Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Colegiado de Sistemas de Informação

Uso de Técnicas de Mineração de Dados para Classificação das Ocorrências de Casos de Dengue nos Municípios Brasileiros

Renato Avilez Vilarinho

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Janniele Aparecida Soares Araujo

Março, 2017
João Monlevade/MG

Renato Avilez Vilarinho

Uso de Técnicas de Mineração de Dados para Classificação das Ocorrências de Casos de Dengue nos Municípios Brasileiros

Orientador: Janniele Aparecida Soares Araujo

Coorientador: Samuel Souza Brito

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Computação e Sistemas da Universidade Federal de Ouro Preto como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação

Universidade Federal de Ouro Preto

João Monlevade

Março de 2017

V697u Vilarinho, Renato Avilez.
Uso de técnicas de mineração de dados para classificação das ocorrências de casos de dengue nos municípios brasileiros [manuscrito] / Renato Avilez Vilarinho. - 2017.

50f.: il.: color; tabs.

Orientador: Prof. Me. Janniele Aparecida S. Araujo.
Coorientador: Prof. Me. Samuel S. Brito.

Monografia (Graduação). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Departamento de Computação e Sistemas de Informação.

1. Mineração de dados (Computação). . 2. Classificação (dados). 3. Tratamento de dados. 4. Processamento de dados. 5. Algoritmos . I. Araujo, Janniele Aparecida S. . II. Brito, Samuel S.. III. Universidade Federal de Ouro Preto. IV. Titulo.

Catalogação: ficha@sisbin.ufop.br

CDU: 004.62:007.52

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E APLICADAS
COLEGIADO DO CURSO DE SISTEMAS DE INFORMAÇÃO

TERMO DE RESPONSABILIDADE

Eu, Renato Avilez Vilarinho, CPF: 001.239.101-88, declaro que o texto do trabalho de conclusão de curso intitulado *"Uso de Técnicas de Mineração de Dados para Classificação das Ocorrências de Casos de Dengue nos Municípios Brasileiros"* é de minha inteira responsabilidade e que não há utilização de texto, material fotográfico, código fonte de programa ou qualquer outro material pertencente a terceiros sem as devidas referências ou consentimento dos respectivos autores.

João Monlevade, 24 de março de 2017

Renato Avilez Vilarinho

Assinatura do aluno

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E APLICADAS
COLEGIADO DO CURSO DE SISTEMAS DE INFORMAÇÃO

ATA DE DEFESA

Aos 24 dias do mês de Março de 2017, às 18 horas e 30 minutos, na sala C 204 do Instituto de Ciências Exatas e Aplicadas, foi realizada a defesa de Monografia pelo aluno **Renato Avilez Vilarinho,** sendo a Comissão Examinadora constituída pelos professores: Prof. Ms. Janniele Aparecida Soares Araujo, Prof. Ms. Samuel Souza Brito e Prof. Ms. Helen de Cássia Souza da Costa Lima e Prof. Ms. Theo Silva Lins.

O candidato apresentou a monografia intitulada: *"Uso de Técnicas de Mineração de Dados para Classificação das Ocorrências de Casos de Dengue nos Municípios Brasileiros".* A comissão examinadora deliberou, por unanimidade, pela aprovação do candidato, com nota 9,5 (nove e meio), concedendo-lhe o prazo de 15 dias para incorporação das alterações sugeridas ao texto final.

Na forma regulamentar, foi lavrada a presente ata que é assinada pelos membros da Comissão Examinadora e pelo graduando.

João Monlevade, 24 de Março de 2017.

Janniele Aparecida Soares Araujo
Prof. Ms. Janniele Aparecida Soares Araujo,
Professor Orientador

Samuel Souza Brito
Prof. Ms. Samuel Souza Brito
Professor Coorientador

Helen de Cássia S. da Costa
Prof. Ms. Helen de Cássia Souza da Costa Lima
Professor Convidado

Theo Silva Lins
Prof. Ms. Theo Silva Lins
Professor Convidado

Renato Avilez Vilarinho
Renato Avilez Vilarinho
Graduando

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E APLICADAS
COLEGIADO DO CURSO DE SISTEMAS DE INFORMAÇÃO

Curso de Sistemas de Informação

FOLHA DE APROVAÇÃO DA BANCA EXAMINADORA

Uso de Técnicas de Mineração de Dados para Classificação das Ocorrências de Casos de Dengue nos Municípios Brasileiros

Renato Avilez Vilarinho

Monografia apresentada ao Departamento de Ciências Exatas e Aplicadas da Universidade Federal de Ouro Preto como requisito parcial da disciplina CSI499 – Trabalho de Conclusão de Curso II do curso de Bacharelado em Sistemas de Informação e aprovada pela Banca Examinadora abaixo assinada:

Prof. Ms. Janniele Aparecida Soares Araujo
DECSI – Universidade Federal de Ouro Preto

Prof. Ms. Samuel Souza Brito
DECSI – Universidade Federal de Ouro Preto

Prof. Ms. Helen de Cássia Sousa da Costa Lima
DECSI – Universidade Federal de Ouro Preto

Prof. Ms. Théo ~~da~~ Silva Lins
DECSI – Universidade Federal de Ouro Preto

João Monlevade, 24 de Março de 2017

Aos meus pais, Marco e Rosangela

"A vida é arte do encontro,
embora haja tanto desencontro pela vida"
Vinícius de Moraes

RESUMO

Neste trabalho foram exploradas as capacidades de Algoritmos de Mineração de Dados na obtenção de informações úteis relativas aos casos de Dengue nos municípios brasileiros. Características socioeconômicas e os casos notificados de Dengue foram utilizados como atributos para os registros. No primeiro momento foram utilizadas Tarefas de Agrupamento para identificar grupos de municípios a partir da quantidade de casos de Dengue notificados. Os grupos encontrados foram denominados: "Muito Baixo", "Baixo", "Médio", "Alto"e "Muito Alto". Em seguida estes dados foram submetidos a algoritmos de Classificação. As atividades de Classificação tiveram dificuldade de identificar os grupos "Alto"e "Muito Alto", porém apresentaram resultados satisfatórios de acordo com as métricas utilizadas.

Palavras-chaves: Dengue, Classificação, Agrupamento, Weka, Mineração de Dados, KDD, K-means, C4.5.

ABSTRACT

In this paper we explored the capabilities of Data Mining Algorithms in obtaining useful information regarding Dengue Fever cases in Brazilian cities. Socioeconomic characteristics and reported cases of Dengue Fever were used as attributes for the instances. At first moment, we used Clustering Techniques to identify clusters of cities based on the number of reported cases of Dengue Fever. The clusters found were called "Very Low", "Low", "Medium", "High" and "Very High". Then, these data were submitted to Classification algorithms. The Classification tasks didn't identify "High" and "Very High" clusters as expected, but still presented satisfactory results according to the used metrics.

Key-words: Dengue Fever, Clustering algorithms, Classification algorithms, Weka, Data Mining, KDD, K-means, J48.

AGRADECIMENTOS

Agradeço ao meu pai Marco Antonio Cardoso Vilarinho por sempre ter acreditado no meu potencial e apostado em mim. À minha mãe Rosangela Avilez Vilarinho, que mesmo longe, nunca deixou me faltar atenção ou carinho. E também à minha irmã Luciana Avilez Vilarinho por todo carinho e por sempre ter sido uma grande amiga.

Agredeço à Rafaella Dias, companheira, confidente e namorada. Obrigado por ter me ajudado a manter o foco e motivado.

Aos amigos da República DuBodi, família que me ofereceu um lar durante toda essa caminhada. Amigos com quem chorei, com quem sorri. Amigos que levarei para toda vida.

À minha professora e orientadora Janniele que sem sua paciência e dedicação este trabalho não seria possível.

À todos os professores e colaboradores da UFOP por terem compartilhado o que permitiu que eu agregasse conhecimento suficiente para chegar até aqui.

Sumário

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

BD	Banco de Dados
CCI	*Correctly Classified Instances*
KDD	*Knowledge Discovery in Database*
KNN	*K-nearest neighbours*
MD	Mineração de Dados
TP	*True Positive*

1 Introdução

A Dengue é uma das doenças mais difíceis de se combater atualmente. De acordo com World Health Organization (2009) estima-se que de cerca 50 milhões de infecções desse tipo ocorram por ano. Especialistas no mundo inteiro estão alarmados com esta doença, pois mesmo com os métodos de combate já existentes, ela continua a se disseminar para novos países, mudando de comportamento e ganhando resistência.

Em Saúde (2015) é afirmado que só no Brasil, de 1994 a 2014, mais de 9 milhões de pessoas foram infectadas, o que resultou em mais de 1800 mortes. A Dengue é uma doença transmitida pelos mosquitos Aedes aegypti e pelo Aëdes albopictus.

Até o momento, a melhor forma de combate a doença é evitando que o mosquito transmissor se reproduza. Ele utiliza locais com água acumulada para depositar seus ovos, como caixas d'águas descobertas, calhas entupidas e pneus, por exemplo. Este mesmo mosquito também é responsável pela transmissão da Febre Amarela, do vírus da Zika e da Febre Chikungunya.

A utilização de métodos eficientes para combater a proliferação do mosquito é muito importante para evitar que essas doenças atinjam a população. Os métodos tradicionais incluem inspeção domiciliar, controle químico (fumacê) e mais recentemente o controle biológico (BRASIL, MINISTÉRIO DA SAÚDE, 2009). Para que esses métodos sejam mais eficazes, é necessário que os focos de reprodução do mosquito sejam identificados, reforçando a importância da utilização de métodos eficientes pelos órgãos responsáveis.

Hoje, a tecnologia nos permite coletar e armazenar cada vez maiores quantidades de dados e uma das formas de transformar esses dados em informações úteis é utilizando técnicas de Mineração de Dados (MD). A MD é uma técnica utilizada para a obtenção de informações a partir de grandes quantidades de dados. Ela é capaz de analisar diferentes tipos de elementos e encontrar diferentes tipos de relações entre eles.

Neste trabalho, serão utilizadas as tarefas de Clusterização e Classificação. A primeira é uma técnica de MD capaz de agrupar cada instância de acordo com a semelhança entre seus atributos. Já a segunda, foi utilizada para identificar perfis e classificar os municípios brasileiros em relação a quantidade de casos notificados de dengue. Assim, será possível demonstrar quais fatores estão relacionados a quantidade de casos da doença. Esses perfis permitirão um maior direcionamento das políticas públicas aos municípios de acordo com a sua classificação.

1.1 Objetivos

O objetivo geral deste trabalho é identificar quais fatores estão mais relacionados aos casos de Dengue de cada município brasileiro. Assim, classificar cada Município de acordo com o seu perfil.

1.1.1 Objetivos Específicos

Para que seja possível atingir o objetivo principal descrito acima, é essencial que durante o processo, sejam alcançados os seguintes objetivos específicos:

- Levantar e preprocessar dados dos municípios brasileiros relativos aos casos de Dengue;
- Levantar e preprocessar dados socioeconômicos dos municípios brasileiros;
- Categorizar cada município em relação a quantidade de casos de dengue;
- Classificar os municípios levando em consideração os dados encontrados e as classes de Dengue;
- Analisar e reportar as análises dos experimentos.

2 Revisão Bibliográfica

A Dengue é uma doença transmitida no Brasil pelos mosquitos *Aedes aegypti* e *Aedes albopictus*. Em Brasil, Ministério da Saúde (2001) ela é caracterizada como uma doença febril aguda com dores musculares e articulares intensas. O seu agente é o arbovírus do gênero Flavivírus da família Flaviviridae, do qual existem quatro sorotipos: DEN-1, DEN-2, DEN-3 e DEN-4. Ela é uma enfermidade de áreas tropicais e subtropicais pois as condições climáticas favorecem a reprodução dos vetores de transmissão.

O mosquito para se tornar um transmissor, primeiro a fêmea deve picar alguma pessoa na fase virêmica da doença. A partir deste momento, este mosquito se torna um potencial transmissor através das suas picadas por toda a vida (BRASIL, MINISTÉRIO DA SAÚDE, 2001).

Ainda sobre a Dengue, na forma clássica, apresenta baixa letalidade, mesmo podendo incapacitar de trabalho o indivíduo infectado. Na forma mais séria, a Dengue Hemorrágica, a febre é alta, apresenta manifestações hemorrágicas, hepatomegalia e insuficiência circulatória. Se não tratada, apresenta uma taxa de letalidade significativamente maior do que na forma clássica.

O mosquito transmissor da Dengue mais presente no Brasil é o *Aedes aegypti*. Ele pertence ao RAMO Arthropoda (pés articulados), CLASSE Hexapoda (três pares de patas), ORDEM Diptera (um par de asas anterior funcional e um par posterior transformado em halteres), FAMÍLIA Culicidae, GÊNERO Aedes. Pode ser encontrado em todo mundo, principalmente nas áreas tropicais e subtropicais entre as latitudes 35°N e 35°S e nas altitudes até 1000 metros acima do mar. Para se reproduzir, o mosquito prefere locais com armazenamento de água limpa, principalmente relacionados a atividades do homem: calhas entupidas, pneus descobertos, garrafas e caixas d'águas destampadas. O mesmo mosquito ainda é capaz de transmitir outras doenças como Febre Amarela, vírus da Zika e da Febre Chikungunya (BRASIL, MINISTÉRIO DA SAÚDE, 2001).

2.1 *Knowledge Discovery in Database*

Com a redução dos custos de armazenamento de dados, várias empresas começaram a armazenar informações sobre seus clientes e serviços. Com o tempo, o acúmulo desses dados permitiu que especialistas pudessem obter informações úteis para as empresas. Estas informações eram utilizadas para otimizar processos dentro das empresas e também melhorar o relacionamento com cliente, resultando em aumento dos lucros empresariais (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

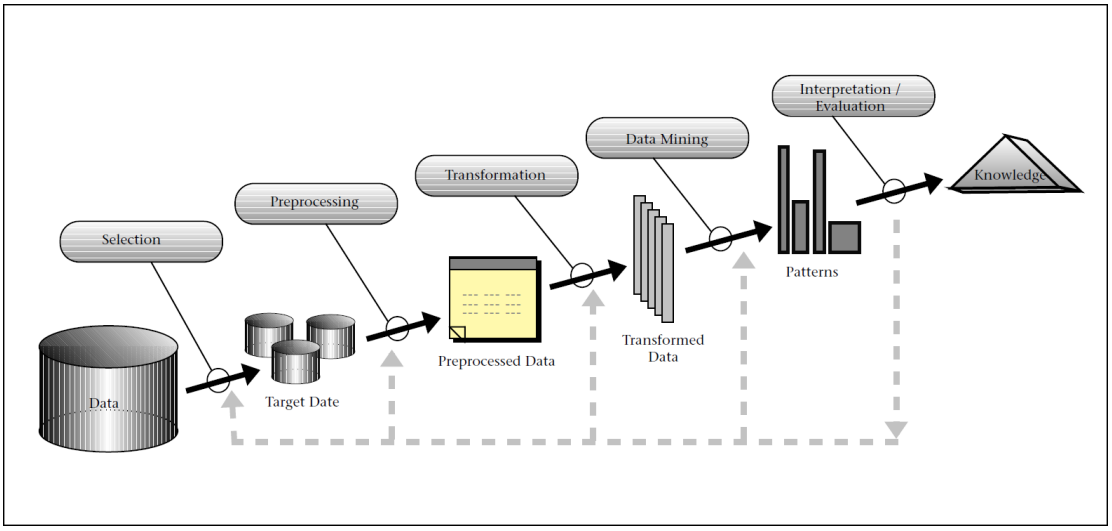
Inicialmente, as análises eram feitas por especialistas de cada área. As técnicas utilizadas ainda eram rudimentares pois eram caras, lentas e como dependiam dos técnicos, muitas vezes subjetivas. Quando a quantidade de dados começou a ficar muito grande, este antigo modelo se tornou inviável. E assim, os pesquisadores propuseram uma nova metodologia capaz de tornar este processo mais eficiente (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

O *Knowledge Discovery in Database* (KDD) segundo Fayyad, Piatetsky-Shapiro e Smyth (1996) refere-se a um processo não-trivial de descoberta de padrões potencialmente úteis a partir de uma grande quantidade de dados. O processo do KDD é interativo e iterativo, com vários passos a serem tomados decididos pelo usuário. Alguns dos passos que podem ser tomados são:

- Antes do início: Antes de se iniciar o KDD propriamente dito, é muito importante a compreensão do domínio da aplicação, e também quais são os conhecimentos relevantes. Deve-se entender qual tipo de informação o cliente está esperando dos resultados.
- Seleção: Escolher quais dos dados disponíveis são importantes e deverão ser utilizados durante o processo. Possui impacto significativo na qualidade do resultado final.
- Pré-processamento e limpeza dos Dados: Nesta etapa deve-se preparar os dados para que eles possam ser utilizados: eliminação ou correção de ruídos (dados discrepantes ou fora do padrão) e decisão de como serão tratados os dados ausentes. A qualidade dos dados obtidos aqui, definirão a eficiência dos algoritmos de mineração.
- Transformação: Modular os dados de tal forma a obter representações úteis para o estudo. Neste momento, algumas variáveis podem ser combinadas de tal forma a facilitar o processo e também devem ser armazenadas em um único repositório.
- Mineração de Dados: Escolha do algoritmo de mineração de dados apropriado para o objetivo do KDD. Entre as opções, estão: Classificação, Agrupamento e Regressão, por exemplo. Mais detalhes são descritos em 2.2.
- Interpretação: Talvez a etapa mais complexa, tem por objetivo entender os resultados gerados pela MD e como eles podem ser utilizados para ajudar a empresa. São necessários profissionais de várias áreas para validar e interpretar as informações geradas no passo anterior.

Pela observação da Figura 1 pode-se compreender melhor os passos do processo KDD.

Figura 1 – Passos do processo KDD.



Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

2.2 Mineração de Dados

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), Mineração de Dados (MD) é uma das etapas do KDD. Nesta etapa estão envolvidos a aplicação iterativa e repetitiva de métodos particulares de MD para a obtenção de padrões. Os objetivos das tarefas dirão quais os tipos de algoritmos utilizados. Os algoritmos de Classificação possuem o objetivo de "aprender"a classificar registros em classes pré definidas. É explicado em mais detalhes em 2.4. As tarefas de Regressão são utilizadas quando o objetivo é prever resultados futuros a partir de registros antigos. Os algoritmos de Agrupamento procuram identificar grupos ou *clusters* em um conjunto de registros a partir de seus atributos, como descrito em 3.2. Já quando se deseja procurar relacionamentos entre os registro de um BD, utiliza-se os algoritmos de Associação.

2.3 Tarefas de Agrupamento

Segundo Han, Pei e Kamber (2011) a tarefa de Agrupamento de dados ou *Clustering* tem como objetivo descobrir subconjuntos, ou *clusters*, de registros dentro de um conjunto de dados. Os subconjuntos são determinados pela semelhança entre os seus registros e pela discrepância em relação aos atributos dos registros de diferentes *clusters*.

Essa tarefa possui aplicação em várias áreas, tais como: biologia, onde pode ser utilizada para encontrar genes com funções semelhantes; área de telecomunicação, com o intuito de determinar o melhor posicionamento de antenas; setores de marketing e relacionamento com o cliente, onde pode ser utilizada para segmentar os clientes de acordo com os seus interesses ou necessidades; e, outras áreas.

Entre os vários algoritmos de agrupamento existentes, Witten et al. (2011) descreve a técnica *K-Means*. No primeiro momento, é necessário especificar o valor de K, que representa a quantidade de *clusters* que se deseja encontrar. Em seguida, K pontos iniciais são escolhidos aleatoriamente. Cada registro é então associado ao ponto mais próximo. Depois, utiliza-se a distância Euclidiana para recalcular o ponto central de cada *cluster*, denominados centroides, estes são utilizados como os novos centros dos *clusters*. Este processo é repetido até que os centroides não mudem de posição.

Neste estudo, o algoritmo *k-means* foi utilizado para identificar as classes relacionadas as notificações dos casos de dengue de cada município. Este foi um importante passo para que pudesse ser realizada a próxima tarefa, a de Classificação dos municípios.

2.4 Tarefas de Classificação

Em seu livro, Tan, Steinbach e Kumar (2005) caracteriza a tarefa de Classificação como a determinação de uma função capaz de relacionar cada registro de um banco de dados a uma classe a partir dos seus atributos. As classes devem ser pré-definidas e nominais. Desta forma, seria possível dizer a qual classe um novo registro pertenceria.

Ela pode ser utiliza para vários propósitos como detecção de spans em e-mails a partir dos cabeçalhos e conteúdo, classificação de galáxias baseadas nos formatos, ou classificação de municípios a partir de alguns dos seus atributos, como é o objetivo deste trabalho.

As classes utilizadas neste trabalho serão definidas de acordo com a Tarefa de Agrupamento detalhada em 3.2.

2.4.1 kNN

O algoritmo kNN (*k-Nearest neighbours*) é um algoritmo de Classificação de vizinhos mais próximos 1. Para predizer a classificação de um novo registro, o algoritmo compara este registro às instâncias já classificadas. A nova instância será atribuída a classe dos k registros mais próximos (FERRERO, 2009). Seu pseudocódigo é apresentado em 1.

Algoritmo 1: `kNN`

Data: k: número de vizinhos mais próximos, D o conjunto de exemplos de treinamento

for *(cada exemplo de teste* $z = (x', y')$*)* **do**

  Calcule $d(x', x)$, a distância entre z e cada exemplo $(x, y) \in D$;

  Seleciona $D_z \subseteq D$ o conjunto dos k exemplos de treinamento para z;

  $y' = argmax \sum_i (x_i, y_i) \in D_z \forall\ I(v = y_i)$;

end

retorna y'

Onde:

- k: número de vizinhos mais próximos dado pelo usuário;
- D: conjunto de treinamento;
- z: conjunto de teste;
- $d(x', x)$: cálculo de distância entre o item de teste e todas as instâncias de treinamento;
- $y' = argmax \sum (x_i, y_i) \in D_z I(v = y_i)$: votação majoritária, onde o item de teste é classificado de acordo com a classe predominante dentro dos K vizinhos mais próximos.

Existem alguma funções que podem ser utilizadas para calcular a distância entre instâncias, como a Distância Euclidiana, Distância de Manhattan e *city-block*. Witten et al. (2011) afirma que a fórmula mais utilizada é a Euclidiana, e por isso ela foi adotada neste trabalho. A distancia entre uma instância com atributos de valores ${a_1}^{(1)}, {a_2}^{(1)}, ..., {a_k}^{(1)}$ (aonde k é o número de atributos) e outra com os valores ${a_1}^{(2)}, {a_2}^{(2)}, ..., {a_k}^{(2)}$ é dada por:

$$\sqrt{({a_1}^{(1)} - {a_1}^{(2)})^2 + ({a_2}^{(1)} - {a_2}^{(2)})^2 + ... + ({a_k}^{(1)} - {a_k}^{(2)})^2}$$

Para exemplificar o algoritmo, podemos citar o exemplo dado por Ferrero (2009). Em um consultório médico deseja-se classificar os pacientes de acordo com os tipos de sintomas apresentados. Primeiramente deve-se criar um conjunto de treinamento (*Training Set*) a partir dos registros de pacientes antigos. Este conjunto possuirá registros nos quais os atributos serão os sintomas e a classificação corresponde ao fado do paciente possuir ou não determinada doença. Na Figura 2, as bolinhas azuis representam registros em que o paciente estava de fato doente. As verdes representam que o paciente não estava doente. A bola vermelha representa o objeto que se deseja classificar, no caso, o novo paciente.

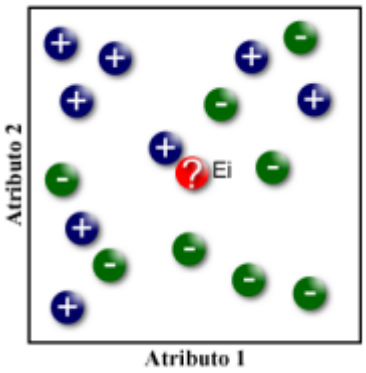
Duas configurações poderiam influenciar no próximo passo. O valor de K representa quantos vizinhos deverão ser levado em consideração. A segunda é relacionada a qual Função de Distância seria utilizada. Para os valores de k = 1 e utilizando a Distância Euclidiana, a esfera vermelha, representando o novo paciente, seria considerado Doente, por estar mais próxima de uma bola azul.

2.4.2 C4.5

O algoritmo C4.5 permite a descoberta de padrões na forma de um classificador representado em formato de Árvore de Decisão.

Salzberg (1994) explica que os algoritmos de árvore de decisão começam com um conjunto de casos, ou registros, e criam uma estrutura de dados (em formato de árvore) que

Figura 2 – Exemplo do Algoritmo Knn.



Fonte: elaborado pelo autor

pode ser utilizada para classificar novos casos. Cada caso é descrito como um conjunto de atributos que podem ser numéricos ou nominais (registros). Um identificador é associado a cada caso de treinamento para representar a classe. Desta forma, cada nó interno da árvore de decisão possui um teste. Este teste é utilizado para determinar qual o próximo passo a ser tomado, seja o nó esquerdo ou direito. Para decidir quais testes serão utilizados em cada nó, o algoritmo C4.5 utiliza fórmulas baseadas na Teoria da Informação para avaliar o quão bom um teste é. Ele escolhe os testes que extraem a máximo de informações de um conjunto de registros, dada a restrição que apenas um atributo será testado de cada vez.

Para evitar o problema de *overfitting*, vários algoritmos de árvore de decisão utilizam técnicas de *pruning*, ou "poda". Ou seja, após crescer a árvore, pedaços dela são cortados. Salzberg (1994) diz que o processo de poda do algoritmo C4.5 é baseado na estimativa da taxa de erro de cada sub-árvore, quando a folha possuir uma estiva menor do que a sub-árvore, ela é substituída pela folha menor.

2.4.3 kStar e bayvesNet

Cleary, Trigg et al. (1995) define kStar(k*) como um classificador baseado em instância, ou seja, baseado na comparação com um BD de instâncias pré-classificadas. Desta forma, instâncias similares, receberão classificações semelhantes. Este algoritmo utiliza a Entropia baseada na Teoria da Informação, para calcular a semelhança entre suas instâncias. O valor da Entropia pode ser definido como a complexidade de uma instância se transformar em outra.

O programa WEKA (2.7) oferece vários algoritmos para a implementação de Redes Bayesianas com diferentes métricas. Neste estudo foi utilizado o algoritmo de subida de encosta K2.

As Tarefas de Classificação foram as mais exploradas neste trabalho. Elas foram utilizadas para gerar classificadores para os municípios brasileiros de acordo com os casos

de dengue apresentados.

2.5 Métricas de Avaliação

Comparar e avaliar o desempenho de algoritmos de MD não é uma tarefa simples e é importante para se evitar falsos positivos. Witten et al. (2011) descreve as métricas que foram utilizadas neste trabalho:

- Taxa de acerto (*Correctly Classified Instances*): Porcentagem de instâncias classificadas corretamente.
- Taxa de erro (*Incorrectly Classified Instances*): Porcentagem de instâncias classificadas incorretamente.
- *TP Rate* (*True Positive*): Instâncias pertencentes a uma classe "A"classificadas corretamente divididas pela quantidade total de instâncias classificadas como "A".
- *FP Rate* (*False Positive*): Proporção de instâncias classificadas erroneamente como uma determinada classe "A".
- Precisão: Proporção das instâncias que realmente pertencem a uma classe "A", dividida pela quantidade de instâncias classificadas na mesma classe "A".
- *Recall*: Taxa que mostra quantos dados foram classificados corretamente. Calculado pela seguinte fórmula: True Positive / quantidade de instâncias de registros dessa classe.
- *F-measure*: Medida de combinação calculada pela fórmula 2*Precisão**Recall* / (Precisão + *Recall*)
- *Kappa*: Medida de concordância utilizada junto com escalas nominais. Seu objetivo é fornecer uma noção de quanto as predições se afastam das classificações corretas. Essa estatística é uma forte avaliação dos resultados de um algoritmo de MD. Seu valor varia de zero a um. Quanto mais próximo de um, melhor a capacidade do algoritmo na classificação dos dados.
- Matriz de Confusão (*Confusion Matrix*): Demonstra a quantidade de instâncias classificadas em cada classe. Na tabela 1, podemos perceber que todas as instâncias pertencentes ao *Cluster* A (a = Iris-setosa), foram classificadas corretamente. Pois todas as instancias estão na coluna "a", e as outras estão vazias. Já em relação as instâncias do *Cluster* C (c = Iris-virginica), 46 foram classificadas corretamente, porém quatro foram classificadas erroneamente como pertencentes ao *Cluster* B. O mesmo ocorreu em relação ao *Cluster* B (b = Iris-versicolor), 48 classificações

corretas e duas erradas. Esta métrica nos permite visualizar com mais facilidade os erros e acertos dos algoritmos.

Tabela 1 – Exemplo de uma Matriz de Confusão

a	b	c	<– Classificado como
50	0	0	a = Iris-setosa
0	48	2	b = Iris-versicolor
0	4	46	c = Iris-virginica

Fonte: elaborado pelo autor

Outras métricas como *Root mean squared error*, *Relative absolute error* e *Root relative squared error* são mais utilizadas nas tarefas de Predição, a qual não é o foco deste estudo.

2.6 *Cross-Validation*

A validação dos resultados de um algoritmo de Classificação não é uma atividade trivial. Witten et al. (2011) afirma que a Taxa de erro não é suficiente para validar a maioria dos casos, pois ela representa apenas a capacidade do algoritmo em classificar um registro em relação ao conjunto de treinamento (*Training Set*). Como na maioria das vezes ao utilizarmos tarefas de Classificação estamos interessados em classificar novos registros, faz-se necessário utilizar outras metodologias.

Em situações em que possuíamos um grande BD para utilização, podemos dividi-lo em duas ou três partes: Conjunto de Treino (*Training Set*), Conjunto de Teste (*Test Set*) e Conjunto de Validação (*Validation Set*), caso necessário. No primeiro conjunto o algoritmo realiza o seu processo de aprendizado, determinando a melhor forma de classificar cada instância. O segundo conjunto é utilizado para testar as funções que foram aprendidas no passo anterior. E o terceiro grupo é utilizado para validar estas funções.

O problema está no fato de que nem sempre possuímos um conjunto grande o bastante para ser dividido em tantas partes. Subconjuntos podem não possuir a capacidade de representar o conjunto inicial o que causa um mau desempenho do algoritmo.

Entre as formas de mitigar os efeitos deste problema, Witten et al. (2011) destaca a técnica de *Cross-Validation* (Validação Cruzada). Neste procedimento é escolhido o número de *Folds* ou partições dos dados. Se considerarmos o número de *Folds* igual a quatro, teremos os dados divididos em quatro partes. Desta forma, a técnica realizará o treinamento do algoritmo de Classificação utilizando cada vez, três quartos como Conjunto de Treinamento, e um quarto como conjunto de teste. Esse procedimento será repetido até que cada registro tenha sido utilizado uma vez para testes. Devido a vários estudos

utilizando diferentes técnicas de aprendizado e bancos, o valor que apresentou os melhores resultados foi 10, porém, em algumas situações este fato pode não se repetir.

2.7 WEKA

Witten et al. (2016) descreve o WEKA como uma coleção de algoritmos de aprendizado de máquina e ferramentas úteis para o processamento de dados. Desenvolvida na Universidade de Waikato, Nova Zelândia, de forma que fosse possível testar diferentes técnicas em diferentes bancos de dados rapidamente.

Em relação a MD, o WEKA possui ferramentas para tarefas de regressão, classificação, agrupamento (*clustering*), associação e seleção de atributos. Esta ferramenta foi escolhida para a realização das tarefas de MD envolvendo Agrupamento e Classificação de dados. A facilidade de uso, vasta documentação disponível, concentração de algoritmos, gratuidade e possibilidade de conexão direta com banco de dados foram essenciais para a escolha desse software.

2.8 Trabalhos Relacionados

O estudo de Vianna et al. (2010) não possui relação direta com a Dengue, porém utilizou a metodologia e os conceitos de Mineração de Dados que desejamos aplicar em nossa pesquisa. Seu objetivo foi identificar padrões de características maternais e fetais que estivessem relacionadas a mortalidade infantil, utilizando a Mineração de Dados como ferramenta. O artigo teve como foco a análise dos óbitos infantis ocorridos no Paraná, no período de 2000 a 2004.

Vianna et al. (2010) obteve seus dados a partir de três Sistemas de Informação em Saúde (SIS): o Sistema de Informações sobre Mortalidade (SIM), o Sistema de Informações sobre Nascidos Vivos (SINASC) e o Sistema de Investigação da Mortalidade Infantil (SIMI) e os consolidou em um único banco de dados para a posterior mineração de dados. A segunda etapa pré processou os dados, neste momento os dados foram categorizados de acordo com as necessidades apontadas por especialistas da área médica. Também foi feita uma filtragem nos dados para detectar e corrigir erros de preenchimento. A mineração de dados em si, começou com a escolha do algoritmo J48 do WEKA para identificar as regras que relacionam as diversas variáveis do banco de dados. Este algoritmo permite a descoberta de padrões na forma de um classificador representado como uma árvore de decisão, o que facilita o entendimento tanto acadêmico, quanto comercial. Um segundo painel de especialistas foi responsável por escolher as variáveis que seriam submetidas à mineração, tendo como foco a caracterização da evitabilidade dos óbitos e a especificação da causa do óbito, de tal forma que a informação obtida facilitasse o combate. A tarefa de

mineração utilizada foi a classificação, com o intuito de se obter uma previsão, padrões que se repetem de tal forma que, dadas as mesmas, ou parecidas, circunstâncias, o evento se repetirá. No caso, dado os mesmos atributos a mãe, e ao bebê, a sua morte poderá, ou não, ser evitada. Ainda durante a mineração de dados, 22 especialistas foram consultados para verificar se os padrões obtidos condiziam com as suas próprias experiências e com as literaturas existentes.

Em seus resultados, Vianna et al. (2010) observou que 55% dos óbitos teriam sido evitados caso estas gestações tivessem recebido uma atenção diferenciada. A integração de todos esses sistemas de saúde permite uma visão mais ampla da realidade, disponibilizando dados mais úteis para a sociedade e às autoridades. As informações obtidas por esse estudo são importantes para a construção de mapas de risco, os quais seriam utilizados como uma bússola durante a definição das estratégias de ação governamental. Este trabalho reiterou a importância e o quão necessário é a utilização de novos recursos tecnológicos na área de Saúde Pública.

O trabalho de Pessanha et al. (2012) estudou a distribuição dos casos de Dengue em Belo Horizonte, Minas Gerais, durante o intervalo entre 1996 e 2011. Nesse período, a cidade foi vítima de 186.000 casos registrados de Dengue. Alguns programas nacionais de combate ao mosquito foram desenvolvidos, com destaque ao Programa de Erradicação do Aedes aegypti (PEAa), aplicado na cidade em 1998. Com epidemias acontecendo anualmente, desde o primeiro ano da pesquisa, estudos destacam que mesmo em cidades onde as medidas de combate à doença foram bem aplicadas, os programas ainda não conseguiram demonstrar resultados satisfatórios.

Pessanha et al. (2012) utilizou os seguintes dados em sua pesquisa: notificações de casos de dengue, informações sobre coletas em ovitrampas, altitude, umidade relativa do ar, quantidade de chuva e temperatura média. A correlação de Pearson permitiu que fosse criada uma correspondência entre os dados climáticos e os casos de dengue entre 2001 e 2010. A estimativa de densidade Kernel foi utilizada como ferramenta na análise estatística espacial. E o índice de Infestação Predial estimou a proporção de casas com larvas do mosquito a partir dos dados obtidos nas ovitrampas.

O estudo dos padrões de disseminação da Dengue e os fatores que o influenciam são importantes, afirma Pessanha et al. (2012), porém, para que esta técnica funcione, é necessário que as medidas de combate ao vetor sejam contínuas, e direcionadas às regiões em que são detectados focos do mosquito. A utilização das ovitrampas se mostrou fundamental, para a identificação e quantificação dos focos.

A análise temporal dos fatores climáticos identificou que a dengue geralmente ocorre quando a temperatura média sobe, durante o início da época da chuva e quando a umidade está mais alta. Pessanha et al. (2012) ainda reiterou que o acompanhamento dos casos de Dengue durante os meses frios e secos também devem ser observados a fim de

contribuir com os estudos para o controle do vetor.

A pesquisa encontrou relações entre os argumentos utilizados, demonstrando que pesquisas de autocorrelação espacial podem ser utilizadas no planejamento das políticas de combate ao mosquito, uma vez que conseguem identificar os locais com maior probabilidade de reprodução do Aedes Aegypti. Pessanha et al. (2012) afirma que a utilização de notificações de casos de Dengue por si só, são insuficientes para um estudo mais aprofundado deste tipo, pois muitas pessoas infectadas podem não apresentar sintomas ou não entrarem para a da estatística oficial, mesmo que apresentem sintomas.

Poucos municípios brasileiros dispõem de todos os dados utilizados na pesquisa de Pessanha et al. (2012). Os dados climáticos e sobre ovitrampas encontrados, não foram suficientes para fazer parte desta pesquisa, por isso, para realizar este estudo, foi dado o foco aos dados socioeconômicos e populacionais das municípios.

Em 2015, Shaukat et al. (2015) publicou um estudo comparando diferentes tipos de algoritmos de agrupamento de dados aplicados a tarefa de agrupamento dos casos de Dengue em uma região do Paquistão. A Dengue se tornou uma grande ameaça para este país, devido à inexistência de vacinas preventivas na época e a baixa qualidade dos sistemas de saneamento e esgoto em áreas altamente povoadas, condições ideais para a proliferação do mosquito e consequentemente a disseminação da doença.

Muitas pesquisas não levam em consideração que algumas pessoas podem ser infectadas em outras regiões, e não somente nos locais em que residem. Os algoritmos utilizados por Shaukat et al. (2015) são baseados em densidade e por isso, são menos sensíveis às influências causadas pelos "carregadores de Dengue". A área estudada se limitou ao distrito de Jhelum, na província de Punjabm, Paquistão.

Existem vários algoritmos que podem ser utilizados nas tarefas de classificação, porém, cada um apresenta vantagens e desvantagens de acordo com o tipo de tarefa a que é submetido. A pesquisa de Shaukat et al. (2015) fez a comparação entre quatro deles e obteve algumas observações importantes. O algoritmo *K-mediods* resolveu o problema de visualização ao "reduzir o cerco" de cada *cluster*, porém ainda não apresentou resultados satisfatórios. O outro algoritmo testado foi o DBSCAN, ele foi capaz de identificar um *cluster* em cada *tehsil* (subdivisão administrativa de distritos utilizada no Paquistão). O melhor resultado obtido foi o do algoritmo OPTICS, ele foi capaz de encontrar a maior quantidade de *clusters*, permitindo a visualização em maior detalhes dos casos de Dengue dentro de cada *tehsil*, informação ideal para facilitar medidas de combate ao mosquito. De acordo com as observações de Shaukat et al. (2015) serão evitados os algoritmos *k-mediods* e DBSCAN no desenvolvimento do nosso trabalho devido à sua má performance nas tarefas de agrupamento.

Shaukat et al. (2015) conseguiu demonstrar a importância da utilização de algorit-

mos baseados em densidade no estudo de agrupamento de doenças a partir da comparação de diferentes algoritmos. Esta técnica permite uma melhor visualização da distribuição dos casos de Dengue, e assim ser utilizada de forma mais eficiente nas medidas de combate ao mosquito. Esta técnica não foi utilizada pois seu foco é obter informações sobre a distribuição de dengue em áreas menores, como os *tehsils*. O objetivo deste estudo está na obtenção de informações para todo o Brasil.

3 Desenvolvimento

O combate a dengue é uma tarefa simples e que todos devem praticar. Esvaziar e cobrir recipientes que possam armazenar água e manter calhas de água limpas são atitudes essenciais ao combate. Porém, até o momento, essas medidas não tem sido suficientes.

Resta então ao governo, desenvolver políticas de prevenção e combate à dengue. Por isso, identificar os locais mais suscetíveis a reprodução do mosquito e consequentemente a propagação da doença é uma tarefa fundamental para auxiliar os órgãos responsáveis no direcionamento dos recursos.

A proposta desse trabalho é fornecer mais um meio de auxílio de combate ao mosquito.

3.1 Pré-Processamento de Dados

A coleta de dados teve como objetivo reunir a maior quantidade de dados que pudessem ser relacionados ao mosquito transmissor e aos casos de dengue, a nível municipal e entre os anos 2000 e 2015. Foram utilizados as informações provenientes dos seguintes sites: Atlas IDHM [1], ipeadata [2], Acesso à Informação Governo Federal [3] e IBGE [4]. A tabela 2 relaciona cada um dos atributos escolhidos aos anos disponíveis. Na primeira linha da tabela são representados os anos.

Tabela 2 – Atributos disponíveis

	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
População	x										x					
IDHM	x										x					
IDHM - Longevidade	x										x					
IDHM - Educação	x										x					
IDHM - Renda	x										x					
Tx pop água encanada	x										x					
Tx pop coleta lixo	x										x					
Tx pop energia elétrica	x										x					
Tx desemprego pop <= 18	x										x					
Dengue	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Bolsa Família					x	x	x	x	x	x	x	x	x			

Fonte: elaborado pelo autor

Como demonstrado na Tabela 2, as pesquisas de âmbito municipal não são realizadas todos os anos. Por isso, os anos de 2000 e 2010 foram selecionados como foco do estudo pois possuíam maior disponibilidade de dados. Os atributos escolhidos foram:

[1] http://www.atlasbrasil.org.br/2013/pt/
[2] http://www.ipeadata.gov.br/Default.aspx
[3] http://www.consultaesic.cgu.gov.br/busca/SitePages/principal.aspx
[4] http://seriesestatisticas.ibge.gov.br/

1. População: quantidade de moradores de um determinado município.
2. IDHM (Índice de Desenvolvimento Humano Municipal): adaptação do IDH Global para calcular o IDH dos municípios brasileiros. Utilizou como fonte, os dados dos censos demográficos do IBGE de 1991, 2000 e 2010. Possui três dimensões, assim como o IDH Global, Renda, Longevidade e Educação, que também foram selecionados para o estudo (PNUD, IPEA, FUNDAÇÃO JOÃO PINHEIRO, 2013).
3. IDHM Longevidade: representa qual é a expectativa de vida média ao nascer dos moradores de um município, mantidos os mesmos padrões de mortalidade.
4. IDHM Educação: relação entre o percentual das pessoas com 18 anos ou mais com ensino fundamental completo e o fluxo escolar da população jovem.
5. IDHM Renda: representa a soma da renda média de todos os residentes de um determinado município dividida pela quantidade de moradores deste mesmo município.
6. Taxa da população que vive em domicílios com água encanada.
7. Taxa da população que vive em domicílios com serviço de coleta de lixo.
8. Taxa da população que vive em domicílios com energia elétrica.
9. Taxa de Desocupação de pessoas com 18 anos ou mais: representa a quantidade de indivíduos maiores de idade que estavam desocupados, mas haviam procurado trabalho ao longo do mês anterior a pesquisa.
10. Casos de Dengue notificados: Quantidade de casos de dengue notificados (confirmados, ou não) ao órgão responsável de cada município em um período de um ano.
11. Bolsa Família: Valor total transferidos através do programa Bolsa Família em Dezembro do ano escolhido. O Bolsa Família é um programa com objetivo de transferência de renda para famílias pobres. Foi instituído a partir de 2004. O valor do benefício varia conforme a renda domiciliar per capita da família, número e idade dos filhos.

Os dados foram obtidos de diversas fontes, e por isso, possuíam formatos, escalas e padrões diferentes. O pré-processamento envolveu transportar esses dados para planilhas no Excel, aonde os dados foram formatos, padronizados e os ruídos eliminados.

Na modulação, um dos processos realizados envolveu os dados relacionados a casos de dengue por semana epidemiológica. Eles foram condensados em total por ano e transportados para uma única planilha, contendo o total de casos notificados por município e por ano. Essa e todas as outras tabelas foram transferidas para um banco de dados MySQL. Para permitir que diferentes tipos de análise pudessem ser feitas, foram criadas *views* relacionando os diferentes atributos. Uma vez os dados no Banco de Dados (BD), foi possível seguir para o passo de MD.

3.2 Agrupamento

A tarefa de Agrupamento foi escolhida com o objetivo de identificar *clusters* na distribuição dos casos de dengue pois é capaz de identificar grupos a partir dos atributos de cada registro.

Os valores dos casos de dengue por si só, não devem ser comparados com os de outros municípios. Devemos lembrar que cada município possui uma população diferente, desta forma, 100 casos de dengue em um município com população igual a 1000, possui um peso muito maior (10%) do que 100 casos de dengue, em um município com 1 milhão de moradores (0,01%). Desta forma, antes de rodar o algoritmo, todos atributos de notificação de dengue foram divididos pela população da sua respectiva cidade. E para fins de melhorar o desempenho do algoritmo, esses valores ainda foram multiplicados por 10000, permitindo que o algoritmo rodasse com valores mais altos e mais casas decimais. Para o ano de 2000, o valor obtido é representado pela variável pct_2000, para o ano de 2010, a variável é chamada pct_2010.

O algoritmo utilizado foi o *K-Means*, descrito em 3.2. A função de distância entre os pontos e os centroides escolhida foi a Distância Euclidiana. Para determinar o valor de K, (quantidade de *clusters*) foram realizados testes variando o seu valor de um a dez.

A melhor distribuição de registros por *clusters* foi com K igual a cinco. Desta forma, os grupos ficaram definidos da seguinte forma: Muito Baixo, Baixo, Médio, Alto e Muito Alto, de acordo com as Tabelas 4 e 5. Para K menor que cinco, os grupos encontrados eram pequenos demais para serem utilizados, e para K maior que cinco, a quantidade de grupos aumentava muito, porém o tamanho deles se mantinha pequeno.

Como o KDD é um processo iterativo e interativo, os grupos obtidos nesta etapa, puderam ser adicionados ao BD e as *views* adaptadas para a nova disposição de dados.

3.2.1 Análise

A tarefa de agrupamento por si só, já oferece várias informações para interpretação. A Tabela 3 mostra a soma das notificações de dengue por município em cada *cluster* no seu respectivo ano. O segundo ano apresenta um aumento de mais de sete vezes dos casos de dengue de 2000, informação confirmada por Roriz-Filho (2010) que afirma que 2010 foi um ano de muitos municípios em situação de epidemia de dengue.

Por outro lado, a distribuição dos casos de dengue entre os grupos não se manteve a mesma. Ao se analisar as Tabelas 4 e 5 percebemos que em 2000, 92,45% dos municípios eram classificadas como “muito baixo”, já em 2010, apenas 76,19% estiveram classificadas nesse grupo. Logo, 16,26% dos municípios tiveram que ser redistribuídos nos outros grupos. Como mais pessoas foram infectadas, o grupo “baixo” e “médio” recebeu a maior parte

Tabela 3 – Casos de Dengue

	2000	2010
Muito Baixo	16701	138804
Baixo	41254	219092
Médio	41371	273707
Alto	24608	266684
Muito Alto	11256	113231
Total	135190	1011518

Fonte: elaborado pelo autor

desses casos. Os valores de pct_2000 e pct_2010 são representados no "Limite Inferior"para os menores valores encontrados dentro do seu respectivo grupo, o mesmo vale para "Limite Superior", representam os maiores valores encontrados dentro de cada grupo.

Tabela 4 – Distribuição dos *Clusters* de Dengue. Ano 2000

	Limite Inferior	Limite Superior	Quantidade (%)
Muito Baixo	0	14,25	92,45
Baixo	14,25	52,99	5,19
Médio	52,99	127,0	1,50
Alto	127,0	280,0	0,68
Muito Alto	280,0	521,3	0,16
		Total	5565

Fonte: elaborado pelo autor

Tabela 5 – Distribuição do *Clusters* de Dengue. Ano 2010

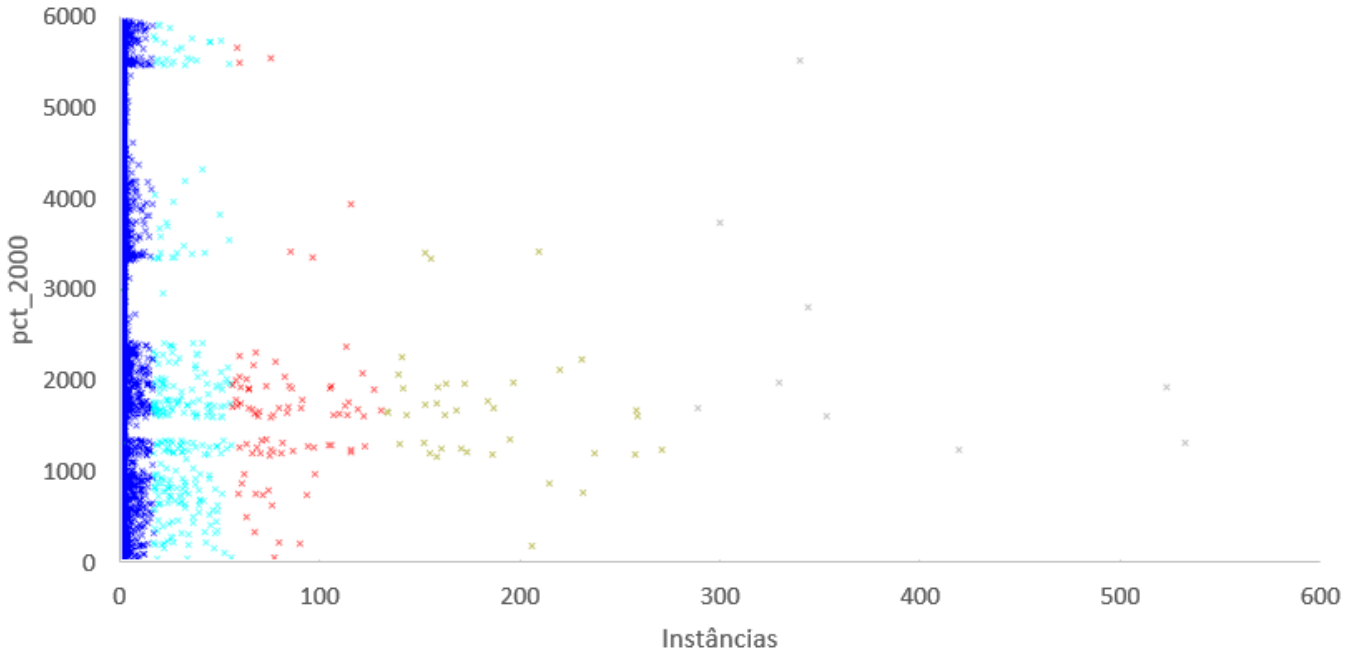
	Limite Inferior	Limite Superior	Quantidade(%)
Muito Baixo	0	42,46	76,19
Baixo	42,46	135,53	14,51
Médio	135,53	278,44	5,69
Alto	278,44	533,8	2,87
Muito Alto	533,8	957,6	0,71
		Total	5565

Fonte: elaborado pelo autor

Ainda em relação as Tabelas 4 e 5, percebemos o aumento dos limites entre os grupos. Com a maior quantidade de casos, o algoritmo se reajustou para agrupar melhor os casos semelhantes. Ou seja, as categorias, mesmo com os mesmos nomes, passaram a conter mais casos. Essa capacidade de autoajuste do algoritmo é demonstrada nas Figuras 3 e 4. Em 2000, a maioria dos casos estavam concentrados em “muito baixo”, a faixa era mais estreita e muito mais densa, enquanto as outras eram curtas, com exceção da faixa “alta” que era longa e pouco densa, como um grupo das "exceções". Em 2010, as faixas

ficaram mais extensas e mais densas. A distribuição estava mais uniforme em todos os municípios e em todos os grupos, ou seja, mais cidades com mais casos.

Figura 3 – Distribuição de Municípios em *clusters* por cor. Ano 2000



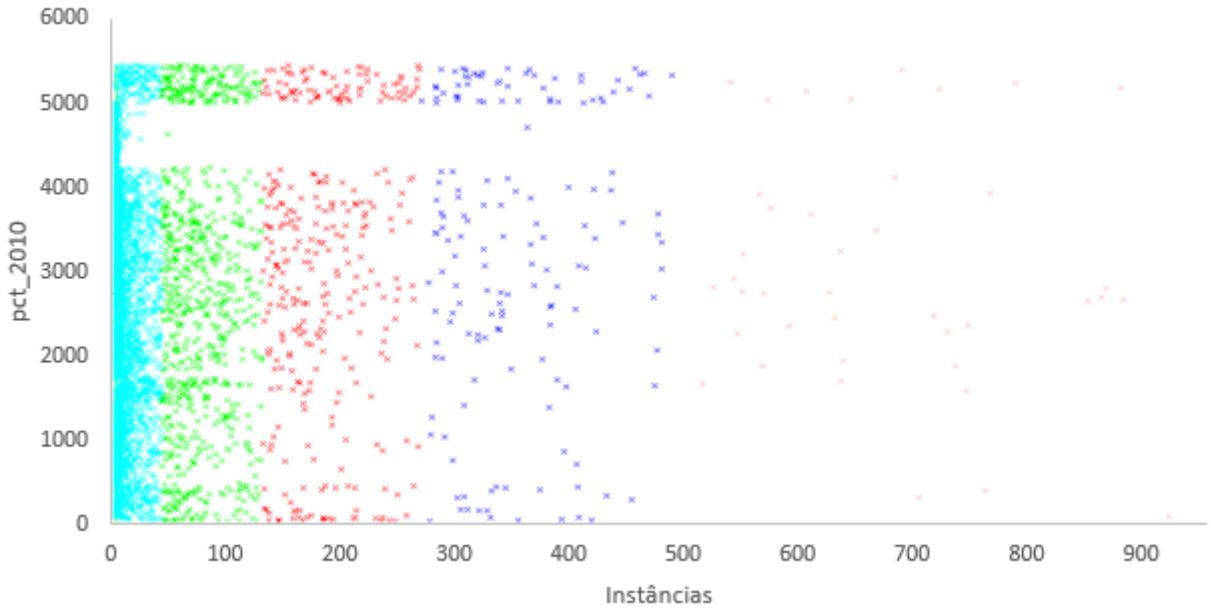
Fonte: elaborado pelo autor

3.3 Dificuldades

As pesquisas a nível municipal não ocorrem todos os anos, e isto dificulta a realização de análises temporais destes dados. Também, não foi possível encontrar a quantidade de casos confirmados de dengue por município, uma vez os órgãos responsáveis utilizam a quantidade de casos notificados de dengue, confirmados, ou não. De acordo com Saúde (2015) esta é uma forma de aumentar a sensibilidade dos dados estatísticos.

Alguns fatores climáticos como temperatura média, quantidade de chuvas, e velocidade do vento são fatores que influenciam diretamente na reprodução do mosquito. Estes atributos permitiriam classificar as cidades considerando-se também as suas características físicas. Porém, são poucas as cidades que possuem estações meteorológicas capazes de medir esses valores. Entre as que possuem, os dados não são disponibilizados em formatos amigáveis. Geralmente são disponibilizados em mapas ou tabelas individuais por cidade.

Figura 4 – Distribuição de Municípios em *clusters* por cor. Ano 2010



Fonte: elaborado pelo autor

4 Resultados e Experimentos

Foram realizados dois experimentos principais. No primeiro, foram selecionados algoritmos e definidas as melhores configurações utilizando os dados relativos ao ano de 2010. No segundo, os algoritmos foram aplicados nos dados do ano 2000. As considerações sobre os resultados obtidos ficaram na seção 3.2.1.

4.1 Experimento I

Para o experimento inicial, foram escolhidos os dados de 2010, pois, como descrito em 3.2.1, neste ano, mais municípios tiveram mais casos de dengue. Com os *clusters* menos densos e mais distintos entre si, foi mais fácil determinar os melhores algoritmos de classificação para o caso em estudo.

Os algoritmos testados no primeiro experimento foram: kNN, BayesNet, NaiveBayes, Kstar e C4.5, descritos em mais detalhes em 2.4. A primeira métrica utilizada para avaliar o desempenho desses algoritmos foi baseada no CCI. Todos eles utilizaram *Cross-Validation* com dez *folds*. Os resultados obtidos estão descritos na Tabela 6.

Tabela 6 – Resumo Experimento 1

Algoritmo	CCI	Muito Baixo	Baixo	Médio	Alto	Muito Alto
kNN	64.19%	0,794	0,184	0,103	0,046	0,023
Kstar	63.38%	0,787	0,176	0,088	0,031	0,046
BayesNet	75.92%	0,863	0,028	0	0	0
NaiveBayes	45.73%	0,672	0,032	0,040	0,079	0,012
C4.5	71.77%	0,846	0,170	0,106	0,047	0

Fonte: elaborado pelo autor

O classificador kNN foi escolhido para o segundo experimento pois obteve os melhores resultados na classificação dos registros, apesar de ser o terceiro colocado quanto ao CCI. Já o algoritmo C4.5 apresentou o segundo melhor CCI e deixou de classificar apenas a classe "Muito Alto". Este resultado pode ser melhorado alterando-se as configurações do algoritmo, por isso ele também foi selecionado para a segunda etapa. O algoritmo BayesNet apresentou o melhor CCI, porém foi incapaz de classificar atributos em três classes, e por isso não foi selecionado para o próximo experimento.

No segundo momento, foram testadas algumas variações nas configurações dos algoritmos. Para o Algoritmo C4.5 foram testados vários valores de *Seed*, porém, nenhum

deles teve impacto nos resultados obtidos. A matriz de confusão para o valor de *Seed* = 1 é apresentada na Tabela 7.

Tabela 7 – Matriz de Confusão: Algoritmo C4.5. Ano 2010

a	b	c	d	e	<– Classificado como
3859	234	57	86	4	a = MuitoBaixo
638	103	22	44	1	b = Baixo
125	19	6	10	0	c = Alto
236	41	11	26	3	d = Medio
29	5	0	6	0	e = MuitoAlto

Fonte: elaborado pelo autor

O algoritmo kNN (*K-nearest neighbors*) permite a configuração do valor de "K", referente a quantidade de vizinhos que deve ser utilizada para a classificação. Utilizando-se três *folds*, a Tabela 8 mostra os resultados obtidos. O melhor resultado obtidos, considerando-se o valor da estatística Kappa, foi K = 1. A matriz de confusão é representada na Tabela 9.

Tabela 8 – Algoritmo kNN - Variação de N

KNN	1	2	3	4	5
CCI	64.20%	73.15%	72.70%	73.41%	73.93%
Kappa	0.0918	0.0639	0.0765	0.0594	0.0535

Fonte: elaborado pelo autor

Tabela 9 – Matriz de Confusão: Algorítimo kNN, Ano 2010, Cross-validation 3

a	b	c	d	e	<– Classificado como
3367	538	101	207	27	a = MuitoBaixo
549	162	30	63	4	b = Baixo
100	26	11	19	4	c = Alto
204	57	19	32	5	d = Medio
24	6	5	4	1	e = MuitoAlto

Fonte: elaborado pelo autor

A Tabela 10 condensa os resultados das melhores configurações obtidas no Experimento 1. Os valores apresentados sob as Classes são as medidas de *F-measure*. O algoritmo kNN apresenta um excelente desempenho quando o conjunto de teste é o mesmo conjunto do conjunto de treinamento. Este seria um ótimo resultado caso o objetivo da pesquisa fosse a detecção de anomalias, como descrito por Witten et al. (2011), aonde este tipo de teste é utilizado para detectar fraudes em cartões de crédito, por exemplo.

Tabela 10 – Algoritmos kNN e C4.5 aplicados no ano 2010

Algoritmo	*Folds*	CCI	*Kappa*	Muito Baixo	Baixo	Médio	Alto	Muito Alto
kNN	0	99,98%	0.9995	1	0,999	1	1	1
kNN	3	64,20%	0.0918	0,794	0,203	0,100	0,067	0,025
C4.5	0	85,43%	0.5427	0,917	0,564	0,590	0,586	0,370
C4.5	10	71,24%	0.0978	0,843	0,152	0,081	0,054	0

Fonte: elaborado pelo autor

4.2 Experimento II

O segundo experimento baseou-se em aplicar algoritmos escolhidos no Experimento I, ao conjunto de dados de 2000. As configurações que obtiveram melhor resultado estão demonstradas na Tabela 11. Dos testes realizados com a variação de *Folds* para o algoritmo kNN, nenhum deles obteve melhoria nos resultados. Por isso o valor padrão "dez"foi utilizado.

Tabela 11 – Algoritmos kNN e C4.5 aplicados no ano 2000

Algoritmo	*Folds*	CCI	*Kappa*	Muito Baixo	Baixo	Medio	Alto	Muito Alto
kNN	0	99.78%	0.9847	0,999	0,979	1	1	1
kNN	10	86,70%	0.0827	0,932	0,115	0,035	0	0
C4.5	0	92,45%	0	0,961	0	0	0	0
C4.5	10	91,84%	0.0204	0,958	0,019	0,043	0	0

Fonte: elaborado pelo autor

A matriz de confusão do algoritmo C4.5 (Tabela 13) para o ano de 2000 é muito diferente da matriz de 2010 (Tabela 9). Em 2010 existiram muito mais casos de dengue, os valores se encontram mais espalhados e apenas a categoria "Muito Alto"não obteve nenhuma classificação correta. No ano 2000, o resultado foi menos satisfatório. Nenhuma instância foi classificada corretamente nas categorias "Muio Alto"e "Alto". E apenas duas (20%) instâncias foram classificadas corretamente como "Médio". Estes valores podem ser confirmados através da estatística Kappa, com o valor obtido em 2010 sendo 4,79 vezes maior que o de 2000.

Os resultados obtidos pelo algoritmo kNN para o ano 2000 foram melhores que os obtidos pelo C4.5. O CCI mostra que menos instâncias foram classificadas corretamente, porém o valor de Kappa demonstra que o algoritmo foi capaz de classificar mais registros nas outras classes. Como visto o aumento dos TPs das classes "Baixo"e "Médio".

A realização do segundo experimento é importante para verificar se um determinado algoritmo é capaz de melhorar ou manter suas estatísticas independente do BD utilizado. Como observado nesta seção, os resultados foram diferentes. Para determinar qual algoritmo

Tabela 12 – Matriz de Confusão: Algorítimo C4.5, Ano 2000

a	b	c	d	e	<– Classificado como
5106	26	6	7	0	a = MuitoBaixo
285	3	1	0	0	b = Baixo
81	1	2	0	0	c = Medio
35	2	1	0	0	d = Alto
9	0	0	0	0	e = MuitoAlto

Fonte: elaborado pelo autor

é capaz de manter resultados melhores e mais constantes para as diversas situações, exigiria mais experimentos com Banco de Dados maiores.

Tabela 13 – Matriz de Confusão: Algorítimo kNN, Ano 2000

a	b	c	d	e	<– Classificado como
4788	251	71	31	4	a = MuitoBaixo
232	34	12	10	1	b = Baixo
71	9	3	0	1	c = Medio
31	6	1	0	0	d = Alto
6	3	0	0	0	e = MuitoAlto

Fonte: elaborado pelo autor

4.3 Discussão

Selecionar o melhor algoritmo de MD para uma determinada tarefa é uma atividade que envolve a realização de muitos testes. Existem inúmeros Algoritmos de Classificação, e cada um deles pode ser configurado de acordo com as necessidades da pesquisa. No primeiro experimento, tanto a Tabela 10 quanto a Tabela 11 mostram que os melhores resultados foram obtidos utilizando o conjunto de treinamento também como conjunto de teste (*folds* = 0). Esta informação deve ser analisada com cuidado, conforme descrito em 2.4. Como estamos interessados na classificação de novos registros, os valores obtidos utilizando a técnica de *Cross-validation* devem ser priorizadas. Desta forma, o algoritmo C4.5 utilizando 10 *folds* apresentou o melhor desempenho para o ano de 2010, mesmo não tendo sido capaz de classificar corretamente os registros em "Muito Alto"e "Alto".

Na análise do ano 2000, o algoritmo C4.5 não apresentou o mesmo desempenho, provavelmente devido a alta densidade de registros da classe "Muito Baixo". Mesmo alterando-se as configurações do algorítimo, o desempenho não obteve melhora. Por outro lado, o algoritmo kNN mesmo com um CCI menor, obteve melhor pontuação na estatística *Kappa*, representando uma melhor capacidade em classificar os registros. A Tabela 14 condensa os resultados obtidos nos dois anos.

Tabela 14 – Comparação dos resultados do Experimento 1 e 2

	2000		2010	
	CCI	Kappa	CCI	Kappa
C4.5	91,85%	0,0204	71,25%	0.0978
kNN	86,70%	0.0827	64,19%	0.0786

Fonte: elaborado pelo autor

5 Conclusão

A utilização de técnicas de Mineração de Dados para identificar padrões no comportamento e transmissão de doenças pode ser de grande utilidade para os órgãos públicos responsáveis no combate a doenças. Este trabalho alcançou o objetivo proposto e demonstrou a capacidade dos algoritmos de Classificação em desempenhar esta tarefa.

No primeiro momento, foram coletados dados relativos aos índices de Dengue e dados socioeconômicos e físicos sobre cada município brasileiro. Alguns fatores dificultaram a coleta de mais dados: pesquisas de nível municipal não são realizadas todos os anos, nem todos os municípios possuem dados climáticos históricos disponíveis e nem informações sobre ovitrampas, por exemplo. Os resultados dos algoritmos mostraram que estes fatores organizacionais e políticos interferem diretamente nos resultados.

O próximo passo envolveu o tratamento das informações coletas. Elas foram filtradas, padronizadas e organizadas para que pudessem ser adicionadas a um Banco de Dados. Foram escolhidos os anos de 2000 e 2010 devido a maior disponibilidade de dados. Este fator também teve grande impacto na pesquisa, pois reduziu a quantidade de dados disponíveis para análise. Como apontado em diversos momentos, para melhorar os resultados da tarefas de MD, é muito importante que se tenha um grande conjuntos de dados treinados.

Uma vez o Banco de Dados populado, foi possível iniciar as técnicas de Mineração de Dados. A primeira tarefa utilizou o algoritmo *K-means* para agrupar os registros de acordo com os seus atributos. Foram determinados cinco *clusters*: "Muito Baixo", "Baixo", "Médio", "Alto"e "Muito Alto". Estas informações também foram adicionadas ao Banco de Dados.

Em seguida, os dados foram submetidos aos algoritmos de classificação. Depois de vários testes, os melhores resultados obtidos foram através dos algoritmos kNN e C4.5. No ano 2000, com menos casos de Dengue, o algoritmo kNN conseguiu classificar corretamente 86,70% dos municípios, alcançando uma estatística Kappa de 0,0827. Para o ano de 2010, o algoritmo C4.5 obteve o valor de Kappa igual a 0,0978 e classificou corretamente 71,25% das instâncias. A diferença no padrão de distribuição das notificações de Dengue nos dois anos estudados causou oscilações nas métricas obtidas. Porém, com o passar do tempo, quando mais dados forem acumulados, a transição entre os dados de um ano para outro se tornará mais suave, maximizando os resultados dos algoritmos.

Os melhores valores de Kappa foram obtidos quando os conjuntos de treinamento foram utilizados para teste (0,9847 em 2000, e 0,9995 em 2010). Porém, para os testes aonde foi utilizada o *Cross-Validation*, para melhorar a confiança dos resultados, os valores de Kappa caíram para menos de 0,1. A valor de CCI por outro lado, foi importante para

mostrar que os algoritmos possuem potencial. Acreditamos que com a expansão do banco de dados, mais municípios poderão ser classificados nos grupos "Alto"e "Muito Alto", melhorando a qualidade dos *clusters* e consequentemente aumentando os valores de Kappa obtidos na Classificação.

5.1 Trabalhos Futuros

Algoritmos de Mineração de Dados dependem de grandes quantidades de registros com muitos atributos para obter melhores resultados. A utilização de mais atributos físicos, como os dados climáticos, poderá melhorar o desempenho dos algoritmos, uma vez que eles são importantes fatores para a reprodução do mosquito. Análises temporais não puderam ser realizadas devido ao tempo de desenvolvimento do trabalho. Elas seriam capazes de levar em consideração a ordem cronológica dos registros. Assim, seria possível até mesmo estimar os anos com mais chances de epidemias. O desenvolvimento de uma ferramenta capaz de receber os dados, tratá-los, e oferecer diferentes algoritmos para serem testados e comparados também seria de grande utilidade.

Referências

BRASIL, MINISTÉRIO DA SAÚDE. *Dengue instruções para pessoal de combate ao vetor: manual de normas técnicas*. 3. ed. Brasília: Ministério da saúde: Fundação Nacional de saúde, 2001. Citado na página 17.

BRASIL, MINISTÉRIO DA SAÚDE. *Diretrizes Nacionais para Prevenção e Controle de Epidemias de Dengue*. Brasília: Ministério da saúde, 2009. Citado na página 15.

CLEARY, J. G.; TRIGG, L. E. et al. K*: An instance-based learner using an entropic distance measure. In: *Proceedings of the 12th International Conference on Machine learning*. [S.l.: s.n.], 1995. v. 5, p. 108–114. Citado na página 22.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996. Citado 3 vezes nas páginas 17, 18 e 19.

FERRERO, C. A. *Algoritmo kNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia*. 2009. Citado 2 vezes nas páginas 20 e 21.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011. Citado na página 19.

PESSANHA, J. E. M. P. et al. Diffusion pattern and hotspot detection of dengue in belo horizonte, minas gerais, brazil. *Journal of tropical medicine*, Hindawi Publishing Corporation, v. 2012, 2012. Citado 2 vezes nas páginas 26 e 27.

PNUD, IPEA, FUNDAÇÃO JOÃO PINHEIRO. *O Índice de Desenvolvimento Humano Municipal brasileiro*. Brasília: PNUD, Ipea, FJP, 2013. Citado na página 30.

RORIZ-FILHO, L. D. e Sérgio Almeida e Tissiana Haes e Letícia Mota e J. Dengue: transmissão, aspectos clínicos, diagnóstico e tratamento. *Medicina (Ribeirao Preto. Online)*, v. 43, n. 2, p. 143–152, 2010. ISSN 2176-7262. Citado na página 31.

SALZBERG, S. L. C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, v. 16, n. 3, p. 235–240, 1994. ISSN 1573-0565. Disponível em: <http://dx.doi.org/10.1007/BF00993309>. Citado 2 vezes nas páginas 21 e 22.

SAÚDE, M. da. *Casos de Dengue. Brasil, Grandes Regiões e Unidades Federadas, 1990 a 2014**. [S.l.], 2015. Último acesso em 20 de Março, 2017. Disponível em: <https://goo.gl/FQr35G>. Citado na página 15.

SAÚDE, M. da. *Dados: casos e óbitos de Dengue - Pedido 25820002781201539*. 2015. Disponível em: <https://goo.gl/f2WxKn>. Citado na página 33.

SHAUKAT, K. et al. Dengue fever in perspective of clustering algorithms. *CoRR*, abs/1511.07353, 2015. Disponível em: <http://arxiv.org/abs/1511.07353>. Citado na página 27.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data mining*. [S.l.]: Addison Wesleys, 2005. Citado na página 20.

VIANNA, R. C. X. F. et al. Mineração de dados e características da mortalidade infantil data mining and characteristics of infant mortality. *Caderno saúde pública*, scielo, v. 26, p. 535 – 542, 03 2010. ISSN 0102-311X. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2010000300011&nrm=iso>. Citado 2 vezes nas páginas 25 e 26.

WITTEN, I. H. et al. *Data mining : practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2011. Citado 5 vezes nas páginas 20, 21, 23, 24 e 36.

WITTEN, I. H. et al. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques*. [S.l.]: Morgan Kaufmann, 2016. Citado na página 25.

WORLD HEALTH ORGANIZATION. *Dengue: guidelines for diagnosis, treatment, prevention and control- New edition*. [S.l.], 2009. Citado na página 15.