possíveis, o algoritmo C4.5 se baseia no atributo mais informativo, escolhido entre todos os atributos ainda não considerados no caminho desde a raiz. O algoritmo seleciona como sendo o atributo mais informativo aquele que possuir o maior ganho de informação, resultante da diferença do valor da informação do atributo categórico e do valor da informação (entropia) do atributo em questão.

Para cada atributo é calculado seu ganho de informação. O atributo que tiver o maior valor, será considerado pelo algoritmo como o próximo nodo da árvore. Assim, a partição começa pelo nodo raiz e continua pelos nodos filhos da mesma maneira, até que todos os exemplos desta partição possuam a mesma classe, rotulando-se este nodo como folha e recebendo sua respectiva classe. Na Figura 3, podemos observar o funcionamento principal do C4.5.

Figura 3 – Núcleo do Algoritmo C4.5

```
Função C4.5   (R: conjunto de atributos não categóricos,
                    C: atributo categórico,
                 S: conjunto de treino): árvore de decisão
Início
        Se S está vazio
                retorna um nodo com valor negativo (falha);
                Se S contém exemplos de mesma classe
                retorna uma folha com o valor do atributo categórico;
        Se R está vazio
                retorna uma folha com o valor mais freqüente do atributo categórico;
                Senão
                calcula o atributo D com maior ganho de informação,  entre os
                                atributos restantes de R;
                        acrescenta nodo D;
                divide S em subconjuntos de acordo com os valores de D;
                        chama novamente a Função C4.5 para cada um desses valores;
Fim.
```

Fonte: (FELDENS; CASTILHO, 1997).

O algoritmo recebe o atributo categorico *C* (que indica a classe), um conjunto de atributos não categóricos *R* (demais atributos), e um conjunto de treinamento *S* (que contém os exemplos). Ele verifica qual é o atributo *D* mais informativo de *R*, ou seja, o atributo com maior ganho de informação para o conjunto de treinamento *S* e, então, subdivide este conjunto *S* de acordo com cada um dos valores deste atributo mais informativo *D* (FELDENS; CASTILHO, 1997).

De forma recursiva o algoritmo é chamado para cada subconjunto obtido, até atingir os seguintes objetivos:

- Não exista exemplos para classificar o conjunto de treinamento *S*. Neste caso, a árvore de decisão é um nodo com valor negativo;
- Que os exemplos no conjunto de treinamento *S* pertençam todos à mesma classe *C*. A árvore de decisão é um folha identificando a classe C_i;
- Não exista atributos não categóricos *R* para classificação. A árvore é uma folha identificando a classe C_i mais frequente de *S*.

Quando o conjunto de treinamentos *S* contém casos pertencentes a várias classes, a ideia é refinar *S* em subconjuntos de casos, que tendem a permanecer em apenas uma classe. A árvore de decisão para *S* consiste de um nodo de decisão que identifica o teste e uma ligação de cada valor possível do atributo. O mesmo processo de construção é aplicado recursivamente em cada subconjunto de casos de treinamento até que a *i-ésima* ligação tenha a árvore de decisão construída do subconjunto S_i de casos de treinamento. Quando o algoritmo parar, basta percorrer os caminhos desde a raiz até as folhas para verificar as descobertas extraídas da base de dados.

Como observado em (HALMENSCHLAGER, 2002), os critérios que fazem parte do algoritmo C4.5 são:

- Eleição do melhor atributo: utiliza o critério do ganho de informação;

- Tratamento dos atributos discretos: atribui uma ligação distinta a cada valor ou forma de agrupamentos de valores em vários conjuntos;
- Tratamento de atributos contínuos: Utiliza a técnica do teste simples para partição, escolhendo como ponto de cisão o ponto médio entre os valores;
- Tratamento de valores desconhecidos: Desconsidera os atributos com valores deconhecidos, utilizando apenas aqueles com valores totalmente deconhecidos;
- Determinação da classe associada à folha: É efetuada por atribuição da classe mais provável nesta folha;
- Método de poda: Utiliza a técnica de pós-poda baseada no erro, examinando a árvore de forma *bottom-up* e subtituindo uma subárvore por uma folha;
- Complexidade do algoritmo: É dada por *O(*mn^2*)*, em que *m* é o número de atributos e *n* é o número de instâncias do conjunto de treinamento.

2.4.1.2 Algoritmo *RandomTree*

Segundo (OSHIRO, 2013), considerando um conjunto de treinamento *T* com *a* atributos e *n* exemplos, seja T_k uma amostra *bootstrap* do conjunto de treinamento a partir de *T* com reposição, contendo *n* exemplos e usando*m* atributos aleatórios *(*$m \leq a$*)* em cada nó das árvores.

RandomTree é uma árvore induzida aleatoriamente a partir de um conjunto de árvores possíves, usando *m* atributos aleatórios em cada nó. O termo "aleatoriamente" significa que cada árvore tem uma chance igual de ser utilizada na amostrada. *Random Trees* podem ser geradas eficientemente e a combinação de grandes conjuntos de *Random Trees* geralmente leva a modelos precisos (ZHAO; ZHANG, 2008).

2.4.1.3 Algoritmo *REPTree*

O algoritmo *REPTree* constrói árvores de decisão para classificação ou regressão com base no ganho de informação/variância e poda esta árvore usando uma poda guiada por erro. Otimizado para velocidade, só classifica valores para atributos numéricos uma vez. Os valores são tratadas dividindo as instâncias correspondentes em pedaços (como no algoritmo C4.5) (WITTEN; FRANK; HALL, 2011).

2.4.2 Teorema de Bayes

É um teorema usado para calcular a probabilidade condicional , usando em estatística, probabilidade e outras aplicações

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 ou da presença de determinados fatores de confusão que afetam a classificação mas que não são incluídos na análise.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \tag{2.1}$$

Sendo $P(A \mid B)$ a probabilidade **posteriori** condicional de A a B, tem-se que:

- $P(B \mid A)$ a probabilidade **posteriori** condicional de B a A.
- $P(A)$ probabilidade **apriori** de A.
- $P(B)$ probabilidade **apriori** de B.

2.4.2.1 Algoritmo NaiveBayes

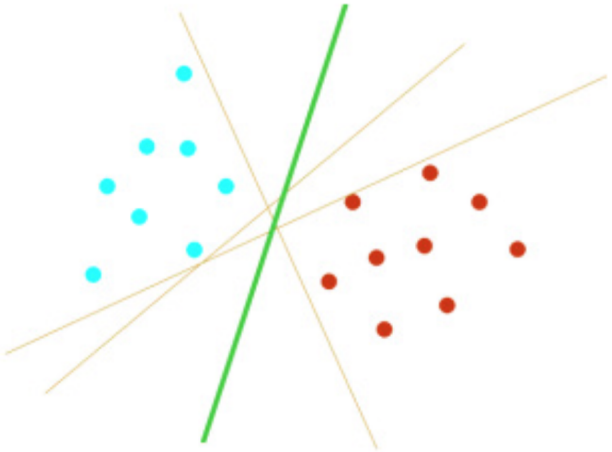
O Classificador *NaiveBayes* é uma técnica probabilística baseada no teorema de *Bayes*, expressão, para calcular a probabilidade Posteriori da classe C Em (JOHN; LANGLEY, 1995)

2.4.3 Máquinas de Vetores de Suporte (SVM)

Uma técnica de classificação que tem recebido considerável atenção pois esta técnica possui seus fundamentos na teoria de aprendizagem estatística e tem mostrado resultados empíricos promissores em muitas aplicações práticas, desde o reconhecimento de dígitos escritos à mão até a categorização de textos. SVM também funciona bem com dados de alta dimensionalidade e evita o problema da dimensionalidade. Outro aspecto único desta abordagem é que ela representa o limite da decisão usando um subconjunto dos exemplos de treinamento, conhecido com vetores de suporte (TAN; STEINBACH; KUMAR, 2009).

Na Figura 4, existe um conjunto de classificadores lineares que separam duas classes, mas apenas um (em destaque) que maximiza a margem de separação (distância da instância mais próxima ao hiperplano de separação das classes). O hiperplano com margem máxima é chamado de hiperplano ótimo (JUNIOR, 2010).

Figura 4 – Conjunto de hiperplanos possíveis



Fonte: (JUNIOR, 2010).

2.4.3.1 Algoritmo SMO

O Algoritmo SMO (*Sequential Minimal Optimization*) (PLATT, 1998), é um algoritmo simples que pode resolver rapidamente o problema quadrático encontrado em SVM (*Support Vector Machines*) sem qualquer armazenamento extra de matriz e sem o uso de passos de otimização numérica.

O SMO (*Sequential Minimal Optimization*) é um algoritmo de decomposição que usa uma solução analítica para otimizar um par de Multiplicadores de Lagrange por iteração. Em cada uma delas, três tarefas são executadas: seleção de um par de coeficientes, otimização do par de coeficientes selecionados e atualização de dados globais. A parada do algoritmo ocorre quando todos os coeficientes satisfazem o conjunto de condições. O conjunto de trabalho é formado pelo par de coeficientes a otimizar (HERNÁNDEZ, 2009).

2.4.4 Métricas para Avaliação de Modelo

A performance de um modelo de classificação pode ser medida utilizando a matriz de confusão, mostrada na Tabela 1. Nesta tabela pode ser observado que a distribuição entre as classes (proporção entre exemplos positivos e negativos) é o relacionamento entre a primeira e segunda linha. Assim, qualquer medida de desempenho que utilize valores de ambas as colunas será, necessariamente, sensível a desproporção entre as classes. Havendo mudança na distribuição das classes, os valores das métricas também mudarão, mesmo que o desempenho global do modelo não melhore (PRATI; BATISTA; MONARD, 2008).

Tabela 1 – Matriz de confusão para problemas de classes binárias

	Predição positiva	**Predição negativa**
Classe positiva	Verdadeiro positivo *(TP)*	Falso negativo *(FN)*
Classe negativa	Falso positivo *(FP)*	Verdadeito negativo *(TN)*

Fonte: Adaptado de (PRATI; BATISTA; MONARD, 2008).

- Precisão na classificação
- Taxa de erro na classificação

2.5 WEKA (*Waikato Environment for Knowledge Analysis*)

A ferramenta WEKA (Waikato Environment for Knowledge Analysis) conforme descrito em (HALL et al., 2009), visa proporcionar uma coleção abrangente de algoritmos de aprendizagem de máquina e ferramentas de pré-processamento de dados tanto para pesquisadores como para profissionais. Permite aos usuários experimentar rapidamente e comparar diferentes métodos de aprendizagem de máquina em diferentes conjuntos de dados. Sua arquitetura modular, extensível, permite que os processos de mineração de dados pode ser construido

a partir da vasta coleção de algoritmos de aprendizado e diversas ferramentas oferecidas.

2.6 Dados Abertos

Utilizamos a base de dados disponibilizada pelo portal de dados abertos da prefeitura do Município de Fortaleza de casos notificados de dengue. O portal de dados abertos de Fortaleza é um espaço desenvolvido pela Coordenadoria de Ciência, Tecnologia e Inovação da Prefeitura de Fortaleza (CITINOVA) para que a sociedade possa encontrar e utlizar os dados e informações públicas da cidade de Fortaleza. Os dados são publicados em formatos abertos que permitem sua reutilização em aplicativos digitais desenvolvidos por e para qualquer pessoa. Além disso, o portal serve como uma ferramenta de interlocução com a sociedade fortalezense para pensar e promover a inovação e a criatividade em prol da melhoria de serviços e da vida na cidade de Fortaleza. O portal de dados abertos de Fortaleza está em conformidade com os princípios da administração pública e observâncias às recomendações aceitas internacionalmente, como as emitidas pela Open Knowledge Foundation e pelo Consórcio W3C Internacional.

2.7 SINAN

O Sistema de Informações de Agravos de Notificação - SINAN é o principal sistema de informações que tem como objetivo os dados referentes a morbidade, sendo fundamental no processo de trabalho da Vigilância em Saúde, estando envolvido não somente nas ações de Vigilância Epidemiológica, mas também na Vigilância Ambiental em Saúde e Vigilância em Saúde do Trabalhador (CONASS, 2015). O SINAN foi desenvolvido no início da década de 90, com o objetivo de padronizar a coleta e o processamento de dados sobre agravos de notificação obrigatória em todo território nacional. Construido de maneira hierarquizada, mantendo coerência com a organização do SUS (Sistem Único

de Saúde), pretende ser suficiente ágil na viabilização de análises de situações de saúde em curto espaço de tempo. O SINAN fornece dados para a análise do perfil da morbidade e contribui para a tomada de decisões nos níveis municipal, estadual e federal (SAÚDE, 2008). A dengue é uma das doenças de notificação compulsória, devendo todo caso suspeito ou confirmado ser notificado ao Serviço de Vigilância Epidemiológica, por meio do SINAN nas fichas de notificação de investigação (SAÚDE, 2008).

3 Experimentos

Neste capítulo, avaliamos os modelos gerados através de uma série de experimentos. Apresentaremos a base de dados utilizada para os experimentos, a metodologia de experimentação e os experimentos em si realizados.

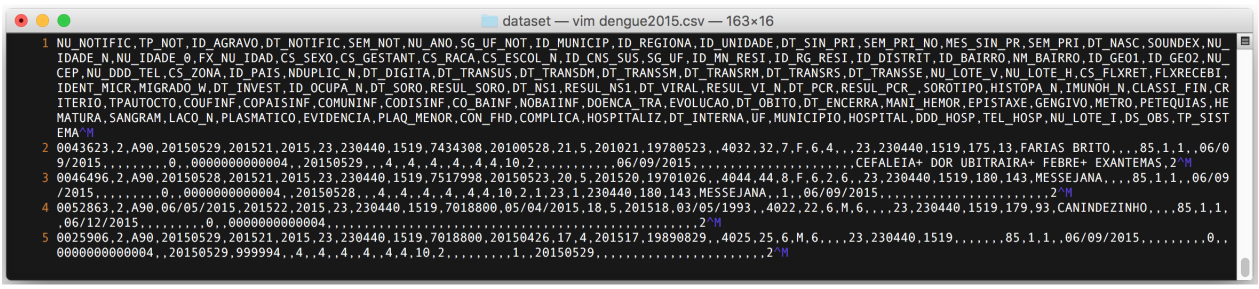
3.1 Base de Dados

Para avaliar o objetivo geral neste trabalho, usamos a base de dados do Município de Fortaleza, diposnibilizada pelo portal de dados abertos. Os dados são publicados em formatos abertos que permitem sua reutilização em aplicativos digitais desenvolvidos por e para qualquer pessoa (NOTIFICAÇÕES. . . , 2015). Este conjunto de dados contém notificações de dengue do Município de Fortaleza no período de janeiro à junho de 2015, contendo 26.568 instâncias.

3.2 Pré-processamento

Após o download da base de dados em (NOTIFICAÇÕES. . . , 2015), iniciamos a importação do arquivo no formato CSV (*Comma Separated Value*), conforme mostrado na Figura 5.

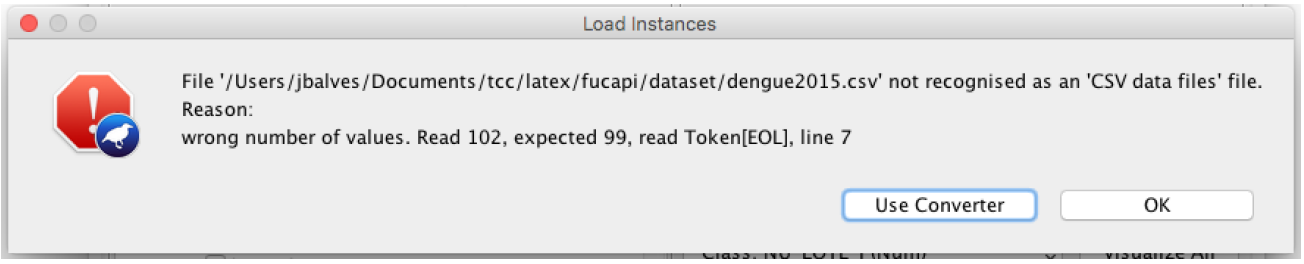
Figura 5 – Data Set Original




```
NU_NOTIFIC,TP_NOT,ID_AGRAVO,DT_NOTIFIC,SEM_NOT,NU_ANO,SG_UF_NOT,ID_MUNICIP,ID_REGIONA,ID_UNIDADE,DT_SIN_PRI,SEM_PRI_NO,MES_SIN_PR,SEM_PRI,DT_NASC,SOUNDEX,NU_
IDADE_N,NU_IDADE_0,FX_NU_IDAD,CS_SEXO,CS_GESTANT,CS_RACA,CS_ESCOL_N,ID_CNS_SUS,SG_UF,ID_MN_RESI,ID_RG_RESI,ID_DISTRIT,ID_BAIRRO,NM_BAIRRO,ID_GEO1,ID_GEO2,NU_
CEP,NU_DDD_TEL,CS_ZONA,ID_PAIS,NDUPLIC_N,DT_DIGITA,DT_TRANSUS,DT_TRANSDM,DT_TRANSSM,DT_TRANSRM,DT_TRANSRS,DT_TRANSSE,NU_LOTE_V,NU_LOTE_H,CS_FLXRET,FLXRECEBI,
IDENT_MICR,MIGRADO_W,DT_INVEST,ID_OCUPA_N,DT_SORO,RESUL_SORO,DT_NS1,RESUL_NS1,DT_VIRAL,RESUL_VI_N,DT_PCR,RESUL_PCR_,SOROTIPO,HISTOPA_N,IMUNOH_N,CLASSI_FIN,CR
ITERIO,TPAUTOCTO,COUFINF,COPAISINF,COMUNINF,CODISINF,CO_BAINF,NOBAIINF,DOENCA_TRA,EVOLUCAO,DT_OBITO,DT_ENCERRA,MANI_HEMOR,EPISTAXE,GENGIVO,METRO,PETEQUIAS,HE
MATURA,SANGRAM,LACO_N,PLASMATICO,EVIDENCIA,PLAQ_MENOR,CON_FHD,COMPLICA,HOSPITALIZ,DT_INTERNA,UF,MUNICIPIO,HOSPITAL,DDD_HOSP,TEL_HOSP,NU_LOTE_I,DS_OBS,TP_SIST
EMA^M
0043623,2,A90,20150529,201521,2015,23,230440,1519,7434308,20100528,21,5,201021,19780523,,4032,32,7,F,6,4,,,23,230440,1519,175,13,FARIAS BRITO,,,,85,1,1,,06/0
9/2015,,,,,,,,,0,,0000000000004,,20150529,,,4,,4,,4,,4,,4,4,10,2,,,,,,,,,,06/09/2015,,,,,,,,,,,,,,,,,,,,,,CEFALEIA+ DOR UBITRAIRA+ FEBRE+ EXANTEMAS,2^M
0046496,2,A90,20150528,201521,2015,23,230440,1519,7517998,20150523,20,5,201520,19701026,,4044,44,8,F,6,2,6,,23,230440,1519,180,143,MESSEJANA,,,,85,1,1,,06/09
/2015,,,,,,,,,0,,0000000000004,,20150528,,,4,,4,,4,,4,,4,4,10,2,1,23,1,230440,180,143,MESSEJANA,,1,,06/09/2015,,,,,,,,,,,,,,,,,,,,,,2^M
0052863,2,A90,06/05/2015,201522,2015,23,230440,1519,7018800,05/04/2015,18,5,201518,03/05/1993,,4022,22,6,M,6,,,,23,230440,1519,179,93,CANINDEZINHO,,,,85,1,1,
,06/12/2015,,,,,,,,,0,,0000000000004,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,2^M
0025906,2,A90,20150529,201521,2015,23,230440,1519,7018800,20150426,17,4,201517,19890829,,4025,25,6,M,6,,,,23,230440,1519,,,,,,,85,1,1,,06/09/2015,,,,,,,,,0,,
0000000000004,,20150529,999994,,4,,4,,4,,4,,4,4,10,2,,,,,,,,,1,,20150529,,,,,,,,,,,,,,,,,,,,,,2^M
```

Fonte: Próprio Autor.

Figura 6 – Erro ao Carregar Instâncias



Fonte: Próprio Autor.

Onde nos deparamos com erros no campos DS_OBS, onde o WEKA informa que está lendo 102 atributos em vez de 99, conforme mostrado na Figura 6. Já na Figura 7, podemos observar que houve falha na inserção dos dados no atributo DS_OBS, pois o mesmo foi preenchido com vírgula em seu conteúdo da seguinte "CEFALEIA, FEBRE, DOR NOS OLHOS, VOMITO", além do caracter "M̂" que representa quebra de linha. Pelo fato do arquivo ser contruído no formato CSV (*Comma Separated Value*), essas vírgulas excedentes serão consideradas como início de novos atributos.

Figura 7 – Erro Mencionado pelo WEKA

```
0040967,2,A90,20150522,201520,2015,23,230440,1519,7434308,20150519,20,5,20
1520,05/06/1993,,4022,22,6,F,5,4,,,23,230440,1519,177,66,PRESIDENTE KENNED
Y,,,,85,1,1,,20150613,,,,,,,,0,,0000000000004,,20150522,,,4,,4,,4,,4,,4,4
,10,2,1,23,1,230440,177,66,PRESIDENTE KENNEDY,,1,,20150529,,,,,,,,,,,,,,
,,,,,,CEFALEIA,FEBRE,DOR NOS OLHOS,VOMITO,2^M
```

Fonte: Próprio Autor.

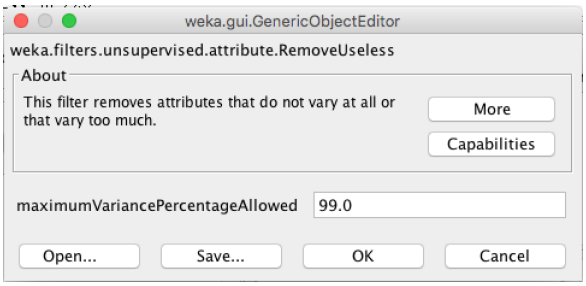
Para resolver este problema editamos o arquivo de tal forma que os atributos "DS_OBS" e "TP_SISTEMA" foram removidos.

Em seguida aplicamos os seguintes filtros:

- *RemoveUseless*: Este filtro remove atributos que não variam em tudo ou que variam muito. Todos os atributos constantes são excluídos automaticamente, juntamente com qualquer que exceder o percentual máximo de parâmetro de variação. O teste de variância máxima só é aplicada aos

atributos nominais. Utilizamos os parâmetros padrões conforme Figura 8.

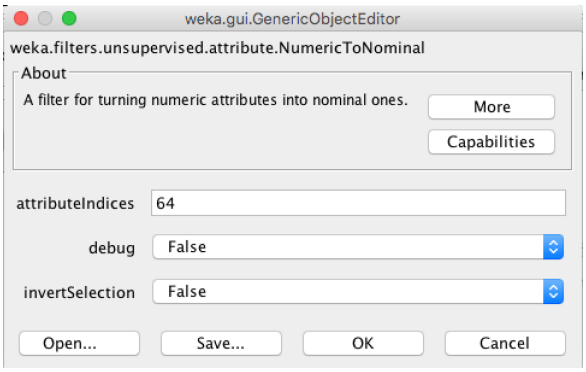
Figura 8 – Filtro *RemoveUseless*



Fonte: Próprio Autor.

- *NumericToNominal*: Um filtro para transformar numérico atributos em uns nominais. Ao contrário da discretização, apenas toma todos os valores numéricos e os adiciona à lista de valores nominais do que atributo. Útil após a importação CSV, para impor certos atributos para se tornar nominal. Utilizamos o parâmetro mostrado na Figura 9.

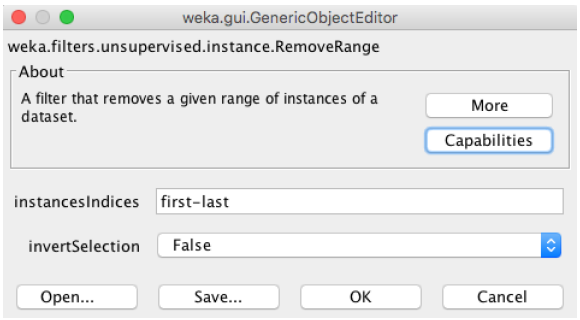
Figura 9 – Filtro *NumericToNominal*



Fonte: Próprio Autor.

- *RemoveRange*: Um filtro que remove um determinado intervalo de ocorrências de um conjunto de dados. Utilizamos o parâmetro mostrado na Figura 10. O parâmetro "instancesIndices" foi utilizado cinco vezes para criarmos 05 (cinco) conjuntos de dados com as seguintes quantidades de instâncias: 5.000, 10.000, 15.000, 20.000 e 25.000. Que foram utilizados durantes os experimentos.

Figura 10 – Filtro *RemoveRange*



Fonte: Próprio Autor.

3.3 Metodologia do Experimento

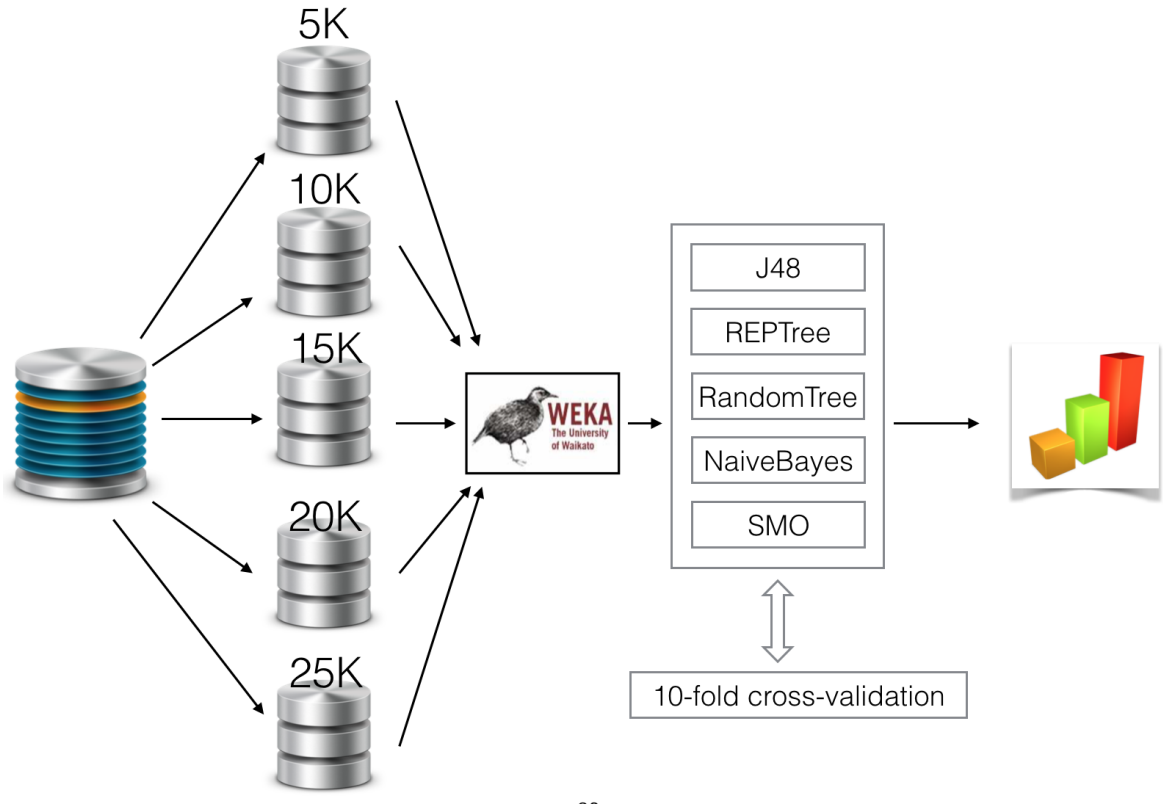
Nossos experimentos foram realizar a partir dos conjuntos de dados divididos nas seguintes quantidades de instâncias: 5.000, 10.000, 15.000, 20.000 e 25.000. A cada conjunto de instâncias, aplicamos os processos de classificação utilizando 5 classificadores que estão contidos em 3 técnicas de classificação. Como resultado dos classificadores obtivemos 25 modelos.

Todos os testes foram feitos utilizando o processo de validação conhecido como *Cross Validation* diposnibilizado pelo WEKA (*Waikato Environment for Knowledge Analysis*), com o funcionamento da seguinte forma:

- Defini-se o parâmetro *folds*, que utilizaremos com o valor 10;
- O conjunto de dados é aleatoriamente reordenado e depois divide-se em conjuntos de tamanho = *folds*;

- Em cada iteração, um conjunto é usado para testes e os outros *n -1* são utilizados para o treinamento do classificador;
- Os resultados do teste são coletados e calculados sobre todos os conjuntos.

Figura 11 – Metodologia Processo dos Experimentos



Fonte: Próprio Autor.

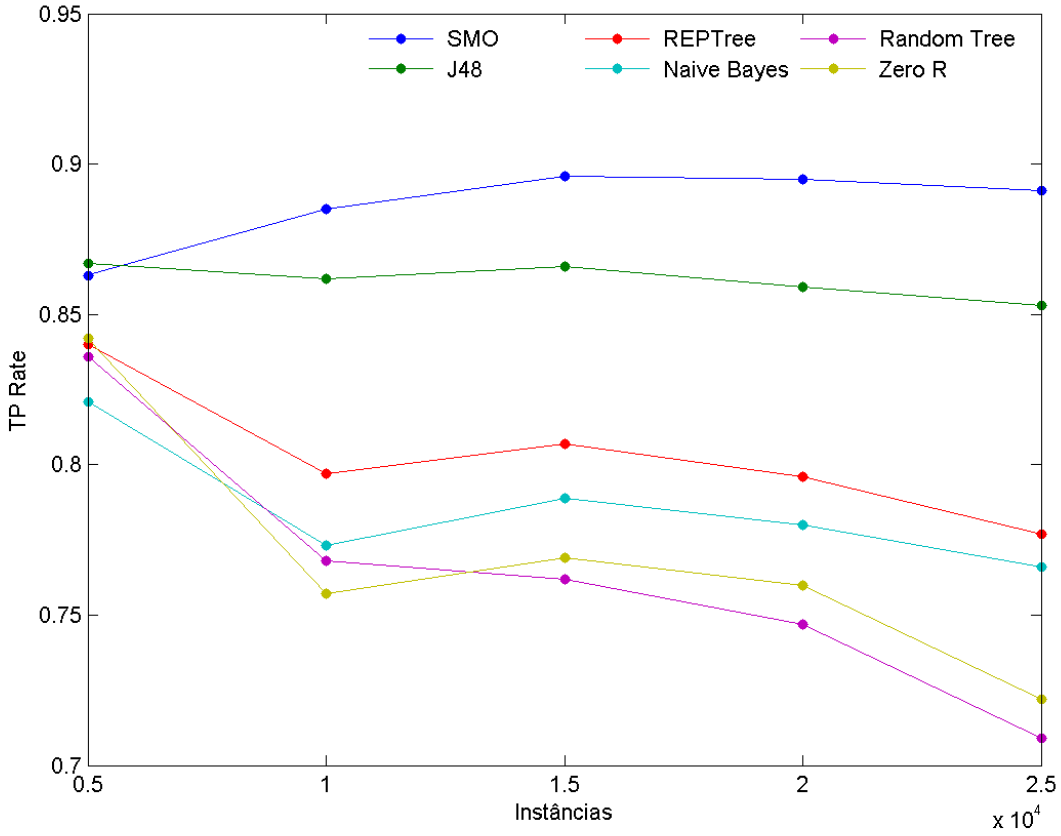
3.4 Resultados

Nesta seção apresentamos os resultados obtidos na aplicação dos classificadores em 5 conjuntos de dados e as medidas resultantes que utilizamos:

Tabela 2 – Resultados de Classificação

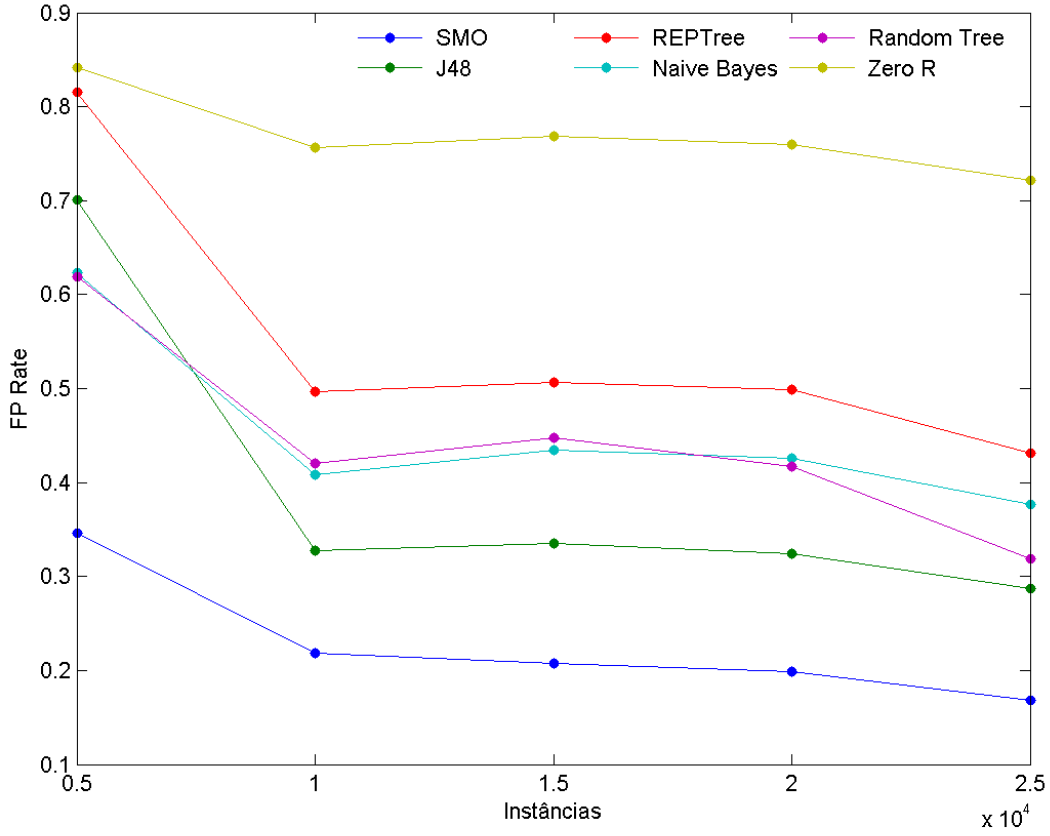
Classificador	**Correctly Classified Instances**				
	5k instâncias	10k instâncias	15k instâncias	20k instâncias	25k instâncias
SMO	89,2866%	88,5354%	89,6185%	89,4714%	89,1407%
J48	86,653%	86,2315%	86,6322%	85,9331%	85,3096%
REPTree	83,9683%	79,7414%	80,7076%	79,6344%	77,7052%
NaiveBayes	83,5592%	76,8103%	76,1682%	74,9666%	70,8689%
RandomTree	82,0762%	77,2906%	78,8551%	78,0289%	76,6155%
ZeroR	84,2496%	75,6773%	76,8608%	75,9664%	72,1555%

Figura 12 – TP Rate



Fonte: Próprio Autor.

Figura 13 – FP Rate



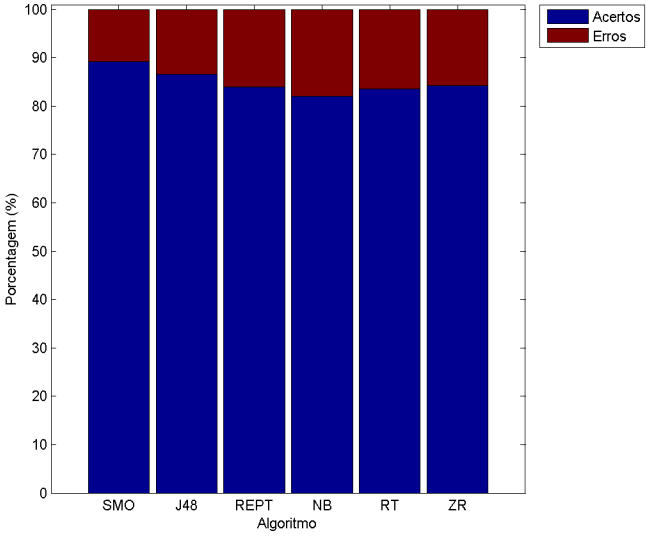
Fonte: Próprio Autor.

4 Conclusões e Trabalhos Futuros

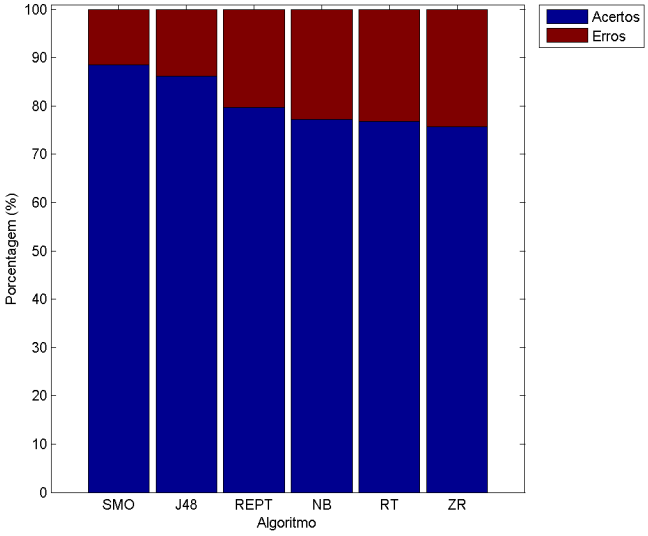
Neste Capítulo, apresentamos as conclusões do nosso trabalho, incluindo suas limitações de nossa pesquisa e futuros estudos que possam ser explorados.

4.1 Resultados Obtidos

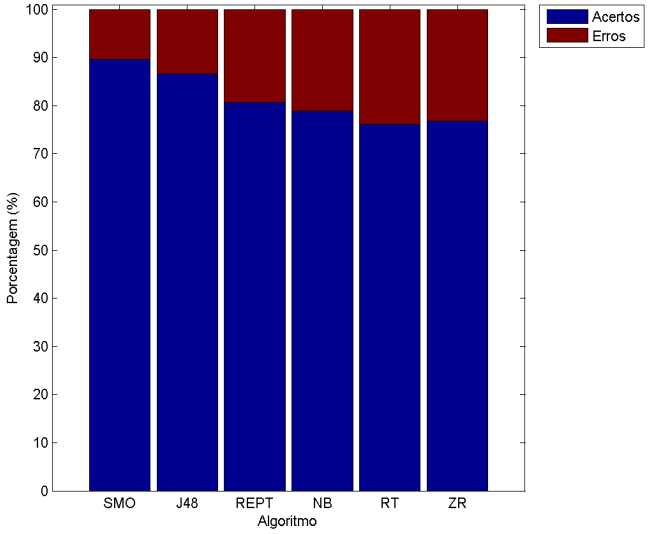
As medidas de desepenho resultantes dos experimentos na classificação foram:



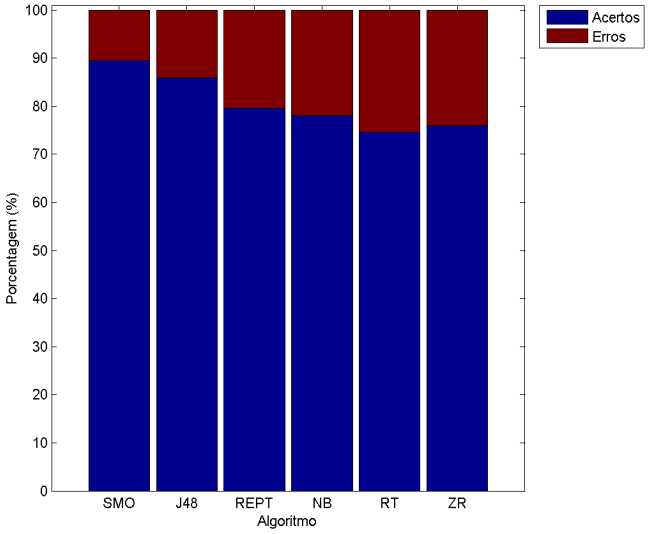
(a) Figura 1



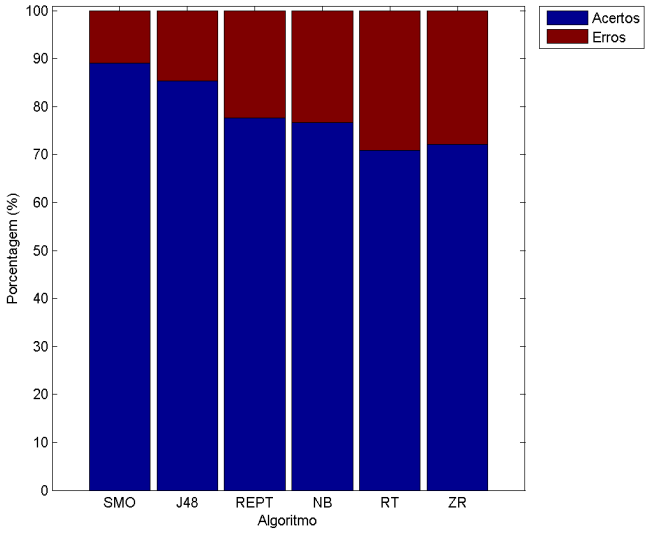
(b) Figura 2



(c) Figura 3

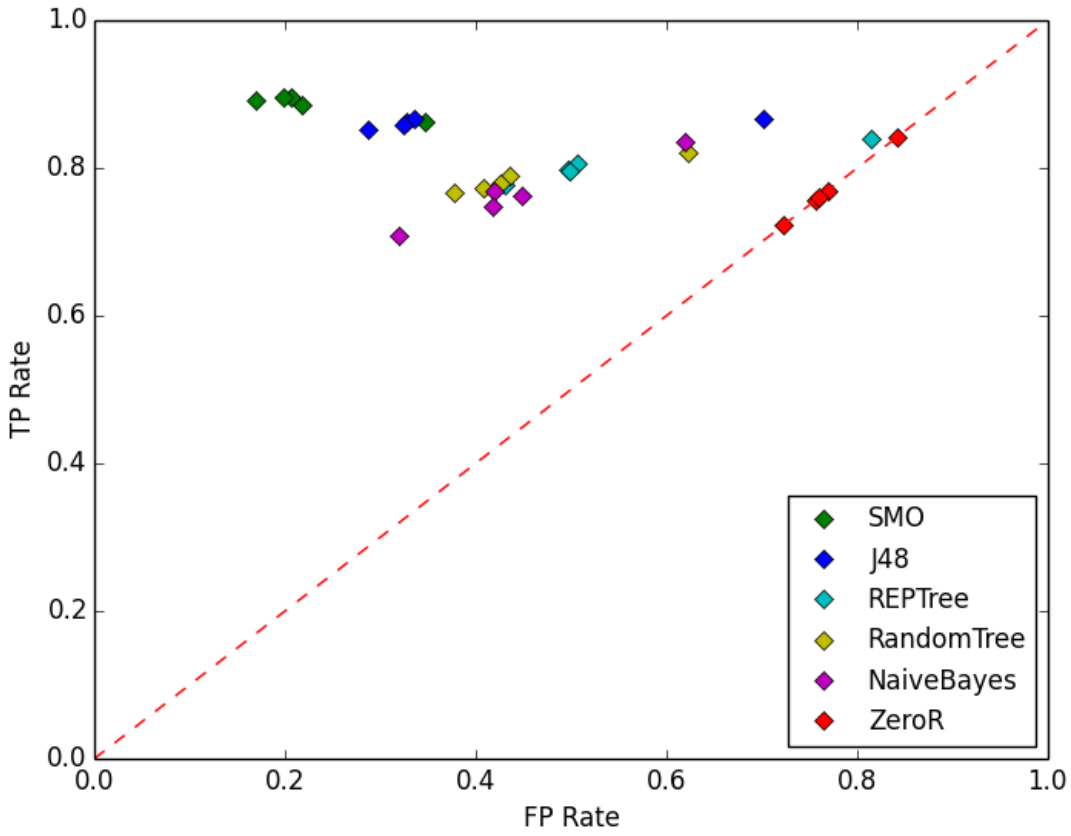


(d) Figura 4



(e) Figura 5

Figura 14 – Gráfico ROC dos Modelos Gerados



Fonte: Próprio Autor.

Como pode-se observar na Figura 14, os modelos gerados pelo algoritmos *SMO* e *J48* obtiveram melhor qualidade.

4.2 Limitações

Neste trabalho não estamos levando em consideração o tempo de criação de cada modelo gerado.

4.3 Trabalhos Futuros

Como trabalhos futuros, podem ser explorados na continuação desta pesquisa as oportunidades descritas a seguir:

- Implementar remoção de instâncias que influenciam na qualidade do classificador.
- Adaptar a construção de modelos em tempo real.
- Integrar informações de precipitação de chuva da região onde os dados foram coletados.
- Usar outros algoritmos de classificação.

Referências

ADRIAANS, P.; ZANTINGE, D. Data mining. *Addision-Wesley, Harlow*, 1996. páginas 18

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 12

BERRY, M. J.; LINOFF, G. *Data mining techniques: for marketing, sales, and customer support*. [S.l.]: John Wiley & Sons, Inc., 1997. páginas 18

BHATT, S. et al. The global distribution and burden of dengue. *Nature*, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., v. 496, n. 7446, p. 504–507, 04 2013. Disponível em: <http://dx.doi.org/10.1038/nature12060>. páginas 13

CONASS. *Nota Técnica - 45 - 2013*. 2015. Disponível em: <http://www.conass.org.br>. páginas 28

COSTA, A. F. *Mineração de imagens médicas utilizando características de forma*. Tese (Doutorado) — Universidade de São Paulo, 2012. páginas 11, 12

DAMASCENO, M. Introdução a mineração de dados utilizando o weka. *Instituto AAAFederal de Educação, Ciência e Tecnologia do Rio Grande do Norte*, 2005. páginas 18, 19

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 11, 15, 17, 18

FELDENS, M. A.; CASTILHO, J. M. V. d. Engenharia da descoberta de conhecimento em bases de dados: estudo e aplicação na área de saúde. *Porto Alegre*, 1997. páginas 22

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 17

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 15, 27

HALMENSCHLAGER, C. *Um algoritmo para indução de árvores e regras de decisão*. Tese (Doutorado) — UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, 2002. páginas 21, 23

HERNÁNDEZ, R. A. *MP-SMO: um algoritmo para a implementação VLSI do treinamento de máquinas de vetores de suporte.* Tese (Doutorado) — Universidade de São Paulo, 2009. páginas 26

INITIATIVE, O. S. 2015. Disponível em: <https://opensource.org/definition>. páginas 15

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: MORGAN KAUFMANN PUBLISHERS INC. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. [S.l.], 1995. p. 338–345. páginas 25

JUNIOR, G. M. d. O. Máquina de vetores suporte: estudo e análise de parâmetros para otimização de resultado, 2010. *Trabalho de Graduação em Ciência da Computação*, 2010. páginas 25, 26

NOTIFICAÇÕES de Dengue - Ref. 01 a 06/2015. 2015. <http://dados.fortaleza.ce.gov.br/catalogo/dataset/notificacoes-de-dengue-ref-01-06-2015>. Acessado em 2015. páginas 29

ORACLE. 2015. Disponível em: <https://www.java.com/pt_BR/download/faq/whatis_java.xml>. páginas 14, 21

OSHIRO, T. M. *Mestre em Bioinformática*. Tese (Doutorado) — Universidade de São Paulo, 2013. páginas 23

PLATT, J. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. [S.l.], 1998. 21 p. Disponível em: <http://research.microsoft.com/apps/pubs/default.aspx?id=69644>. páginas 26

PRATI, R. C.; BATISTA, G. E. d. A. P. A.; MONARD, M. C. Curvas roc para avaliação de classificadores. *Revista IEEE América Latina*, v. 6, n. 2, p. 215–222, 2008. páginas 26, 27

SALZBERG, S. C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, Kluwer Academic Publishers, v. 16, n. 3, p. 235–240, 1994. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1007/BF00993309>. páginas 21

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. 2011. páginas 14

SAÚDE, M. da. Vigilância em saúde: Dengue, esquistossomose, hanseníase, malária, tracoma e tuberculose. *Departamento de Atenção Básica*, 2008. páginas 13, 28

SAÚDE, M. da. Dengue: diagnóstico e manejo clínico: adulto e criança. *Departamento de Atenção Básica*, 2013. páginas 12

SAÚDE, M. da. 2015. Disponível em: <http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/svs/dengue>. páginas 14

SHAKIL, K. A.; ANIS, S.; ALAM, M. Dengue disease prediction using weka data mining tool. *arXiv preprint arXiv:1502.05167*, 2015. páginas 14

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. [S.l.]: Ciencia Moderna, 2009. páginas 14, 20, 21, 25

THITIPRAYOONWONGSE, D.; SURIYAPHOL, P.; SOONTHORNPHISAJ, N. Data mining of dengue infection using decision tree. *Entropy*, v. 2, p. 2, 2012. páginas 14

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 17

WAIKATO, T. U. of. 2015. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. páginas 15, 21

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123748569, 9780123748560. páginas 24

ZHAO, Y.; ZHANG, Y. Comparison of decision tree methods for finding active objects. *Advances in Space Research*, Elsevier, v. 41, n. 12, p. 1955–1959, 2008. páginas 24