UNIVERSIDADE DO ESTADO DA BAHIA

DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA

CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

CALISON RIBEIRO DOS SANTOS

UTILIZAÇÃO DE MINERAÇÃO DE DADOS PARA A BUSCA DE PADRÕES DE ESPALHAMENTO GEOGRÁFICO DE GENÓTIPOS VIRAIS

SALVADOR - BAHIA

2017

CALISON RIBEIRO DOS SANTOS

UTILIZAÇÃO DE MINERAÇÃO DE DADOS PARA A BUSCA DE PADRÕES DE ESPALHAMENTO GEOGRÁFICO DE GENÓTIPOS VIRAIS

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia- UNEB, como requisito parcial à obtenção do grau de bacharel em Sistemas de Informação. Área de Concentração: Inteligência Artificial

Orientadora: Trícia Souto Santos

Co-Orientadora: Maria Inés Valderrama Restović

SALVADOR - BAHIA

2017

CALISON RIBEIRO DOS SANTOS

UTILIZAÇÃO DE MINERAÇÃO DE DADOS PARA A BUSCA DE PADRÕES DE ESPALHAMENTO GEOGRÁFICO DE GENÓTIPOS VIRAIS

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia- UNEB, como requisito parcial à obtenção do grau de bacharel em Sistemas de Informação. Área de Concentração: Inteligência Artificial

Aprovada em:

BANCA EXAMINADORA

Trícia Souto Santos (Orientadora)
Universidade do Estado da Bahia – UNEB

Maria Inés Valderrama Restović (Co-Orientadora)
Universidade Do Estado Da Bahia - UNEB

Membro da Banca Dois
Faculdade de Filosofia Dom Aureliano Matos – FAFIDAM
Universidade do Membro da Banca Dois - SIGLA

Membro da Banca Três
Centro de Ciências e Tecnologia - CCT
Universidade do Membro da Banca Três - SIGLA

RESUMO

Os arbovírus (arthropod-bone virus) compõem um extenso grupo de vírus zoonóticos que infectam mosquitos capazes de transmitir doenças para os seres humanos por meio de suas picadas. Os arbovírus são responsáveis por transmitir os vírus da Dengue, Zika e Chikungunya que se tornaram uma grande ameaça a saúde humana, visto que têm evoluído ao longo do tempo criando linhagens mais complexas, retardando o processo de identificar uma cura e transformando esse vírus em um problema em nível mundial. Com isso, o objetivo desta pesquisa será buscar a existência de características em uma região que possam determinar a maior incidência dos vírus da Dengue, Zika e Chikungunya, levando em consideração seus atributos genéticos. Para se tentar encontrar essas características uma base de dados com informações genéticas sobre os arbovírus foram criadas, esses dados foram extraídos do GenBank que consiste em uma base de dados pública contendo informações genéticas. A presente pesquisa busca utilizar mineração de dados juntamente com sua técnica de Clustering que irá buscar padrões capazes de determinar a maior incidência dos arbovírus levando em consideração as características climáticas e geográficas de uma região e as características genéticas dos arbovírus. Além disso será inserido mais informações na base de dados utilizando a técnica de scraping, extraindo essas informações do GenBank.

Palavras-chave: Arbovírus. Dengue. Zika. Chikungunya. Mineração de Dados. Clustering.

ABSTRACT

Arboviruses (arthropod-bone virus) make up an extensive group of zoonotic viruses that infect mosquitoes capable of transmitting diseases to humans through their bites. Arboviruses are responsible for transmitting Dengue, Zika and Chikungunya viruses that have become a major threat to human health, since they have evolved over time creating more complex lineages, delaying the process of identifying a cure and turning that virus into a World-wide problem. Therefore, the objective of this research will be to investigate the existence of characteristics in a region that can determine the highest incidence of Dengue, Zika and Chikungunya viruses, taking into account their genetic attributes. To attempt to find these characteristics a database with genetic information about arbovirus were created, this data was extracted from GenBank that consisting of a public database containing genetic information. The present research seeks to use data mining along with its Clustering technique that will search for patterns capable of determining the highest incidence of arboviruses taking into account the climatic and geographic characteristics of a region and the genetic characteristics of arboviruses. In addition, more information will be entered into the database using the scraping technique, extracting this information from GenBank

Keywords: Arbovirus. Dengue. Zika. Chikungunya. Data Mining. Clustering.

SUMÁRIO

1 INTRODUÇÃO

Os arbovírus (arthropod-borne virus) compõem um grupo grande de vírus zoonóticos que infectam artrópodes hematófagos (mosquitos ou parasitas que se alimentam de sangue) e são comumente transmitidos aos seres humanos, principalmente por meio da picada de mosquitos. Os arbovírus são responsáveis pela transmissão da Dengue, Zika e Chikungunya e representam uma grande ameaça à saúde humana. Mudanças climáticas, ambientais em conjunto com o desmatamento florestal, favorecem a disseminação e transmissão desses vírus nas regiões subtropicais.

Esses vírus têm evoluído através do tempo, progressivamente em linhagens mais complexas e virulentas, dificultando descobrir curas para cada linhagem. As implicações causadas pelos arbovírus são alarmantes sendo considerado um problema grave para a saúde pública em nível mundial (FIGUEIREDO, 2007).

Neste contexto nota-se que existem bases de dados públicas com informações sobre vírus. Uma dessas bases de dados pública é o GenBank que persiste informações genéticas (GENBANK, 2013). Neste cenário, surge a seguinte questão: baseada na relação entre as linhagens (genótipos) dos arbovírus é possível determinar quais dados climáticos ou geográficos são responsáveis pela maior ocorrência desses vírus em uma região?

Baseada nessa problemática a pesquisa tem como objetivo aplicar Mineração de Dados em uma base de dados que contém informações sobre os arbovírus. A Mineração de Dados (também conhecida em inglês como Data Mining) é utilizada para descobrir regras, padrões e fatores em informações armazenadas em uma base de dados para auxiliar na tomada de decisão (AGARWAL, 2013). Busca-se integrar nesse processo métodos de Mineração de dados que apóiem o objetivo dessa pesquisa. Portanto essa pesquisa propõe os seguintes objetivos específicos:

• Preencher uma base de dados com informações sobre os arbovírus, informações essas que serão extraídas do GenBank.

• Utilizar a ferramenta Weka e a funcionalidade de clustering para a mineração dos dados.

• Avaliar os resultados gerados pela ferramenta de mineração e concluir se existiram regiões em que foi possível determinar a maior incidência dos arbovírus.

Esta pesquisa está dividida em 6 capítulos, iniciando com a introdução, onde são abordados os conceitos gerais desta pesquisa. No capítulo 2 será contextualizado e explicado

sobre os arbovírus e suas derivações, como Dengue, Zika e Chikungunya. O capítulo 3 será abordado conceitos de mineração de dados e seus algoritmos. No capítulo 4 será informado sobre trabalhos relacionados para com a presente pesquisa. O capítulo 5 fala sobre o projeto, a metodologia e validações utilizadas para o desenvolvimento da pesquisa.

2 ARBOVÍRUS

Os arbovírus (arthropod-borne virus) são vírus RNA zoonóticos que pertencem a várias famílias de vírus e que infectam artrópodes hematófagos (mosquitos ou parasitas que se alimentam de sangue) que transmitem o vírus aos seres vertebrados através de picadas. Países com grande desmatamento e uma grande variedade de flora e fauna são ecossistemas que apresentam condições ideais para a existência de vários tipos de arbovírus. Os arbovírus infectam mamíferos, aves e serpentes. Os mosquitos infectados pelo vírus podem transmitir para seres humanos: dengue, zika, chikungunya, febre amarela e outras enfermidades (JAWETZ et al., 2010).

As enfermidades derivadas do arbovírus tem-se espalhado com rapidez ao redor do mundo e existem vários motivos para a proliferação como o enorme deslocamento de pessoas em todo o planeta, a destruição de ambientes naturais, aquecimento global e a alta taxa de evolução do genoma do arbovírus, facilitando a sua adaptação a vários hospedeiros e a expansão de seus vetores (FIGUEIREDO, 2007).

As implicações causadas pelos arbovírus são alarmantes sendo considerado um problema grave para a saúde pública em nível mundial (GONZáLEZ; RODAS, 2015) devido a sua grande expansão geográfica como é exibido na (Figura 1) onde exibe a distribuição geográfica do vírus da dengue, zika e chikungunya no mundo. Duas famílias do arbovírus tornam-se importantes nesse contexto mundial, as famílias Flarividade e Togaviridade por hospedarem os vírus da dengue, zika e chikungunya.

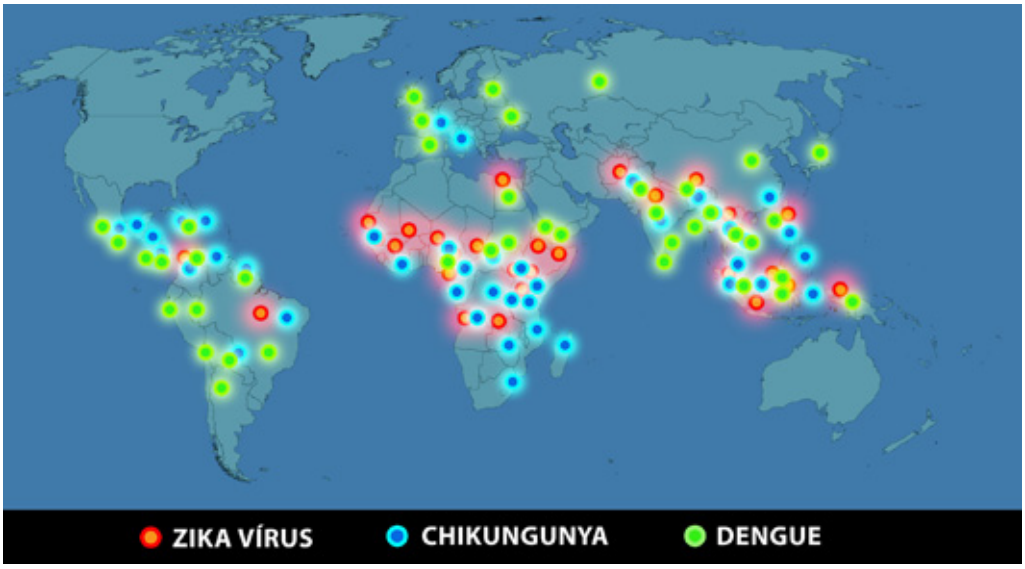

Figura 1 – Distribuição da presença dos vírus dengue, zika e chikungunya
Fonte – Centers for Disease Control and Prevention (CDC)

A Flaviridade do gênero flavivírus contém aproximadamente 70 vírus descobertos compostos de uma cadeia simples de ácido ribonucleico (RNA) de cadeia positiva. Alguns flavivírus são transmitidos entre os vertebrados por mosquitos e carrapatos enquanto outros por roedores e morcegos (JAWETZ et al., 2010).

A Togaviridade do gênero alphavirus contém aproximadamente 30 vírus descobertos compostos de uma cadeia simples de ácido ribonucleico (RNA) de cadeia positiva. Eles são transmitidos entre os vertebrados por artrópodes que se alimentam de sangue (JAWETZ et al., 2010).

2.1 VÍRUS DENGUE

A dengue é um dos principais problemas de saúde pública no mundo. A Organização Mundial da Saúde (OMS) estima que 80 milhões de pessoas se infectem anualmente, em 100 países, de todos os continentes, exceto a Europa. Cerca de 550 mil doentes necessitam de hospitalização e 20 mil morrem em consequência da dengue (SAÚDE, 2002).

O vírus dengue é um vírus do gênero dos flavivírus dentro da família Flaviridade e é classificado em quatro sorotipos (DENV-1, DENV-2, DENV-3 e DENV-4) e são predominantes em áreas tropicais e subtropicais. O número de países que relatam o crescimento da dengue tem aumentado rapidamente. A alta taxa de viagens dos seres humanos sem estarem devidamente vacinados contra a doença colabora com a expansão da dengue em outros habitats. Um outro fato

é a carência de planejamento urbano em algumas cidades, favorecendo a infecção em humanos (VASCONCELOS, 2010–2011).

Os sintomas mais comuns da Dengue são: febre (entre 39 a 40 graus), dores musculares, dores de cabeça e nos olhos. Anorexia, náuseas e vômitos também podem estar presentes, diarréia está presente em 48% dos casos. Além desses sintomas, a dengue também pode ser responsável por hemorragia, quando esse sintoma acontece, comumente é chamado de Dengue Hemorrágica.

Com o propósito de se entender as linhagens (genótipos) evolutivas da dengue uma análise filogenética feita por Rubing Chen foi efetuada para cada sorotipo do vírus da Dengue. O DENV Sorotipo 1 encontraram cinco genótipos (Genótipo I, Genótipo II, Genótipo III, Genótipo IV, Genótipo V). Para o DENV Sorotipo 2 encontram seis genótipos (Genótipo Asiático I, Genótipo Asiático II, Genótipo Sudeste Asiático-Americano, Genótipo Cosmopolita, Genótipo Americano, Genótipo Silvestre). No DENV Sorotipo 3 obteve cinco genótipos (Genótipo I, Genótipo II, Genótipo III, Genótipo IV, Genótipo V). O DENV Sorotipo 4 encontraram dois tipos de genótipos (Genótipo I, Genótipo IV-Silvestre) (CHEN; VASILAKIS, 2011).

2.2 VÍRUS ZIKA

O vírus zika é um arbovírus que pertence ao gênero dos flavivírus dentro da família Flaviridade assim como o vírus da dengue. O vírus possui esse nome porque foi primeiro identificado em macacos rhesus na floresta zika em Uganda em 1947, em humanos foi encontrado em 1954 no oeste da África (PETERSEN, 2016).

O vírus zika causa sintomas adversos como: febre, dores musculares, erupções cutâneas, dores de cabeça e nos olhos. Durante a sua proliferação na América 2015 suspeita-se que mulheres grávidas que foram infectadas pelo mosquito tiveram filhos nascidos com microcefalia (doença responsável pela má formação da cabeça) (YE, 2016). Enquanto no surto acontecido na França entre 2013-2015 houve um aumento no aparecimento de doenças de Síndrome de Guillain-Barré, onde o sistema imunológico ataca o próprio sistema nervoso, indicando que pudesse existir uma relação entre o vírus e a doença (YE, 2016).

Para se entender as linhagens (genótipos) evolutivas do zika uma análise filogenética feita por Qing Ye levando em consideração o gene E e o gene NS5, os resultados dessa análise

apresentaram duas linhagens do zika vírus, sendo elas a linhagem Africana e a linhagem Asiática (YE, 2016).

2.3 VÍRUS CHIKUNGUNYA

O vírus chikungunya é um RNA pertencente ao gênero alphavirus da família Togaviridade, e foi descoberto em 1952 em uma Ilha do Oceano Índico (FIGUEIREDO, 2007). Os primeiros casos de infecção em seres humanos ocorreram no início de 1770. O vírus só foi isolado do soro humano ou de mosquitos na epidemia na Tanzânia em 1952-53. Outras aparições do vírus ocorreram ocasionalmente, entretanto em 2004 um surto originário da costa da Quênia, dispersou-se pelas Ilhas Comoros, Réunion e outras Ilhas do Oceano Índico durante os dois anos seguintes. De 2004 a 2006, ocorreu um número estimado em 500 mil casos (SAÚDE, 2014).

O Aedes aegypti e o Aedes albopictus são os principais mosquitos responsáveis pela distribuição do vírus, com a diferença que o Aedes albopictus também está presente em latitudes mais temperadas. Assim como a dengue e o zika a transmissão do chikungunya se da principalmente em humanos. Em humanos picados pelo mosquito os sintomas de doença costumam aparecer em média em 3-7 dias.

Para se entender melhor a linhagens (genótipos) evolutivas do chikungunya uma análise filogenética foi efetuada por Marion Desdouits levando em consideração os genes E1 e E2. Os resultados apresentaram três genótipos sendo eles: Asia e Caribe, Centro Leste Sul Africano (ECSA) e África Ocidental (DESDOUITS, 2015).

O vírus da chikungunya pode causar doença aguda, subaguda e crônica.

A aguda é caracterizada por febre (usualmente acima de 39 graus) e fortes dores nas articulações. Outros sintomas comuns nesse estágio são: dores de cabeça, dores nas costas, dores musculares, náuseas, vômito, poliartrite, erupção cutânea e conjuntivite, essa fase costuma durar entre 3 a 10 dias (SAÚDE, 2014).

A Subaguda e crônica se inicia com uma recaída, isso costuma acontecer entre os dois e três meses após o início da doença. Alguns pacientes podem desenvolver distúrbios vasculares periféricos, como a síndrome de Raynaud (condição que afeta o fluxo sanguíneo nas extremidades do corpo humano, mãos e pés, assim como os dedos, nariz, lóbulos das orelhas), além de cansaço e fraqueza.

3 MINERAÇÃO DE DADOS

O mundo hoje consome zettabyte de informações por dia, e essa grande quantidade de informações começou a chamar atenção da indústria e da sociedade visando transformar essa grande quantidade de informações "sem utilidades"em conhecimento. Obter conhecimento a partir dessas informações é útil para análises do negócio pessoal, tomadas de decisão e outras aplicações. Nesse cenário surgiu a mineração de dados (HAN; KAMBER, 2006).

Mineração de dados (ou Data Mining do inglês) refere-se a extração ou mineração de informações advindas de uma base de dados, muitas vezes dados redundantes e quando analisados são capazes de identificar padrões, que podem auxiliar na tomada de decisão de um determinado negócio ou aplicação (AGARWAL, 2013) (HAN; KAMBER, 2006).

A base de dados que iremos analisar para aplicação de mineração de dados está construída sobre um modelo relacional por ser o modelo utilizado neste trabalho. O banco de dado relacional (RDB- Relational Database), consiste em um conjunto de dados relacionados conhecido como base de dados (ou em inglês database). A base de dados relacional é uma coleção de tabelas onde possuem nomes únicos e cada tabela possui atributos (colunas ou campos) e são armazenadas em tuplas (registros ou linhas) (HAN; KAMBER, 2006).

Um outro ponto-chave dentro da mineração de dados é saber a quais resultados deseja alcançar, uma vez que se conheça quais resultados desejam ser alcançados é preciso definir qual técnica deve ser aplicada. Através do tipo de dado armazenado pode-se definir que tipo de padrões ou relações desejamos minerar. A funcionalidade na mineração de dados irá especificar os padrões ou relacionamentos entre os registros e suas variáveis (CORTêS et al., 2002).

Existem várias funcionalidades dentro da mineração de dados, e podem ser separadas em dois tipos, Análise Descritiva e Análise de Prognóstico (HAN; KAMBER, 2006). Como podemos ver na Figura 2 ilustra que existem funcionalidades descritivas e de prognósticos para os usuários que desejam utilizar a mineração de dados e dentro das funcionalidades Descritiva e de Prognósticos existem as suas categorizações.

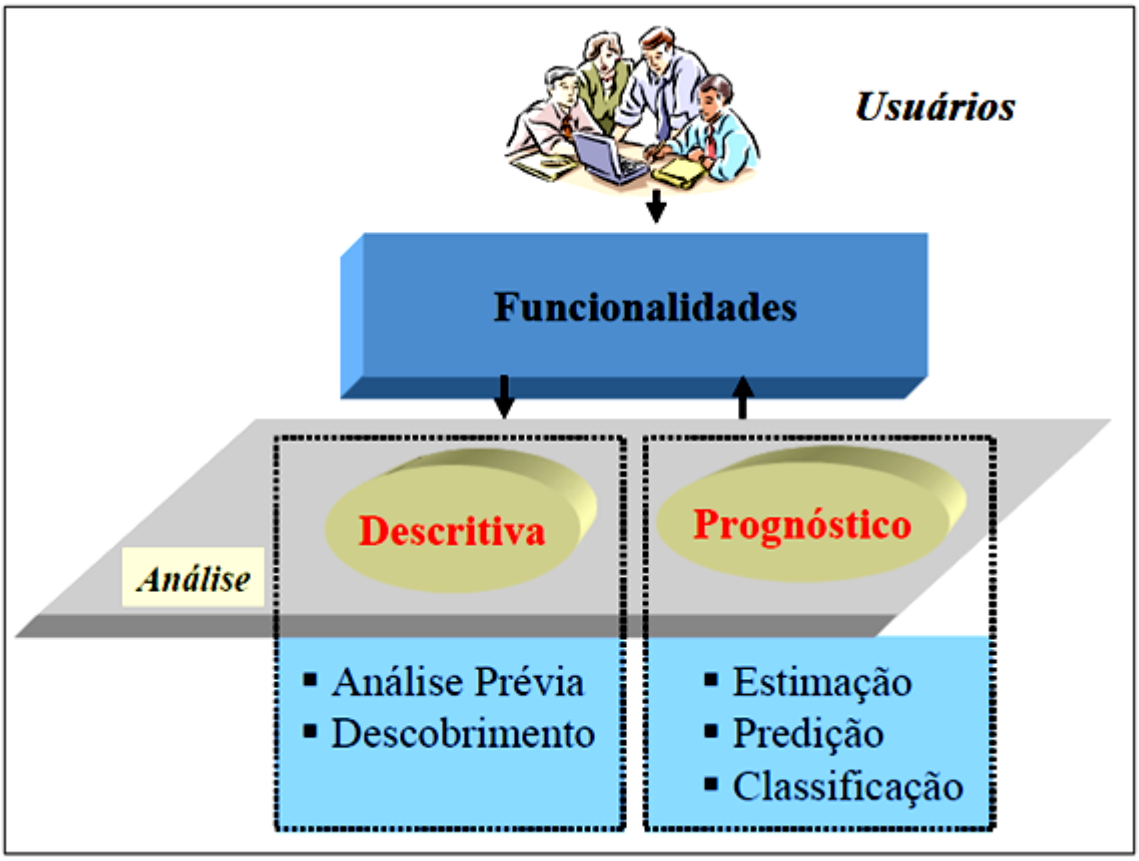

Figura 2 – Funcionalidades em mineração de dados
Fonte – Retirado de (CORTêS et al., 2002)

A Análise Descritiva representa a área de investigação dos dados e busca descrever quais fatos são relevantes e desconhecidos. A análise descritiva pode ser divida em Análise Prévia que consiste em analisar uma base de dados com o objetivo de encontrar anomalias ou seja, valores fora do padrão e a Análise de Descobrimento que tenta encontrar padrões que são desconhecidos pelo usuário como é apresentado na Figura 3.

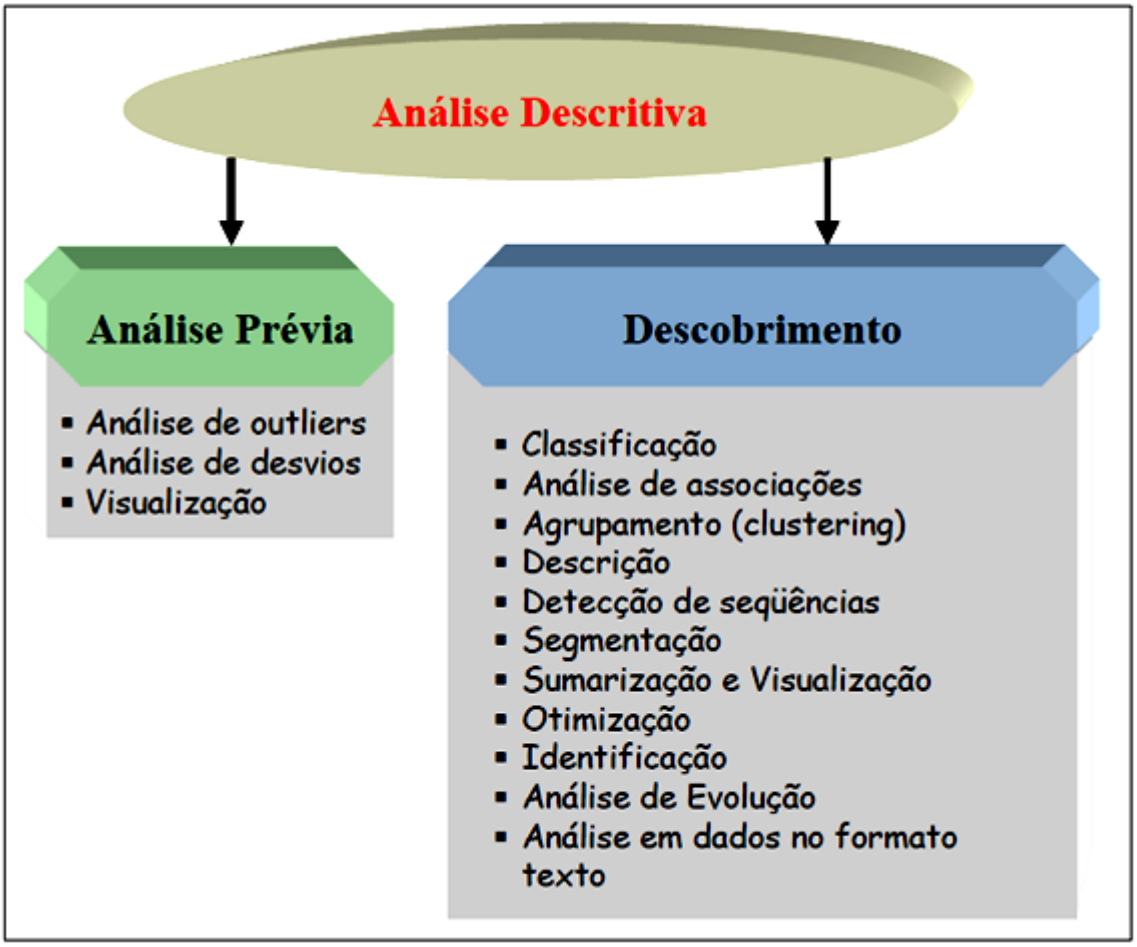

Figura 3 – Sub-funcionalidades da análise prévia e do descobrimento
Fonte – Retirado de (CORTêS et al., 2002)

Dentro da Análise Prévia existem três sub-funcionalidades.

• Análise de Outliers: Buscam encontrar dados que não obedecem ao comportamento do modelo de dados, por isso também é chamada de Detecção de desvios. É uma avaliação probabilística de desvios ou riscos associados aos objetos definidos no início da mineração.

• Análise de desvio: Tem como objetivo identificar que exibe mudanças de comportamento. É uma análise muito utilizada para fraudes de cartões de crédito, uma vez que se possua o padrão de compra de clientes é possível explorar se existe um desvio nesse padrão de compras.

• Visualização: Tem como objetivo encontrar disparidades nos dados uma vez que não se possui conhecimento prévio dos dados.

A outra funcionalidade dentro da Análise Descritiva é a de Descobrimento que possui onze tipos de sub-funcionalidades.

• Classificação: A classificação busca examinar as características dos dados e atribuir classes para esses dados. Esses dados podem ser associados a classes ou a conceitos.

• Análise de associações: Essa sub-funcionalidade investiga quais características estão relacionadas, ou seja, observa regras de associações condicionadas a valores de atributos que ocorrem juntos em conjunto de dados.

• Agrupamento (clustering): Visa segmentar os dados em grupos homogêneos. Essa técnica organiza os dados para formarem grupos o mais similar possível entre si, e que os dados dentro dos seus grupos sejam os mais diverso possível.

• Descrição: Tenta tornar mais clara hipóteses ou fatos que estejam sendo examinados no modelo de dados.

• Detecção de sequências: Essa sub-funcionalidade infere padrões nos dados e através desses padrões tenta determinar os tipos de sequências que desejam ser obtidos.

• Segmentação: Como indica o nome da sub-funcionalidade, os conjuntos de dados são desmembrados em subconjunto de dados menores. E com esses dados podem ser determinados novos agrupamentos (clustering).

• Sumarização e Visualização: A mineração de dados busca oferecer os resultados de forma a ser interpretados pelos seus usuários finais, e essa é uma das características dessa sub-funcionalidade, busca exibir através de uma visualização gráfica a análise de seus dados sumarizados.

• Otimização: Averígua otimizar recursos limitados como tempo, espaço, dinheiro, matéria-prima entre outros. O intuito dessa sub-funcionalidade é maximizar vendas, lucros e distribuição.

• Identificação: Tenta reconhecer um item, evento ou atividade no conjunto de dados inferindo padrões sobre essa base de dados.

• Análise de Evolução: Faz uma análise do modelo, procurando por regularidades ou tendências de comportamentos que mudam ao longo do tempo.

• Análise em dados no formato texto: É utilizada para análise de dados armazenados em formato texto como: narrativas, processos judiciais entre outros, aspirando mudar esses textos

para formas que possam ser extraídos usando técnicas de tratamento e exploração de textos.

Em paralelo a funcionalidade de Análise descritiva e suas sub-funcionalidades (análise prévia e descobrimento) vamos explanar sobre a funcionalidade de Análise de Prognóstico, que se refere a encontrar dados a partir de padrões achados pela Análise descritiva (CORTêS et al., 2002). Existem três categorizações para essa funcionalidade:

• Estimação: Tenta predizer algum valor fundado em algum padrão já conhecido.

• Predição: É o processo de predizer um comportamento futuro baseado em dados previamente estabelecidos.

• Classificação: Prediz algum valor para variáveis categóricas.

3.1 AGRUPAMENTO (CLUSTERING)

Neste capítulo será abordado mais sobre a técnica de Agrupamento ou Clustering, por ser uma técnica utilizada quando não se possui conhecimento prévio sobre os dados, característica presente na atual pesquisa. Essa técnica é comumente utilizada em largas bases de dados por atribuir classificações para objetos que podem possuir um grande custo de processamento (HAN; KAMBER, 2006) (CORTêS et al., 2002).

O clustering é o processo em agrupar classes ou grupos, esse agrupamento tenta associar dados com maiores similaridades de comparação dentro de seu grupo mas possuem uma grande disparidade de comparação com objetos de grupos distintos. Para tornar o processo de clustering automático podemos detectar dados em regiões compactas e dispersas em objetos e descobrir distribuições de padrões globais e correlacionar interesses (HAN; KAMBER, 2006) (CORTêS et al., 2002).

Clustering é uma técnica que está presente em várias áreas como: mineração de dados, biologia e aprendizado de máquina. Na área da biologia clustering é comumente utilizado para, definir os grupos plantas e animais, baseado em suas características comuns e atribui nomes a esses grupos, categorizar genes com funcionalidades similares e obtém conhecimento sobre a herança da população (HAN; KAMBER, 2006).

Clustering podem ser denominados como segmentação de dados em algumas aplicações, pois, particiona grande quantidade de dados em grupos de acordo com suas similaridades.

Clustering também podem ser utilizados para análise de dados fora do padrão determinado e que em alguns casos demonstram mais relevância para os objetivos esperados do que os dados "comuns".

3.2 DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS (KDD)

Existem passos essenciais para esse processo conhecido como descoberta de conhecimento ou em inglês como Knowledge Discovery from Data (KDD).

• Limpeza de dados, elimina dados inconsistentes.

• Integração dos dados, onde fontes de dados podem ser combinadas.

• Seleção de dados, etapa onde os dados relevantes para a análise são recuperados para a base de dados.

• Transformação dos dados, passo onde os dados são transformados de forma que possam ser minerados

• Mineração dos dados, passo essencial no processo, onde são utilizados métodos e técnicas para a extração de padrões dos dados

• Avaliações de Padrões, identifica o padrão que representa o conhecimento baseado nos critérios definidos

• Apresentação do conhecimento, etapa final onde é apresentado o resultado da análise ao usuário

A primeira etapa da mineração é a etapa da limpeza e integração dos dados, essa etapa comumente é necessária quando possuímos as informações advindas de várias bases de dados, então precisamos fazer uma limpeza dessas informações e a integração para podermos concentrá-las em um único repositório o nosso data wharehouse. Quando essas informações já se encontram em um data wharehouse podemos seguir para a etapa de seleção e transformação. Essa etapa consiste em agrupar dados repetidos, remover possíveis erros, generalizar e normalizar as informações. Após a preparação dos dados, eles são submetidos a alguma técnica de mineração. Após serem minerados tenta-se identificar padrões ou tomar decisões como é ilustrado na figura 2 (HAN; KAMBER, 2006).

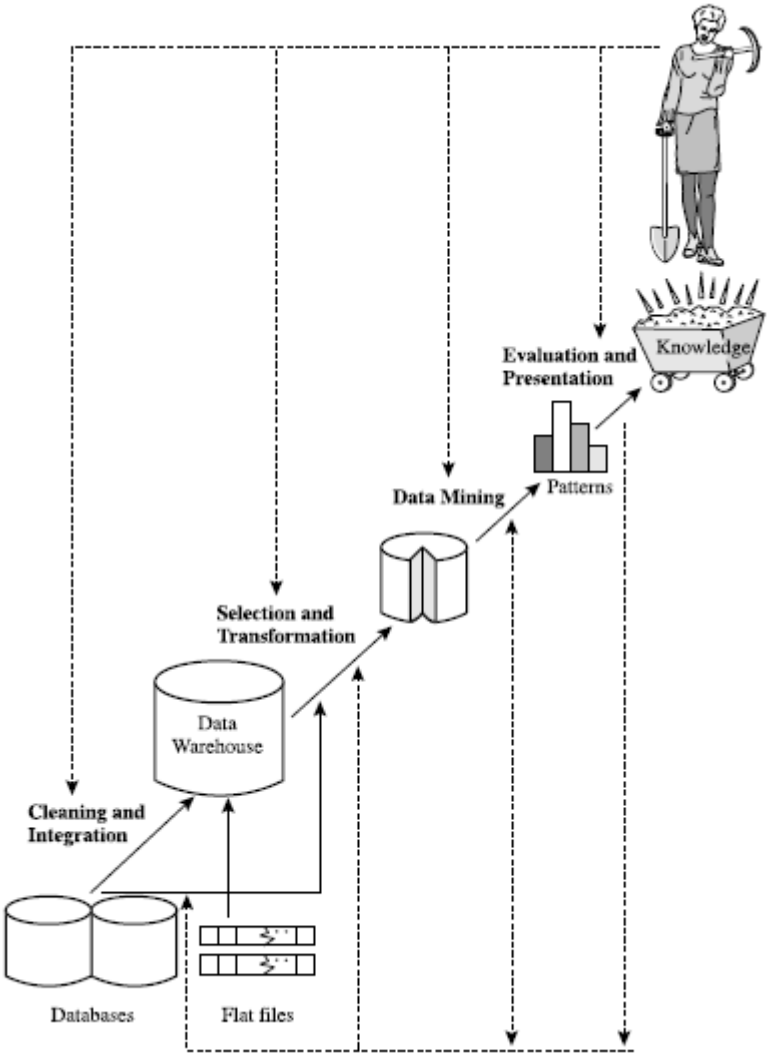

Figura 4 – Mineração de dados como passo no processo de descobrimento de conhecimento
Fonte – Retirado de (HAN; KAMBER, 2006)

4 TRABALHOS RELACIONADOS

Na busca de possíveis soluções para o problema de pesquisa desse trabalho foi efetuada uma pesquisa que visa encontrar trabalhos relacionados com mineração de dados e análise da incidência de vírus levando em consideração características de uma região.

Um dos trabalhos relacionados é a base para o desenvolvimento desse trabalho e é referente a construção de um banco de dados de biologia molecular com informações sobre os vírus dengue, zika e chikungunya (RESTOVIć, 2016).

Além do trabalho relacionado sobre os vírus também foram feitos pesquisas que pudessem auxiliar no entendimento do uso prático da mineração de dados usando a técnica de clustering visto que será a técnica utilizada para esse problema de pesquisa. Também foram realizadas buscar por trabalhos que tentassem correlacionar de alguma maneira a incidência dos vírus dengue, zika e chikungunya com propriedades de uma região.

O trabalho Banco de dados de biologia molecular para os vírus da dengue, zika e chikungunya com integração de tecnologias para filogenia e mineração de dados (RESTOVIć, 2016), busca construir uma base de dados com informações genéticas sobre os arbovírus dengue, zika e chikungunya e as suas linhagens (genótipos). As informações inseridas nessa base de dados foram retiradas do site Genbank e as informações contidas em seus artigos (GENBANK, 2013). A pesquisa também faz um trabalho extenso em documentar quais as linhagens existentes entre os três vírus dengue, zika e chikungunya derivados do arbovírus. O trabalho é essencial para pesquisa visto que a presente pesquisa irá utilizar a base de dados construída nesse trabalho.

O artigo Applied Biological Data Mining Based on Improved K-means Clustering Algorithm And KNN Classifier in Protein Sub-cellular Localization (LEI; WANG, 2016) utiliza o Clustering juntamente com o seu algoritmo K-means para a classificação de proteínas celulares. O artigo utiliza o algoritmo de clustering visando analisar grande quantidade de sequências genômicas em humanos de forma rápida onde ele utilizou o clustering para realizar a mineração desses dados e o KNN classifier para realizar a classificação desses dados após minerados. A pesquisa conseguiu comprovar que a utilização de clustering juntamente com o KNN classifier foi factível para a análise sequencial dos genomas humanos. Esse trabalho se aproxima da presente pesquisa a medida que utiliza a técnica de clustering para classificar proteínas celulares, mas se distância da presente pesquisa por não obter relações com arbovírus.

O trabalho de Review of Data Mining Clustering Techniques to Analyze Data with High Dimensionality as Applied in Gene Expression Data (AOUF et al., 2008) utiliza a técnica de clustering para o diagnóstico de tumores. A pesquisa tenta através de genes identificar técnicas que possam diagnosticar corretamente um tumor através de uma grande quantidade de dados. Esse trabalho se aproxima da presente pesquisa a medida que utiliza a técnica de clustering para tentar encontrar informações que possam ser determinantes no diagnóstico de tumores, mas se distância da presente pesquisa por não obter relações com arbovírus.

Os artigos Predictive model of dengue focus applied to Geographic Information Systems (GONZáLEZ; RODAS, 2015) e Mapping of Dengue Outbreak Distribution Using Spatial Statistics and Geographical Information System (LATIF; MOHAMAD, 2015) são artigos mais próximos do trabalho de pesquisa proposto, eles fazem análises de um determinado problema gerado pelo mosquito da Dengue e tenta identificar onde esse problema irá ocorrer geograficamente. Um artigo utiliza modelos matemáticos para tentar prever focos de mosquitos da dengue geograficamente pela América sem citar nenhuma relação entre arbovírus ou mineração de dados (GONZáLEZ; RODAS, 2015) enquanto o outro artigo tenta prever quais as chances de um surto que está ocorrendo em uma determinada região tem de se espalhar para outras regiões, porém igualmente ao primeiro artigo não faz relação com os arbovírus ou utiliza mineração de dados(LATIF; MOHAMAD, 2015).

5 MINERAÇÃO DE DADOS PARA BUSCA DE PADRÕES VIRAIS

Esta pesquisa será dividida em duas etapas, a primeira etapa consiste popular a base de dados com mais informações sobre os arbovírus dengue, zika e chikungunya e adicionar características climáticas e geográficas das regiões onde esses vírus se encontram. Essas informações serão retiradas do Genbank.

O GenBank é uma base de dados online mantida pela NCBI (National Center for Biotechnology Information) e contém informações genéticas sobre vírus. Existem duas possibilidades para se ter acesso as informações contidas no Genbank a primeira possibilidade é pagar um valor para a NCBI e ter acesso a todos os artigos científicos disponibilizados acessando de qualquer local. A segunda possibilidade é estar acessando de uma rede acadêmica que possua acesso liberado as informações do Genbank. (GENBANK, 2013)

A segunda etapa do projeto será utilizar mineração de dados para buscar insumos que determinem a maior incidência dos arbovírus em alguma região levando em consideração a relação entre as características climáticas e geográficas de uma região e as características genéticas dos arbovírus.

Este capítulo será dividido em cinco sessões contando com esta. Esta sessão dará uma visão geral sobre o problema de pesquisa proposto. A segunda sessão será irá falar sobre as ferramentas, onde serão apresentadas as duas ferramentas para a solução do problema de pesquisa, as duas ferramentas serão: Web Scraping e Weka. A terceira sessão será referente a metodologia abordado nesse trabalho para o objetivo esperado da presente pesquisa. A quarta sessão irá abordar a validação, que irá explicar qual a forma que pretende-se validar os futuros resultados da pesquisa. A quinta e última sessão será referente ao cronograma onde será mostrado cronologicamente o desenvolvimento do trabalho.

5.1 FERRAMENTAS

5.1.1 Weka

Para a execução do processo de mineração de dados foi escolhido a ferramenta Weka, por quatro motivos, a ferramenta é gratuita, possui a técnica de mineração de dados Clustering e é integrada com Java, uma linguagem de conhecimento do autor da presente pesquisa e possui

uma documentação extensa gratuita (WEKA, 2017).

O Weka é uma ferramenta gratuita que consiste em uma compilação de algoritmos de aprendizado de máquina para problemas que envolvem a mineração de dados. Os algoritmos podem ser aplicados diretamente a sua base de dados, ou ser chamado através de um código Java. Essa ferramenta possui funcionalidades como: pré-processamento, classificação, regressão, clustering e visualização.

Existem duas formas de se utilizar o Weka, diretamente em sua IDE que possui inúmeras opções como, realizar conexão a sua base de dados, carregar sua base de dados através de um .CSV, .JSON, ou um .arff (formato específico do Weka). Se decidido carregar sua base de dados a partir de um .CSV ou um .JSON existem alguns campos que devem ser preenchidos para que o Weka reconheça o seu arquivo, como é informado na documentação do Weka (WEKA, 2017). Após escolhido a forma de onde você deseja minerar os seus dados, deve-se escolher através de qual funcionalidade deseja minerar os dados, dentre as funcionalidades já informadas acima que ele disponibiliza, a IDE irá executar o processo de mineração e irá exibir os resultados baseado nos atributos escolhidos a serem minerados e correlacionados.

A segunda forma de se utilizar o Weka para a mineração de dados, é através de um código em Java, metodologia essa que será adotada no presente trabalho. Através do código Java é possível fazer as mesmas funções feitas diretamente na IDE, a escolha de se utilizar o código Java é a facilidade para se manipular, ou modificar informações se necessário.

O Weka disponibiliza uma API que pode ser importada e utilizada em Java como informada em sua documentação oficial (JAVA, 2017). Assim como na IDE é possível realizar o processo de mineração de dados conectando diretamente a sua base de dados através da biblioteca do JDBC, ou converter sua base de dados em um arquivo .CSV, .JSON ou .arff.

Para utilizar a funcionalidade de clustering, será necessário importar a biblioteca de clustering do weka e construir um clustering em código Java. A construção de um clustering usanddo Java é simples, deve-se criar uma classe denominada "EM", informar as opções que consiste em o máximo de iterações a serem executadas e a quantidade de linhas da base de dados a serem lidas e executar o método "buildClusterer"disponível na classe "EM"(JAVA, 2017).

5.1.2 Web Scraping

O Web Scrapping, Web Scrapper ou simplesmente scrapping (do inglês raspagem de dados) como é comumente utilizado, é uma técnica computacional para a extração de dados de uma página em um website evitando o trabalho tedioso de copiar e colar informações que podem ser extraídas automaticamente. e convertidas normalmente em um formato CSV, XML, JSON ou MySql.

Algumas linguagens como PHP, Python e Java proporcionam o serviço de scrapping nativa da linguagem ou utilizando bibliotecas criadas para a linguagem. Para a presente pesquisa iremos utilizar a linguagem PHP por ser uma linguagem de maior domínio do desenvolvedor deste trabalho.

Em PHP existem duas bibliotecas que proporcionam a utilização do scrapping, são elas: Guzzle e Crawler.

5.1.3 Guzzle

É um PHP cliente HTTP que permite realizar requisições para os servidores web exatamente como o seu browser. Essas requisições seguem um padrão de envio para o servidor web, e possuem alguns atributos obrigatórios e opcionais:

• Método de requisição GET ou POST, é um parâmetro obrigatório.

• URL de onde deseja extrair as informações, é um parâmetro obrigatório.

• Headers, é um parâmetro opcional, alguns sites necessitam de uma validação do cliente que está tentando acessar o servidor, em casos como esse são necessários a utilização do parâmetro de headers

• Form Params, são parâmetros opcionais utilizados quando se faz uma requisição do tipo POST, requisições desse tipo são usualmente utilizadas para formulários.

Após feita a requisição para o servidor web, ele irá validar a sua requisição, sendo uma requisição válida ele irá retornar a página HTML.

5.1.4 Crawler

O Crawler é a segunda etapa do processo, onde após a extração da página HTML, ela será consumida pelo Crawler que irá transformar essa informação em uma árvore Dom Document. O Dom Document transforma toda a sua página HTML em nós, esses nós podem ter outros nós filhos e assim sucessivamente como em uma estrutura de árvore. Essa estrutura permite "caminhar"pelo HTML, acessando suas informações através de uma tag HTML ou classes.

Após obter as informações, elas normalmente são transformadas em XML, CSV, JSON ou salvas em alguma base de dados.

5.2 METODOLOGIA

O desenvolvimento do modelo proposto será dividido em quatro etapas: popular a base de dados com mais informações sobre os arbovírus, popular a base de dados com mais informações climáticas e geográficas sobre os locais onde os genótipos serão analisados, minerar os dados na base de dados através do Weka, analisar as informações mineradas para buscar padrões que justifiquem a maior incidência de um genótipo dos arbovírus Dengue, Zika e Chikungunya em uma região.

A primeira etapa será adicionada mais informações sobre os arbovírus na base de dados já existente, essas informações serão extraídas do GenBank através da técnica de Web Scraping.

Na segunda etapa será necessário levantar informações climáticas e geográficas sobre as regiões existentes na base de dados, essas informações também serão coletadas no GenBank através da técnica de Web Scraping.

A terceira etapa irá utilizar a ferramenta do Weka e sua funcionalidade de clustering para minerar as informações relevantes contidas na base de dados para alcançar o objetivo da pesquisa.

A quarta etapa se refere a tomada de decisão, após os dados estarem minerados, essa etapa irá analisar se a mineração de dados foi qualificada para apresentar padrões capazes de inferir que características climáticas ou geográficas são responsáveis pela aparição de um vírus

em uma região.

5.2.1 Validação

Ao final da pesquisa espera-se obter os dados de cada região analisada agrupada pelos arbovírus dengue, zika e chikungunya e informar se foi possível determinar se existem características climáticas ou geográficas responsáveis pela maior incidência desses vírus em uma região. Esses resultados deverão ser apresentados em uma tabela, informando a região, o vírus, a linhagem do vírus, características regionais analisadas e o resultado final, informando se foi possível ou não determinar as padrões para aquela região. Caso existam situações em que foi possível determinar esses padrões, devem ser exibidos o percurso para se chegar a essa conclusão.

5.3 CRONOGRAMA

O cronograma provisório deste projeto é apresentado na tabela 1

	Mar	Abr	Mai	Jun	Jul	Ago	Set	Oct	Nov
Refinar os trabalhos relacionados levantados	•	•	•						
Coletar mais informações sobre os arbovírus				•	•				
Coletar mais informações sociais sobre as regiões				•	•				
Estudar detalhadamente Weka				•	•				
Implementar a mineração de dados em cima da base de dados através do Weka						•	•		
Avaliar resultados produzidos pelo Weka								•	
Apresentação da Monografia									•

Tabela 1 – Cronograma Março de 2017 até Novembro de 2017

6 CONCLUSÕES

Esta pesquisa se iniciou com uma revisão sistemática sobre a utilização de aprendizado de máquina para a busca de padrões que pudessem relacionar as características climáticas e geográficas de uma região e as características genéticas dos arbovírus. A revisão sistemática elucidou que o melhor caminho a ser tomado era a utilização de mineração de dados apenas, ao invés de inserir aprendizado de máquina.

Após esse processo de decisão de qual a melhor tecnologia a ser utilizada, foi necessário realizar um estudo mais avançado para o entendimento de mineração de dados e a técnica de scraping para a inserção dos dados. Esse estudo apresentou a técnica de Clustering como a melhor opção para o problema de pesquisa, pois, não se possui conhecimento prévio dos dados a serem trabalhados. Depois de decidido a técnica, foi necessário realizar uma busca por ferramentas gratuitas de mineração de dados que possam ser capazes de realizar a técnica de Clustering.

Foi necessário realizar um estudo de quais ferramentas ou técnicas computacionais poderiam auxiliar na extração dos dados para preencher com mais informações a base de dados, e qual ferramenta irá auxiliar na mineração de dados utilizando a técnica de Clustering. Esse estudo foi necessário para apresentar a técnica computacional de scraping que irá ser útil para a extração de mais informações genéticas sobre os vírus e climáticas e geográficas sobre as regiões e apresentou a ferramenta gratuita Weka, que será utilizada no processo de mineração de dados.

A pesquisa foi importante para a elaboração de um problema de pesquisa, e a descoberta e o estudo de possíveis ferramentas e técnicas computacionais capazes de solucionar o problema de pesquisa proposto.

REFERÊNCIAS

AGARWAL, S. Data mining. 2013.

AOUF; LYANAGE; HANSEN. Review of data mining clustering techniques to analyze data with high dimensionality as applied in gene expression data. 2008.

CHEN, R.; VASILAKIS, N. Quo tu et quo vadis? 2011.

CORTêS, S. da C.; PORCARO, R. M.; LIFSCHITZ, S. Mineração de dados - funcionalidades, técnicas e abordagens. 2002.

DESDOUITS, M. Genetic characterization of chikungunya virus in the central african republic. 2015.

FIGUEIREDO, L. T. M. Emergent arboviruses in brazil. 2007.

GENBANK. **GenBank**. 2013. Disponível em: <https://www.ncbi.nlm.nih.gov/genbank/>.

GONZáLEZ, M. B.; RODAS, G. G. Predictive model of dengue focus applied to geographic information systems. 2015.

HAN, J.; KAMBER, M. **Data Mining - Concepts and Thechniques**. 2. ed. San Francisco: Diane Cerra, 2006.

JAVA, W. **Weka Java**. 2017. Disponível em: <https://weka.wikispaces.com/Use+WEKA+in+your+Java+code>.

JAWETZ; MELNICK; ADELBERG. **Microbiologia Médica**. 22. ed. Rio de janeiro: McGraw-Hill, 2010.

LATIF, Z. A.; MOHAMAD, M. H. Mapping of dengue outbreak distribution using spatial statistics and geographical information system. 2015.

LEI, Z.; WANG, S. fang. Applied biological data mining based on improved k-means clustering algorithm and knn classifier in protein sub-cellular localization. 2016.

PETERSEN, E. Rapid spread of zika virus in the americas - implications for public health preparedness for mass gatherings at the 2016 brazil olympic games. 2016.

RESTOVIć, M. I. V. Banco de dados de biologia molecular para os vírus da dengue, zika e chikungunya com integração de tecnologias para filogenia e mineração de dados. 2016.

SAÚDE, M. D. **Programa Nacional de Controle a Dengue**. 2002. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/pncd_2002.pdf>.

SAÚDE, M. D. **Preparação e Resposta à Introdução do Vírus Chikungunya no Brasil**. 2014. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/preparacao_resposta_virus_chikungunya_brasil.pdf>.

VASCONCELOS, H. B. Phylogeography of dengue virus serotype 4. 2010–2011.

WEKA. **Weka**. 2017. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>.

YE, Q. Genomic characterization and phylogenetic analysis of zika viruscirculating in the americas. 2016.