UNIVERSIDADE DO ESTADO DA BAHIA

DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA

CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

FILIPE BOMFIM SANTOS FURTADO

ANÁLISE DA EPIDEMIOLOGIA MOLECULAR DO VÍRUS CHIKUNGUNYA RELACIONANDO VARIÁVEIS SOCIOECONÔMICAS E SINTOMATOLÓGICAS UTILIZANDO O K-PROTOTYPE

SALVADOR

2021

FILIPE BOMFIM SANTOS FURTADO

ANÁLISE DA EPIDEMIOLOGIA MOLECULAR DO VÍRUS CHIKUNGUNYA RELACIONANDO VARIÁVEIS SOCIOECONÔMICAS E SINTOMATOLÓGICAS UTILIZANDO O K-PROTOTYPE

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 final à obtenção do grau de bacharel em Sistemas de Informação.
Área de Concentração: Sistemas de Informação

Orientadora: Profa. Dra. Maria Inês Valderrama Restovic

SALVADOR

2021

FICHA CATALOGRÁFICA
Sistema de Bibliotecas da UNEB

F992a Furtado, Filipe Bomfim Santos

Análise da epidemiologia molecular do vírus chikungunya relacionando variáveis socioeconômicas e sintomatológicas utilizando o K-prototype / Filipe Bomfim Santos Furtado. - Salvador, 2021.
93 fls : il.

Orientador(a): Profa. Dra. Maria Inês Valderrama Restovic.
Inclui Referências
TCC (Graduação - Sistemas de Informação) - Universidade do Estado da Bahia. Departamento de Ciências Exatas e da Terra. Campus I. 2021.

1.Clustering. 2.Data Science. 3.Epidemiologia Molecular. 4.Chikungunya.

CDD: 604

FILIPE BOMFIM SANTOS FURTADO

ANÁLISE DA EPIDEMIOLOGIA MOLECULAR DO VÍRUS CHIKUNGUNYA RELACIONANDO VARIÁVEIS SOCIOECONÔMICAS E SINTOMATOLÓGICAS UTILIZANDO O K-PROTOTYPE

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 final à obtenção do grau de bacharel em Sistemas de Informação. Área de Concentração: Sistemas de Informação

Aprovada em: 06 de Julho de 2021

BANCA EXAMINADORA

Profa. Dra. Maria Inês Valderrama Restovic (Orientadora)
Universidade do Estado da Bahia – UNEB

Prof. Dr. Diego Gervásio Frias Suarez
Universidade do Estado da Bahia - UNEB

Prof. Dr. Ernesto de Souza Massa Neto
Universidade do Estado da Bahia - UNEB

AGRADECIMENTOS

Primeiramente, agradeço a Deus por ter me dado coragem, força, determinação e persistência para concluir este trabalho em um período tão turbulento das nossas vidas, em meio a uma pandemia.

Posteriormente, agradeço a meus pais, pelo cuidado, incentivo, acolhimento e torcida em meio às situações mais difíceis. Sou o que sou hoje muito devido a Ailton e Maria do Carmo, e infelizmente, um parágrafo apenas não é capaz de expressar o quão grato sou pela presença deles na minha vida. Então, um muito obrigado!

Agradeço também a Fernanda, que se mostrou uma namorada presente, compreensiva e companheira. Me motivou nos momentos em que eu precisei, e me incentivou até a chegada desta etapa, que é uma das mais importantes da minha vida.

Quero agradecer também a um amigo que a universidade me deu. Passamos por muitas dificuldade juntos desde os primeiros semestres, e chegamos na reta final, superando diversos obstáculos, apoiando um ao outro e descontraindo em meio a jogos de video game. A Adriano, meu muito obrigado.

Um muito obrigado para a professora Inês, a qual fico feliz pelas oportunidades que tive em tê-la como minha orientadora e como professora de outras disciplinas. Sou muito grato pela paciência, compreensão, auxílio e torcida no desenvolvimento de todo este projeto. Da mesma forma, aos professores Débora e Ernesto, pelo auxílio e força passados na disciplina de TCC 2.

Aos meus outros familiares e amigos que se fizeram presente neste período, o meu muito obrigado.

E por fim, e não menos importante, um obrigado especial a mim. Descobri uma força que não sabia que tinha, além de mostrar persistência e determinação em meio a uma mente cheia de preocupações e anseios, com os quais aprendi a viver para chegar até aqui. Um aprendizado importante que obtive durante este projeto é que eu sou mais forte e mais determinado do que eu penso. A mim, um muito obrigado por se provar melhor a cada dia.

"Tente uma, duas, três vezes e se possível tente a quarta, a quinta e quantas vezes for necessário. Só não desista nas primeiras tentativas, a persistência é amiga da conquista. Se você quer chegar aonde a maioria não chega, faça o que a maioria não faz."

(Bill Gates)

RESUMO

Baseado em diversos estudos para se entender o comportamento da febre chikungunya, existe uma precariedade, tanto na coleta de dados de casos confirmados de infecção, quanto no entendimento de quais fatores diretos e indiretos estão correlacionados com esta doença, que pode acarretar problemas sérios de saúde no indivíduo contaminado. Com a finalidade de se obter um conhecimento utilizável a respeito desta arbovirose, o presente projeto apresenta uma metodologia quantitativa baseada em *Data Science* para analisar os casos confirmados do vírus em um banco de dados de epidemiologia molecular. O objetivo desta análise foi buscar padrões entre a variedade dos genótipos da chikungunya, aspectos clínicos dos pacientes infectados, e fatores socioeconômicos das regiões de ocorrência dos casos. O desenvolvimento deste trabalho seguiu uma metodologia quantitativa, fundamentada no processo de descoberta de conhecimento criado por Fayyad et al. (1996), e com o uso de um algoritmo de *clustering* chamado de *K-prototype*, Huang (1997), para aplicação em dados do tipo misto. Para avaliar o desempenho do algoritmo em relação à similaridade entre os *clusters* gerados, duas métricas foram utilizadas: o Índice de Rand Ajustado e o Índice de Jaccard. Com os resultados levantados após a avaliação do algoritmo, foram encontradas tendências importantes, a partir dos padrões identificados entre os genótipos, o gênero e a idade dos pacientes infectados, além de características como o IDH, por exemplo, de regiões onde houveram casos confirmados de infecção pelo vírus chikungunya.

Palavras-chave: Ciência de Dados. Descoberta de Conhecimento. Banco de Dados. Epidemiologia Molecular. Mineração de Dados. K-Prototype. Vírus. Arbovírus. Chikungunya. Sintomas. Socioeconômico.

ABSTRACT

Based on several studies to understand the behavior of chikungunya fever, there is a precariousness both in the collection of data from confirmed cases of infection, and in the understanding of which direct and indirect factors are correlated with this disease, that can lead to serious health problems in the individual contaminated. In order to obtain a usable knowledge about this arbovirus, this project presents a quantitative methodology based on Data Science to analyze confirmed cases of the virus in a molecular epidemiology database. The objective of this analysis was to seek patterns between the variety of chikungunya genotypes, clinical aspects of infected patients, and socioeconomic factors in the regions where the cases occur. The development of this work followed a quantitative methodology, based on the knowledge discovery process created by Fayyad et al. (1996), and with the use of a clustering algorithm called K-prototype, Huang (1997), for application in mixed type data. To evaluate the performance of the algorithm in relation to the similarity between the generated clusters, two metrics were used: the Adjusted Rand Index and the Jaccard Index. With the results raised after the evaluation of the algorithm, important trends were found, based on patterns identified between genotypes, gender and age of infected patients, as well as characteristics such as the HDI, for example, from regions where there were confirmed cases of chikungunya virus infection.

Keywords: Data Science. Knowledge Discovery. Database. Molecular Epidemiology. Data Mining. K-Prototype. Virus. Arbovirus. Chikungunya. Symptoms. Socioeconomic.

LISTA DE FIGURAS

Figura 1 – Etapas do processo de KDD para análise de dados. 26
Figura 2 – Representação simples dos dados em cinco grupos após a aplicação do método de *clustering*. 28
Figura 3 – Diagrama do processo de análise da Chikungunya com o uso de Data Science. 38
Figura 4 – Visualização do conjunto de dados do arquivo baixado do ABVdb, com os 20 primeiros registros. 44
Figura 5 – Exibição dos primeiros e últimos registros do *dataframe* que representa o arquivo do ABVdb já formatado. 44
Figura 6 – Visualização dos primeiros registros do arquivo original que representa o conjunto de dados do IDH. 45
Figura 7 – Visualização dos primeiros e últimos registros do *dataframe* com os dados do Índice de Desenvolvimento Humano (IDH) até 1995. 46
Figura 8 – Exibição dos primeiros e últimos registros do *dataframe* do IDH entre os anos de 1990 e 2004, após a remoção das colunas. 47
Figura 9 – Visualização da versão final do *dataframe* do IDH com os primeiros e últimos registros. 48
Figura 10 – Gráfico de barra horizontal com as frequências dos genótipos no conjunto de dados do ABVdb. 49
Figura 11 – Gráfico de barra horizontal com as frequências dos valores para o atributo gênero no conjunto de dados do ABVdb. 50
Figura 12 – Histograma de frequência dos valores do IDH. 50
Figura 13 – Histograma de frequência dos valores do IDH após a remoção das linhas que não serão utilizadas. 51
Figura 14 – Histograma de frequência dos valores do ASB. 51
Figura 15 – Histograma de frequência dos valores do GSPIB. 52
Figura 16 – Histograma de frequência dos valores do ASB após a remoção das linhas que não serão utilizadas. 52

Figura 17 – Histograma de frequência dos valores do GSPIB após a remoção das linhas que não serão utilizadas. 53
Figura 18 – Primeiros e últimos registros do *Dataframe* que representa a primeira amostra a ser utilizada na análise. 55
Figura 19 – Visualização dos primeiros e últimos registros do *Dataframe* que representa a segunda amostra. 55
Figura 20 – Primeiros e últimos registros do *Dataframe* que representa a terceira amostra a ser utilizada na análise. 57
Figura 21 – Visualização dos primeiros e últimos registros do *Dataframe* que representa a quarta amostra. 57
Figura 22 – Resultado da aplicação do método do cotovelo no conjunto de dados da primeira amostra. 60
Figura 23 – Resultado da aplicação do método do cotovelo no conjunto de dados da segunda amostra. 60
Figura 24 – Resultado da aplicação do método do cotovelo no conjunto de dados da terceira amostra. 61
Figura 25 – Resultado da aplicação do método do cotovelo no conjunto de dados da quarta amostra. 61
Figura 26 – Processo de execução em uma rodada de experimentos. Os valores apresentados são ilustrativos. 62
Figura 27 – Processo de aplicação das métricas para avaliar o algoritmo. Os valores apresentados são ilustrativos. 63
Figura 28 – Resultados do Índice de Rand Ajustado para cada uma das amostras em cada um dos experimentos. 66
Figura 29 – Porcentagem dos *clusters* gerados semelhantes ao *cluster* referencial. 67
Figura 30 – Resultados do Índice de Jaccard para cada uma das amostras em cada um dos experimentos. 68
Figura 31 – Porcentagem dos *clusters* que em ambas as métricas obtiveram o valor 1. . . 69
Figura 32 – Gráfico de barras contendo em ordem, de cima para baixo, as variáveis mais significativas para o agrupamento da primeira amostra. 72
Figura 33 – Gráfico de distribuição de dados das variáveis genótipo e IDH nos *clusters* da primeira amostra. 73

Figura 34 – Gráfico de distribuição de dados das variáveis genótipo e ASB nos *clusters* da primeira amostra. 73
Figura 35 – Gráfico de distribuição de dados das variáveis genótipo e País nos *clusters* da primeira amostra. 74
Figura 36 – Gráfico de barras contendo em ordem, de cima para baixo, as variáveis mais significativas para o agrupamento da segunda amostra. 75
Figura 37 – Gráfico de distribuição de dados das variáveis genótipo e idade nos *clusters* da segunda amostra. 76
Figura 38 – Gráfico de distribuição de dados das variáveis gênero e idade nos *clusters* da segunda amostra. 77
Figura 39 – Gráfico de barras com a frequência de infecção dos genótipos da chikungunya em homens de diferentes faixas de idade. 77
Figura 40 – Gráfico de barras com a frequência de infecção dos genótipos da chikungunya em mulheres de diferentes faixas de idade. 78
Figura 41 – Gráfico de barras contendo em ordem, de cima para baixo, as variáveis mais significativas para o agrupamento da terceira amostra. 79
Figura 42 – Primeiros registros do conjunto de dados da terceira amostra, com as informações dos acessos e dos *clusters*. 80
Figura 43 – Gráfico de barras contendo em ordem, de cima para baixo, as variáveis mais significativas para o agrupamento da quarta amostra 81
Figura 44 – Gráfico de distribuição de dados das variáveis genótipo e sintomas nos *clusters* da quarta amostra. 82
Figura 45 – Recorte do código da função execKPrototype() implementada na linguagem Python. 90
Figura 46 – Recorte do código da função *elbowKPrototype()* implementada na linguagem Python. 91
Figura 47 – Recorte do código da função *randIndex()* implementada na linguagem Python. 92
Figura 48 – Recorte do código da função *jaccardIndex()* implementada na linguagem Python. 93

LISTA DE TABELAS

Tabela 1 – Quantidade de linhas e colunas dos *dataframes* em suas versões iniciais. . . 42
Tabela 2 – Quantidade de linhas e colunas em cada *dataframe* após o pré-processamento dos dados. 48
Tabela 3 – Relação entre a quantidade de linhas analisadas, as restantes e as que foram excluídas em cada *dataframe*. 53
Tabela 4 – Estrutura das quatro amostras a serem utilizadas na análise. 57

Tabela 5 – Dados dos protótipos encontrados com o *K-prototype* na primeira amostra . 72
Tabela 6 – Dados dos protótipos encontrados com o *K-prototype* na segunda amostra . 75
Tabela 7 – Dados dos protótipos encontrados com o *K-prototype* na terceira amostra . . 79
Tabela 8 – Dados dos protótipos encontrados com o *K-prototype* na quarta amostra . . 81

LISTA DE QUADROS

Quadro 1 – Bibliotecas da linguagem Python utilizadas no projeto. 40
Quadro 2 – Variáveis selecionadas para o projeto em cada uma das bases de dados. . . 41
Quadro 3 – Sintomas e variações da chikungunya descritos pela OMS e encontrados nos artigos bases do GenBank. 42

LISTA DE ALGORITMOS

Algoritmo 1 – Processo de alocação inicial 32
Algoritmo 2 – Processo de realocação 33

LISTA DE ABREVIATURAS E SIGLAS

ABVdb	*Arthropod Borne Virus Database*
ASB	Acesso ao Saneamento Básico
BI	*Business Intelligence*
CSV	*Comma Separated Values*
ECSA	*East Central South African*
GB	Gigabytes
GSPIB	Gastos na Saúde com o PIB
IDE	*Integrated Development Environment*
IDH	Índice de Desenvolvimento Humano
KDD	*Knowledge discovery in databases*
OMS	Organização Mundial de Saúde
OPAS	Organização Pan-Americana da Saúde
PIB	Produto Interno Bruto
RAM	Random Access Memory
SVM	*Support Vector Machine*
UNDP	*United Nations Development Programme*

LISTA DE SÍMBOLOS

$\sum$	Somatório
δ	delta

SUMÁRIO

1 INTRODUÇÃO 18
2 TÉCNICAS DE CLUSTERING APLICADAS A ARBOVÍRUS 21
3 O VÍRUS CHIKUNGUNYA E AS POSSÍVEIS ESTRATÉGIAS PARA DESCOBERTA DE CONHECIMENTO 23
3.1 Vírus chikungunya 23
3.2 Data Science 24
3.3 KDD e Data Mining 25
3.4 Clustering 27
3.4.1 Classificação dos algoritmos de clustering 28
3.4.2 Clustering de dados mistos 29
3.5 K-prototype 30
3.5.1 Função de Custo e Medida de Similaridade 30
3.5.2 Funcionamento do algoritmo 31
3.6 Métricas avaliativas para a técnica de clustering 34
3.6.1 Índice de Rand 34
3.6.2 Índice de Rand Ajustado 35
3.6.3 Índice de Jaccard 35
4 O DESENVOLVIMENTO DAS ETAPAS PARA A DESCOBERTA DE CONHECIMENTO 37
4.1 Metodologia de Pesquisa Quantitativa 37
4.2 Configurações do Ambiente de Desenvolvimento 39
4.3 Seleção de Dados 39
4.4 Pré-Processamento de Dados 43
4.5 Análise Exploratória 48
4.6 Transformação de Dados 53
4.6.1 Criação da Primeira e Segunda Amostra 54
4.6.2 Criação da Terceira e Quarta Amostra 56
4.7 Mineração de Dados 58

4.7.1 Identificação da quantidade de *clusters* 58
4.7.2 Execução dos Experimentos 60
4.8 Avaliação dos algoritmos **62**
5 RESULTADOS **65**
5.1 Avaliação dos resultados **65**
5.1.1 Avaliação dos resultados para o Índice de Rand Ajustado 66
5.1.2 Avaliação dos resultados para o Índice de Jaccard 68
5.1.3 Avaliação Final das métricas 69
5.2 Interpretação dos resultados **70**
5.2.1 Interpretação dos resultados da primeira amostra. 71
5.2.2 Interpretação dos resultados da segunda amostra. 74
5.2.3 Interpretação dos resultados da terceira amostra. 78
5.2.4 Interpretação dos resultados da quarta amostra. 80
6 CONSIDERAÇÕES FINAIS **83**
REFERÊNCIAS **85**
APÊNDICES **89**
APÊNDICE A – Implementação do algoritmo *K-prototype* na linguagem Python **90**
APÊNDICE B – Implementação do método do cotovelo na linguagem Python **91**
APÊNDICE C – Implementação das métricas **92**

1 INTRODUÇÃO

Os arbovírus são classes de vírus que caracterizam-se pelo desenvolvimento do seu ciclo de vida em insetos artrópodes, e que são transmitidos aos seres humanos através da picada de mosquitos hematófagos. Existem cerca de 545 espécies diferentes de arbovírus, e uma diversidade de problemas já foram identificados em pacientes infectados por eles. São classificados em 5 famílias: *Bunyaviridae, Flaviviridae, Reoviridae, Rhabdoviridae e Togaviridae*. A Dengue, Zika, Febre Amarela e Chikungunya são exemplos de arbovírus (LOPES et al., 2014).

Como um dos membros da família *Togaviridae*, o vírus chikungunya é transmitido através da picada de dois mosquitos da família *Aedes*: o *Aegypti* e o *Albopictus*. A nomenclatura deste arbovírus é derivada de uma palavra africana, que significa "aqueles que se dobram", uma referência ao estado físico dos contaminados pelo vírus que sofrem com os problemas articulares. Indivíduos infectados apresentam sintomas característicos como febre alta e artralgia, e podem vir a desenvolver outros, como vômito, náuseas, conjuntivite, erupção cutânea e dor nas costas (BURT et al., 2017).

Segundo Barata (2009), nem sempre os fatores de risco conseguem esclarecer os reais motivos das ocorrências de problemas de saúde. Muitas vezes, outros aspectos importantes são deixados de lado, e precisam ser melhor observados ao se levar em conta a compreensão de uma doença. Por exemplo, Siriyasatien et al. (2018), relata que dados do período entre 1998 e 2002, em Cingapura, mostraram que o número de larvas do mosquito da dengue não estava correlacionado com a incidência da doença, mas com fatores ecológicos, condições socioeconômicas ou o habitat da população de baixa renda, indicando uma complexidade maior nos fatores que podem estar relacionados à doença. Estes aspectos precisam ser identificados e abordados para um melhor entendimento acerca deste tipo de doença.

Em meio à disseminação e crescimento dos casos de infectados por estas doenças ao redor do mundo, uma imensa quantidade de dados são gerados que contém as mais diversas características tanto dos vírus, quanto das regiões de registro das incidências. No âmbito computacional, a área de *Data Science* (no português, Ciência de Dados) disponibiliza uma metodologia eficaz para se extrair, analisar e gerar informações úteis a partir destes dados,

(SIRIYASATIEN et al., 2018). Incluso neste âmbito, o *Knowledge discovery in databases* (KDD) (no português, Descoberta de Conhecimento em Base de dados) criado por Fayyad et al. (1996), define todo um processo dividido por etapas para se estudar os conjuntos de dados, que vai desde a seleção até a obtenção de um conhecimento específico. Dentre estas etapas, o *Data Mining* (no português, Mineração de Dados) dispõe de ferramentas que auxiliam no processo de modelagem dos dados para o entendimento acerca de arbovírus como a chikungunya. Com a sua utilização, é possível encontrar padrões que não são facilmente perceptíveis, em variáveis diversas, afim de se obter um conhecimento útil a respeito da doença.

Segundo Huang (1998), é característica do *Data Mining* trabalhar com grandes volumes de dados de forma eficiente. O *clustering* (no português, agrupamento) é uma das técnicas que podem ser empregadas pelos algoritmos de *Data Mining*, e consiste em agrupar conjuntos de dados para analisá-los de forma mais aprofundada. Atualmente, a maior parte das informações é gerada das mais diversas fontes e tipos, sejam estes últimos quantitativos (valores numéricos) ou qualitativos (valores textuais). Porém, boa parte dos algoritmos de *clustering* não são capazes de operar com dados de ambos os tipos. Dentre aqueles que o conseguem, a maioria exige um alto custo computacional, tornando inviável a sua utilização com recursos limitados. Foi preciso então a filtragem e avaliação de um algoritmo que conseguisse agrupar os dados do projeto, de maneira eficiente e com a utilização do mínimo de recursos possíveis.

Para a correta aplicação de uma metodologia em torno da área de *Data Science* para analisar o vírus chikungunya, é preciso coletar dados fidedignos com ocorrências reais de casos confirmados da doença ao redor do mundo. Para isso, o *Arthropod Borne Virus Database* (ABVdb)(no português, Banco de dados de Vírus transmitidos por Artrópodes), desenvolvido por Restovic (2018), fornece as informações necessárias para subsidiar uma análise deste tipo. Ele é um banco de dados de epidemiologia molecular que contém características genéticas, clínicas e epidemiológicas dos vírus chikungunya, dengue e zika, o que o torna uma ferramenta de bioinformática propícia para a realização de estudos com dados virais.

Após uma revisão na literatura, identificou-se a ausência de estudos com dados reais que identificassem relações entre variáveis em torno de casos de infecção pelo vírus chikungunya e fatores socioeconômicos. Diante deste problema de pesquisa, este projeto teve como objetivo analisar os genótipos desta arbovirose em casos registrados em um banco de dados de epidemiologia molecular, utilizando uma metodologia de *Data Science* para, a

partir da obtenção de dados de outras fontes confiáveis, buscar padrões entre eles, os pacientes infectados e aspectos socioeconômicos relacionados à região de ocorrência. Para se alcançar este objetivo, tarefas importantes precisaram ser feitas, como a criação de novos conjuntos de dados a partir dos originais baixados, a criação de um conjunto de dados específico contendo os sintomas identificados na literatura relacionados aos casos registrados no ABVdb e a avaliação do algoritmo aplicado a dados do tipo misto.

A motivação para o desenvolvimento deste projeto deu-se a partir da oportunidade de aplicar uma metodologia para entender melhor o vírus chikungunya, a partir de um estudo com dados do mundo real, e que possa vir a ser utilizado como referência para futuras pesquisas sobre esta arbovirose. Como contribuição para a área computacional, o projeto irá utilizar uma linguagem de programação que forneça os recursos necessários para a implementação de cada uma das etapas do KDD, o emprego de um algoritmo de agrupamento para dados mistos e a utilização de métricas para avaliá-lo. Já para a área de epidemiologia molecular, as percepções encontradas podem auxiliar futuros estudos sobre a chikungunya que utilizem o ABVdb como base, além de que, toda a metodologia aplicada pode servir também para a análise de outros vírus.

O desenvolvimento deste projeto está descrito nos capítulos que aparecem a seguir. No capítulo 2, são apresentados os estudos mais significativos que utilizaram técnicas de *clustering* para analisar os arbovírus. No capítulo 3, a referência teórica é detalhada, baseada no processo de descoberta de conhecimento do vírus chikungunya. No capítulo 4, cada etapa do KDD contextualizada ao projeto foi explicada. No capítulo 5, os resultados das métricas avaliativas são discutidos, e as percepções encontradas são levantadas. Por fim, no capítulo 6, são descritas as considerações finais a respeito de todo o projeto.

2 TÉCNICAS DE CLUSTERING APLICADAS A ARBOVÍRUS

Diversas pesquisas já utilizaram a técnica de *clustering* para analisar as ocorrências dos arbovírus ao redor do mundo. Seja com dados fictícios ou reais divulgados publicamente, estes estudos contribuíram para se entender como a aplicação de um algoritmo de *Data Mining* pode auxiliar no entendimento a respeito de uma doença. A seguir será mostrado de forma resumida o que foi feito, as variáveis utilizadas (se disponível) para as análises, a forma como os algoritmos foram aplicados e uma síntese sobre os resultados obtidos.

O presente projeto é uma continuação da linha de pesquisa em torno do ABVdb, sendo um aprofundamento mais específico do estudo desenvolvido por Conceição (2020). Nele, o objetivo foi comparar o desempenho dos algoritmos de *clustering K-modes*, *K-Prototype* de Cao e *K-Prototype* de Huang, para agrupar os dados de tipo misto contidos neste banco. Foram utilizadas três métricas para avaliar os resultados: o Índice de Jaccard, o F-Measure e o Índice de Rand Ajustado, onde o *K-Prototype* de Huang apresentou a maior estabilidade entre os três algoritmos aplicados. A pesquisa desenvolvida por Conceição (2020) serviu como base para a análise a ser desenvolvida neste projeto.

Muzakki e Nhita (2018), utilizaram as técnicas de agrupamento K-means e *Support Vector Machine* (SVM) para tentar predizer a ocorrência da Dengue, na cidade de Bandung, Indonésia. Os dados utilizados foram relacionados à meteorologia da região, os casos da doença e dados populacionais do período de 2009 a 2016. O K-means mostrou uma acurácia de mais de 86% ao identificar uma relação entre os atributos climáticos e a doença.

Já Mathulamuthu et al. (2017), utilizou a metodologia de aprendizagem múltipla para aplicação em fatores climatológicos e a técnica de agrupamento K-means junto à técnica estatística de regressão linear para construir um modelo preditivo para os casos de Dengue no estado de Selangor, Malásia. Foram utilizados dados climáticos e de casos de infecção da doença na região para a análise. O método da silhueta foi utilizado para identificar a melhor quantidade de *clusters* para o K-means e a regressão linear foi utilizada para medir os valores de eficácia para cada um dos *clusters* gerados.

O algoritmo K-means foi utilizado por Agarwal et al. (2018), juntamente com o

algoritmo de Nave Bayes, para analisar ocorrências do vírus da dengue com fatores demográficos e climáticos. Como resultados, foram obtidos padrões entre os casos confirmados do vírus e dos efeitos da chuva, da umidade relativa do ar e da densidade populacional. Foi sugerido o desenvolvimento de um sistema de alerta para auxílio na tomada de decisão dos países ao tratar de surtos relacionados com a doença, além também de novas análises com o uso de outras variáveis demográficas.

Yogapriya e Geetha (2019), buscaram categorizar as idades das pessoas que foram afetadas pelo vírus da Dengue com a construção de uma metodologia para comparar os resultados de 3 técnicas de agrupamento: K-means, Hierárquico e Konohem-SOM. No agrupamento Hierárquico os dados são agrupados juntos com o uso da metodologia baseada em árvores e o Kohonem-SOM acrescenta a idéia de vizinhança entre os grupos. Foram reunidos dados clínicos de pacientes infectados com o vírus da Dengue de áreas metropolitanas e subtropicais. Após a aplicação das técnicas, o K-means e o Konohem-SOM obtiveram os melhores resultados, com uma taxa de acurácia de 61,9% e 65,7%, respectivamente.

É proposto por Thakur e Kaur (2017) um sistema para predição e prevenção da Chikungunya com o uso do agrupamento K-means, para classificar se a pessoa está infectada ou possivelmente infectada com a doença. Foram coletados dados pessoais e clínicos dos pacientes, e o K-means foi utilizado para relacioná-los, de acordo com os atributos pertinentes. Como resultado, o sistema contém alertas para a possibilidade de infecção pelo vírus para a região identificada e para medidas de segurança a serem tomadas. Porém, após toda a análise, pode ser possível que, para comprovar a infecção pelo vírus, o indivíduo precise ir a um médico especializado para identificação do problema.

3 O VÍRUS CHIKUNGUNYA E AS POSSÍVEIS ESTRATÉGIAS PARA DESCOBERTA DE CONHECIMENTO

A área de *Data Science* oferece diversos recursos que auxiliam na busca por conhecimento útil em conjuntos de dados. Com ela, é possível identificar padrões em meio às características específicas em bases de dados reais. Neste capítulo serão detalhadas as bases teóricas a respeitos dos principais tópicos deste projeto, que seguiu baseado em uma metodologia em torno de *Data Science*. Serão detalhadas informações importantes sobre o vírus analisado, as etapas necessárias para a descoberta de conhecimento sobre os dados, os métodos da área de *Data Mining* necessários para a aplicação desta pesquisa, as especificações sobre o algoritmo empregado, e as métricas para avaliá-lo.

3.1 VÍRUS CHIKUNGUNYA

O vírus chikungunya é um arbovírus do tipo *Togaviridae*, pertencente à família dos *Alphavirus*, e que é transmitido aos seres humanos através da picada dos mosquitos *Aedes Aegytpi* e *Aedes Albopictus*. O vírus foi identificado no ano de 1952, no Planalto Maconde, uma região ao sul da África, e desde 2004 um surto se expande pela Europa e Ásia (BURT et al., 2017). No continente americano, um grande surto sucedeu-se em 2015 com um total de 693.489 casos relatados à Organização Pan-Americana da Saúde (OPAS) (OMS, 2017). Seus sintomas incluem principalmente os problemas articulares graves, seguidos por febre superior à 39ºC, dor de cabeça, náuseas e em alguns casos erupções cutâneas (BURT et al., 2017).

No Brasil, em um levantamento realizado pelo Ministério da Saúde (2019b), de Janeiro a Março de 2019, um total de 5.214 municípios notificaram o risco de aumento do surto da dengue, zika e chikungunya. No Boletim Epidemiológico divulgado pelo Ministério da Saúde (2019a), até setembro de 2019, foram registrados 110.627 de possíveis casos de infecção pelo vírus da chikungunya no país, um valor aproximadamente 31% maior do que no mesmo período em 2018 (com um registro de 76.742 casos). As regiões sudeste e nordeste são as mais afetadas, e no geral, 112 óbitos tem alguma relação com o vírus (57 confirmados e 65 em investigação).

Uma característica deste tipo de vírus é possuir uma alta taxa de mutações em sua composição genética dando origem a diversos genótipos. O vírus da chikungunya possui três

genótipos: o *Asian*, o *East Central South African* (ECSA) e o *West African* (BURT et al., 2017). Estas informações estão vinculadas a outra variáveis no banco de dados *ABVdb*.

O *ABVdb* é um banco de dados de epidemiologia molecular corretamente genotipado e sequencialmente curado, desenvolvido por Restovic (2018) na sua tese de doutorado na FIOCRUZ. Nele estão contidas as informações a respeito do genótipo, da região geográfica, da data de coleta da amostra, do hospedeiro, e dos parâmetros epidemiológicos e clínicos de três arbovírus: dengue, zika e chikungunya. Todos os registros no *ABVdb* que serão utilizados neste projeto são de casos 100% confirmados de infecção em seres humanos pelo vírus da chikungunya, o que trás confiabilidade em qualquer relação que seja identificada com outros fatores oriundos de fontes fidedignas. Este banco de dados está disponível publicamente em www.abvdb.uneb.br.

Para se trabalhar com dados a respeito do vírus da chikungunya armazenadas no *ABVdb* e relacioná-los com outras informações do mundo real, é preciso de uma metodologia propícia para analisá-los. A área de *Data Science* procura compreender os mais diversos tipos de dados aplicados em um contexto, a fim de se obter informações que possam auxiliar em algum âmbito. A próxima seção abordará sobre o que é essa área que vem se destacando ao longo dos últimos anos.

3.2 DATA SCIENCE

Ao redor do mundo são gerados a todo o momento uma imensa quantidade de conteúdo digital, de todos os tipos e formatos, e das mais diversas fontes disponíveis. Esse contexto faz referência ao termo de *Big Data* (no português, Megadados ou Grandes Dados), que consiste no volume, na velocidade de atualização, na variedade de formatos, na veracidade e no valor de toda essa gama de dados que são criados em fotos, vídeos, áudios, textos, etc (CAVIQUE, 2014). O conceito de *Big Data* levanta novos desafios no que diz respeito ao tratamento e análise de tudo que é gerado, e uma das áreas que apresenta soluções é a de *Data Science*.

Data Science consiste na prática de analisar dados e levantar informações úteis e consistentes para algum propósito. É uma área que une *Machine Learning* (no português, Aprendizado de Máquina), *Data Mining* e Estatística com o objetivo de obter novas percepções sobre quantidade de dados dos mais variados tamanhos, para auxiliar na tomada de decisões de diversas áreas como médica, financeira, social, ecológica, etc. Na década anterior também era conhecida como *Business Intelligence* (BI), até ser conceituada com o seu atual nome em 2010

(CAVIQUE, 2014). Dispõe de diversas ferramentas e soluções analíticas para transformar os dados em informações, que vão desde a quantidade de horas por dia que um indivíduo dorme até qual a relação entre fatores climáticos e uma epidemia viral.

O processo de *Data Science* pode ser melhor detalhado através de uma sequência de etapas de acordo com um processo chamado KDD. A próxima seção detalhará sobre o que cada uma delas aborda.

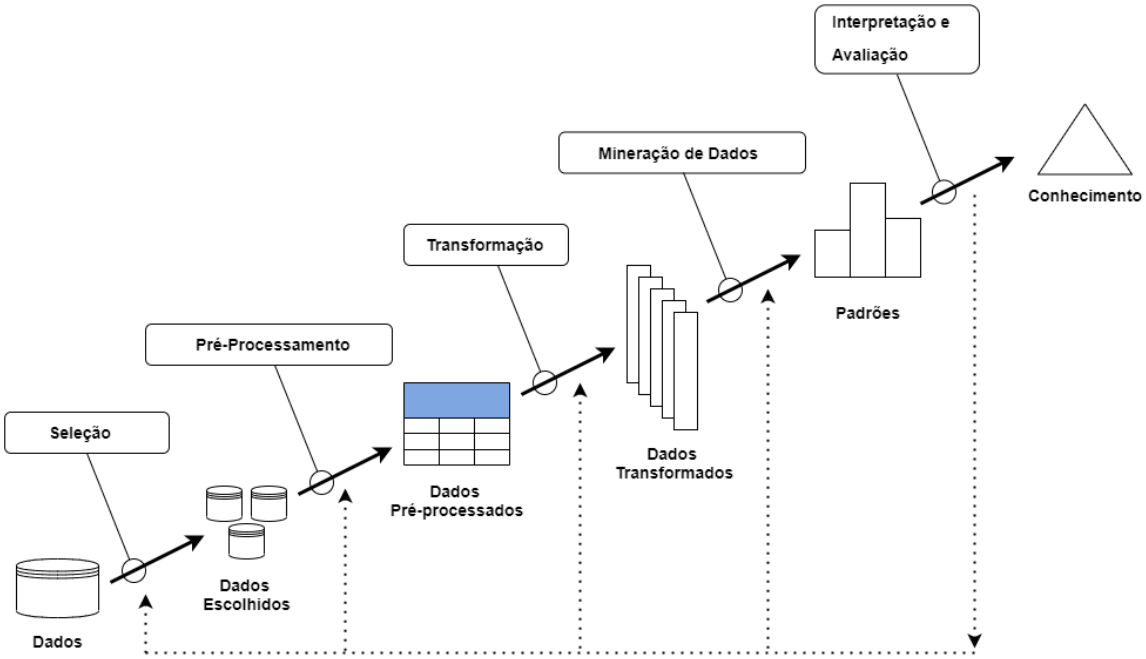
3.3 KDD E DATA MINING

Segundo Fayyad et al. (1996), em diversas áreas de conhecimento, os dados vem sendo coletados e acumulados em uma velocidade muito alta, criando uma necessidade de que existam ferramentas computacionais para auxiliar os humanos a extrair informações úteis de grande volumes de dados. O KDD surge com este propósito, visando detalhar um conjunto de passos que auxilie neste processo, desde a seleção inicial dos dados até a obtenção da informação desejada. Ainda de acordo com Fayyad et al. (1996), o termo surgiu em um *workshop* em 1989 para destacar que o conhecimento é o produto final de uma descoberta baseada em dados, se popularizando nos campos de Inteligência Artificial e *Machine Learning*.

O processo do KDD apresenta duas propriedades importantes: a iteratividade e a interatividade. No que diz respeito ao primeiro termo, existe uma sequência de etapas nas quais o resultado de uma depende diretamente da outra. Já para o segundo, permite a quem está analisando poder intervir em cada uma das etapas (FAYYAD et al., 1996). De uma forma geral, cada etapa pode ser executada diversas vezes, e elas estão divididas em cinco, que são ilustradas na Figura 1 e detalhadas a seguir:

- Etapa 1 - Seleção: É o primeiro estágio do processo, que tem como objetivo selecionar conjuntos de dados de diferentes fontes e formatos, que farão parte da análise.
- Etapa 2 - Pré-processamento de Dados: Tem como operações básicas a limpeza dos dados para a correção ou exclusão de inconsistências que possam atrapalhar a análise, além da identificação e correção dos *outliers*;
- Etapa 3 - Transformação de Dados: Nesta etapa os dados são preparados para a tarefa de mineração, de forma a transformá-los em formatos apropriados com o uso de técnicas específicas.

Figura 1 – Etapas do processo de KDD para análise de dados.



Fonte – Adaptada de Fayyad et al. (1996).

- Etapa 4 - Mineração de Dados: É a etapa essencial do KDD, e consiste na modelagem ou aplicação de técnicas e algoritmos de *Data Mining*. Estas técnicas possuem dois tipos de objetivos distintos: verificar uma hipótese levantada pelo usuário ou descobrir novos padrões de forma autônoma. Este último, ainda pode ser subdividido em uma descoberta preditiva (onde o sistema procura encontrar padrões para prever o futuro comportamento de algumas entidades) ou uma descoberta descritiva (onde o sistema encontra padrões para apresentação a uma usuário de forma compreensiva). A mineração de dados orientada para a descoberta descritiva foi a escolhida por ser a mais apropriada para a análise deste projeto.
- Etapa 5 - Interpretação e Avaliação: A etapa final é a avaliação do desempenho do modelo na identificação dos padrões que representam o conhecimento útil. A validação nesta etapa pode ser feita com a utilização de métricas para avaliar o desempenho da técnica empregada.

Como visto na descrição das etapas, o *Data Mining* é uma das etapas de todo o processo de descobrimento do conhecimento, e envolve a modelagem de dados. De acordo com Mannilat (1996), existe um vínculo entre o *Data Mining* e *Machine Learning*, no qual, além destas grandes áreas terem o objetivo em comum de encontrar regularidades ou padrões em dados

empíricos, o *Data Mining* utiliza o *Machine Learning* em seus algoritmos. Existem algumas diferenças relacionadas aos seus objetivos, onde o foco do *Machine Learning* é ser aplicado em um aprendizado que para o ser humano é complexo de ser realizado, enquanto o do *Data Mining* busca extrair e obter de um conhecimento contextualizado. Também, este último foca em análises de grandes conjuntos de dados, enquanto o primeiro utiliza exemplos específicos para gerar determinadas percepções de um contexto. Mas, de uma forma geral, o *Machine Learning* é um componente essencial para o processo de *Data Mining*.

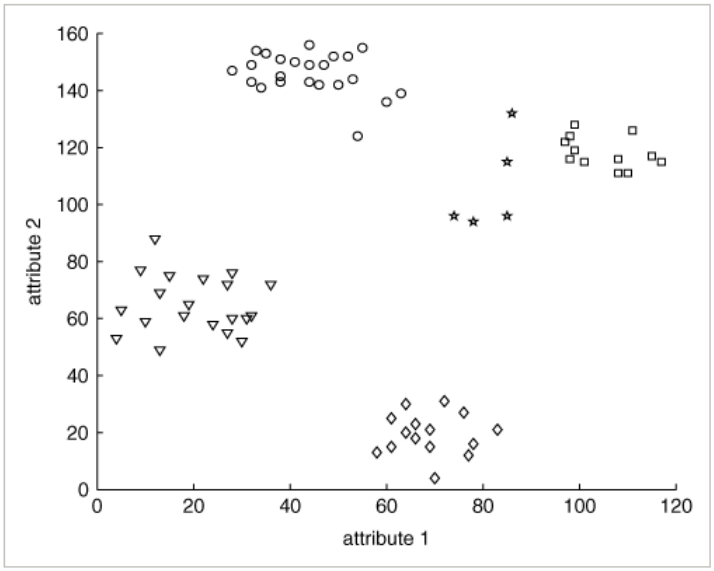
Diante deste contexto, os dados são a matéria-prima fundamental para o desenvolvimento de todas as etapas do KDD. E uma das técnicas de *Data Mining* que busca aplicar a descoberta descritiva sobre eles em uma análise é o *clustering*. Na próxima seção será abordado como é o seu funcionamento.

3.4 CLUSTERING

De acordo com Huang (1998), uma das operações fundamentais resultantes do *Data Mining* é particionar um conjunto de dados em grupos para posteriormente analisá-los de forma mais aprofundada. Esta técnica conhecida como *clustering*, é útil em tarefas como segmentação, agregação e classificação. Ela separa um conjunto de objetos em grupos, de forma que os objetos em um mesmo grupo sejam mais semelhantes entre si do que os objetos em diferentes grupos, de acordo com alguns critérios definidos.

A análise de *clustering* em *Data Mining* tem tido um papel muito importante, e vem sendo utilizada em diversas áreas como no reconhecimento de padrões, na análise de dados espaciais, no processamento de imagens, em análises médicas, na economia e principalmente na bioinformática e biometria (EDNA, 2006). São exemplos de aplicações de *clustering* em um contexto de descoberta de conhecimento a busca por subpopulações homogêneas para consumidores em base de dados de marketing e a identificação de subcategorias de espectros de medições infravermelhas do céu (FAYYAD et al., 1996). A Figura 2 representa um exemplo de *clustering*, onde os dados com características similares encontram-se agrupados, preservando as características de cada grupo.

Figura 2 – Representação simples dos dados em cinco grupos após a aplicação do método de *clustering*.



Fonte – Recortada de Vendramin et al. (2010)

3.4.1 Classificação dos algoritmos de clustering

Cassiano (2015) detalha os 10 tipos principais de algoritmos de *clustering*, que são: Métodos Hierárquicos, Métodos Particionais, Métodos Baseados em Densidade, Métodos Baseados em Grade, Métodos Baseados em Modelos, Métodos Baseados em Redes Neurais, Métodos Baseados em Lógica Fuzzy, Métodos Baseados em Kernel, Métodos Baseados em Grafos e Métodos Baseados em Computação Evolucionária. Alguns algoritmos destes métodos juntam idéias de outros, tornando difícil muitas vezes indicar que aquele algoritmo é exclusivamente pertencente àquele método, Cassiano (2015 apud HAN; KAMBER, 2001). Mas de uma forma geral, são dois os métodos mais tradicionais utilizados e explicados a seguir por Cassiano (2015):

- Métodos Hierárquicos: Nessa classificação, o conjunto de dados é separado em uma estrutura hierárquica baseada no quão próximos estão os seus elementos, e a saída dos algoritmos contidos nesse método é reproduzida em formato de uma árvore binária chamada de dendograma, que segmenta a base de dados em grupos menores e de forma repetitiva. Todo o conjunto de dados é representado pela raíz dessa árvore, enquanto os filhos são os nós folhas, e o resultado do *clustering* pode ser alcançado divindo o dendograma em diferentes níveis baseado em um número de grupos predefinidos.
- Métodos Particionais: Nessa classificação, o conjunto de dados é separado em uma quantidade k de grupos, estes definidos pelo usuário. O processo inicial de separação dos elementos é detalhado da seguinte forma:

> Inicialmente, o algoritmo escolhe k objetos como sendo os centros dos k clusters. Os objetos são divididos entre os k clusters de acordo com a medida de similaridade adotada, de modo que cada objeto fique no cluster que forneça o menor valor de distância entre o objeto e o centro do mesmo. Então, o algoritmo utiliza uma estratégia iterativa de controle para determinar que objetos devem mudar de cluster, de forma que a função objetivo usada seja otimizada.

A etapa seguinte a essa separação é escolher os elementos que se tornarão os centros dos clusters definidos. Para isso, pode-se escolher a utilização de uma entre duas formas possíveis: a abordagem *k-medoids* em que o elemento escolhido como representante está mais próximo ao centro de gravidade de cada cluster, ou a abordagem *k-means*, onde o elemento é escolhido a partir da média aritmética dos elementos contidos no centro de gravidade de cada cluster (CASSIANO, 2015).

3.4.2 Clustering de dados mistos

A especialidade do *Data Mining* é trabalhar com gigabytes ou até mesmo terabytes de dados de forma eficiente. Os conjuntos de dados normalmente possuem vários elementos com dois tipos de variáveis: as quantitativas, com valores em formato numérico e as qualitativas com valores em formato de atributos ou qualidades representados textualmente. No *clustering*, é preciso que softwares possam agrupar dados do tipo misto, onde seus elementos possuem variáveis de ambas as categorias. Mas a maioria deles não conseguem fazê-lo de forma eficiente, e operam com limitações relacionadas ao tipo de atributo da variável, (HUANG, 1998).

Segundo Huang (1998), alguns métodos de *clustering* do tipo hierárquico podem trabalhar com dados do tipo misto, mas o custo computacional torna o uso deles inviável para agrupar grandes conjuntos de dados. Já na categoria dos métodos particionais, o algoritmo *K-means* é eficiente para processar essa quantidade de dados, porém, ele é limitado ao uso somente de dados onde suas variáveis são do tipo quantitativas. O mesmo acontece com o algoritmo *k-modes*, mas com variáveis qualitativas. Outro problema também é que o tratamento tradicional de converter dados categóricos para numéricos, e utilizá-los no algoritmo não necessariamente produz resultados significativos. Diante deste cenário, o autor propõe uma solução com o uso do algoritmo *K-Prototype* para a utilização do método de *clustering* em grandes conjuntos de dados de ambos os tipos com eficiência.

3.5 K-PROTOTYPE

O *K-prototype* de Huang (1997) é um algoritmo de *clustering* que resolve problemas de partição de dados em *Data Mining*. Ele é baseado no algoritmo *K-means*, mas remove a limitação de utilizar somente variáveis numéricas, sem perder a sua eficiência. O seu nome se dá devido ao fato de que os objetos são agrupados em uma quantidade k de protótipos. Segundo Huang (1997), em testes com dados reais, ele foi capaz de particionar, em algumas horas, conjuntos de dados com cerca de cem mil registros, descritos por mais ou menos 20 atributos numéricos e categóricos, em 100 grupos. Nesta seção serão descritos os conceitos necessários para compreender o funcionamento deste algoritmo.

3.5.1 Função de Custo e Medida de Similaridade

Inicialmente é preciso entender algumas noções matemáticas acerca da função de custo e da medida de similaridade que são aplicada internamente ao algoritmo. Segundo Huang (1997), seja $X = \{X_1, X_2, X_3, ..., X_n\}$ um conjunto de n objetos, $X_i = \{X_{i1}, X_{i2}, X_{i3}, ..., X_{im}\}$ um conjunto de m atributos dos objetos e k um valor inteiro positivo. O objetivo deste método de *clustering* é encontrar uma partição que separe os objetos dentro de X em uma quantidade k de grupos diferentes.

Ainda de acordo com o autor, o número de partições possíveis para um determinado valor n é extremamente grande, tornando inviável analisar cada uma delas para se encontrar a melhor. Uma possível solução para esse problema é escolher um critério de agrupamento, também chamado de função de custo, para auxiliar na busca por uma partição específica. A função de custo utilizada por Huang (1997) é definida por:

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{il} d(X_i, Q_l) \tag{3.1}$$

onde $Q_1 = [q_{l1}, q_{l2}, ..., q_{lm}]$ é o vetor representativo, ou o *prototype* (centro) do cluster l; y_{il} é um elemento de valor binário (0 ou 1) que representa a pertinência de X_i. Já $d(X_i, Q_l)$ indica a medida de similaridade, na qual o primeiro termo é o quadrado da distância Euclidiana dos atributos numéricos, e o segundo termo é a medida de dissimilaridade nos

atributos categóricos. Esta medida é representada por:

$$d(X_i, Q_l) = \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^2 + y_l \sum_{j=1}^{m_c} \delta(x_{ij}^c, q_{lj}^c) \tag{3.2}$$

onde x_{ij}^r são aos valores dos atributos numéricos e x_{ij}^c os valores dos atributos categóricos. Como explicado por Dey e Ahmad (2007), q_{lj}^r representa a média aritmética para o atributo numérico r de todas as instâncias agrupadas no *cluster* l e q_{lj}^c a moda do atributo categórico c de todas as instâncias agrupadas no *cluster* l. A quantidade de atributos numéricos é representada por m_r, enquanto a de atributos categóricos por m_c. Para os atributos categóricos, $\delta(p,q) = 0$ para $p \equiv q$ e $\delta(p,q) = 1$ para $p \neq q$. Por fim, y_l é um peso para os atributos categóricos do cluster l de modo a não favorecer qualquer tipo de atributo (HUANG, 1997).

3.5.2 Funcionamento do algoritmo

O algoritmo de *clustering* K-Prototype de Huang (1997) busca minimizar a função de custo apresentada na equação (3.1), e sua base é mesclar a técnica de agrupar conjuntos de dados com variáveis numéricas do algoritmo K-means, com a técnica de agrupar dados com variáveis categóricas através da medida de dissimilaridade entre os objetos do algoritmo K-modes. O algoritmo é, na prática, mais útil que estes dois outros, pois a maior parte dos conjuntos de dados encontrados no mundo real são do tipo misto (HUANG, 1998).

Segundo o autor, os seguintes passos descrevem o seu funcionamento:

1. Selecionar uma quantidade k de protótipos iniciais de um conjunto de dados X, um para cada *cluster*.
2. Alocar cada objeto de X no *cluster* onde o protótipo seja o mais próximo a ele, de acordo com a medida de similaridade descrita na Equação (3.2). Depois de cada alocação, atualizar o protótipo do *cluster*.
3. Depois que todos os objetos forem alocados, testar novamente a similaridade deles em relação aos seus protótipos atuais. Se o protótipo mais próximo de um deles pertencer a outro *cluster*, realocar este objeto para este *cluster* e atualizar os protótipos de ambos.
4. Repetir o 3º passo até que nenhum objeto tenha sido trocado de *cluster* após um teste de ciclo completo de X.

Ainda segundo Huang (1997), o algoritmo é baseado em 3 processos: seleção inicial dos protótipos, alocação inicial e realocação. O primeiro deles seleciona randomicamente uma quantidade *k* de objetos como os protótipos iniciais para os *clusters*. Já no processo seguinte, descrito no algoritmo 1, a partir destes protótipos escolhidos, cada objeto é atribuído a um *cluster*, e o protótipo do cluster é atualizado.

Algoritmo 1: Processo de alocação inicial

```
inicialização;
for i=1 TO NumberOfObjects do
    Mindistance=Distance(X[i],O_prototypes[1])+gamma*Sigma(X[i],C_prototypes[1])
     for j=1 TO NumberOfClusters do
        distance=Distance(X[i],O_prototypes[j]+gamma*Sigma(X[i],C_prototypes[j])
         if distance<Mindistance then
            Mindistance=distance;
            cluster=j;
        end
    end
    Clustership[i]=cluster;
    ClusterCount[cluster] + 1;
    for j=1 To NumberOfNumericAttributes do
        SumInCluster[cluster,j]+X[i,j];
        O_prototypes[cluster,j]=SumInCluster[cluster,j]/ClusterCount[cluster];
    end
    for j=1 TO NumberOfCategoricAttributes do
        FrequencyInCluster[cluster,j,X[i,j]]+1;
        C_prototypes[cluster,j]=HighestFreq(FrequencyInCluster,cluster,j);
    end
end
```

Explicando o algoritmo 1, Huang (1997) define X[i] como representante do objeto *i*, enquanto X[i,j] o valor do atributo *j* para o objeto *i*. Duas variáveis foram utilizadas para armazenar os atributos dos protótipos dos *clusters*: O_prototypes[] para os valores numéricos e C_prototypes[] para os valores categóricos. O_prototypes[i,j] e C_prototypes[i,j] são, respectivamente, elementos do protótipo do *cluster i* para valores numéricos e categóricos. **Distance()** é a função do quadrado da distância Euclidiana e a função **Sigma()** é a implementação da função $\delta()$ na medida de similaridade descrita na Equação 3.2. Já Clustership[] armazena a associação de *cluster* de objetos e ClusterCount[] a quantidade de objetos nos *clusters*. SumInCluster[] soma os valores numéricos dos objetos nos *clusters*, além de ser utilizado para atualizar os atributos numéricos dos protótipos, enquanto a função FrequencyInCluster[] grava as frequência de diferentes

valores de atributos categóricos nos *clusters*. E por último, a função **HighestFreq()** atualiza os atributos categóricos dos protótipos, calculando a moda, ou seja, a classe mais frequente do *j-ésimo* atributo categórico no *cluster*.

O último processo, de realocação, é descrito no algoritmo 2, e é semelhante ao processo de alocação inicial, exceto pelo fato de que, após a realocação de um objeto, os protótipos dos *clusters* anteriores e atuais do objeto são atualizados. A variável *moves* grava a quantidade de objetos que mudaram de *clusters* no processo. O custo operacional do K-Prototype é *O((t+1)kn)*, onde *n* é o número de objetos, *k* o número de *clusters* e *t* o número de iterações do processo de realocação (HUANG, 1997).

Algoritmo 2: Processo de realocação

```
inicialização;
moves=0;
for i=1 TO NumberOfObjects do
    ...
    (Para encontrar o cluster cujo protótipo é o mais próximo ao objeto i, como no
     Algoritmo 1)
    ...
    if Clustership[i]<>cluster then
        moves+1;
        oldcluster=Clustership[i];
        ClusterCount[cluster]+1;
        ClusterCount[oldcluster]-1;
        for j=1 To NumberOfNumericAttributes do
            SumInCluster[cluster,j]+X[i,j];
            SumInCluster[oldcluster,j]-X[i,j];
            O_prototypes[cluster,j]=SumInCluster[cluster,j]/ClusterCount[cluster];
            O_prototypes[oldcluster,j]=SumInCluster[oldcluster,j]/ClusterCount[oldcluster];

        end
        for j=1 To NumberOfCategoricAttributes do
            FrequencyInCluster[cluster,j,X[i,j]]+1;
            FrequencyInCluster[oldcluster,j,X[i,j]]-1;
            C_prototypes[cluster,j]=HighestFreq(cluster,j);
            C_prototypes[oldcluster,j]=HighestFreq(oldcluster,j);
        end
    end
end
```

3.6 MÉTRICAS AVALIATIVAS PARA A TÉCNICA DE CLUSTERING

Após a correta implementação do algoritmo *K-prototype*, é preciso avaliar os resultados gerados. Para esta tarefa, existem métricas estatísticas que auxiliam a medir a similaridade dos dados agrupados em cada *clusters*. As próximas sub-seções irão abordar sobre duas destas métricas: o Índice de Rand Ajustado e o Índice de Jaccard.

3.6.1 Índice de Rand

Antes de abordar sobre o Índice de Rand Ajustado, é preciso compreender as noções acerca do Índice de Rand (1971). Esta medida busca um valor de similaridade entre duas partições (**R** e **Q**), onde a primeira será a referência **R** para a avaliação e a segunda é a que será avaliada, **Q**. Esta medida analisa pares de objetos contidos nos agrupamentos, de forma a verificar a qual *cluster* eles pertencem em cada um dos conjuntos de dados. O agrupamento de referência particiona os dados em k^* classes, enquanto o que será avaliado em k *clusters*. Quanto mais próximo do valor 1 for o resultado, mais semelhantes são os *clusters* analisados (VENDRAMIN et al., 2010). O cálculo para este índice é definido por:

$$I_R(\mathbf{R},\mathbf{Q}) = \frac{a+d}{a+b+c+d} \tag{3.3}$$

onde, segundo Vendramin et al. (2010):

- a = Número de pares de objetos do conjunto de dados que pertencem à mesma classe em **R** e o mesmo *cluster* em **Q**;
- b = Número de pares de objetos do conjunto de dados que pertencem à mesma classe em **R** e a *clusters* diferentes em **Q**;
- c = Número de pares de objetos do conjunto de dados que pertencem a diferentes classes em **R** e ao mesmo *cluster* em **Q**;
- d = Número de pares de objetos do conjunto de dados que pertencem a diferentes classes em **R** e a diferentes *clusters* em **Q**.

Outras observações foram feitas pelo autor:

1. $I_R \in [0,1]$

2. $I_R = 0$ se, **Q** é completamente inconsistente, ou seja, $a = d = 0$
3. $I_R = 1$ se, **Q** é exatamente igual a **R** , ou seja, $b = c = 0$ ($\mathbf{C} = \mathbf{R}$)

3.6.2 Índice de Rand Ajustado

Segundo Vendramin et al. (2010), uma das principais críticas ao Índice de Rand original é que o seu valor não é 0 caso as duas partições sejam aleatórias. Diante desta situação, Hubert e Arabie (1985) propuseram o Índice de Rand Ajustado, determinando um valor esperado para o Índice de Rand. O valor esperado para o Índice de Rand foi adaptado na seguinte equação:

$$E[I_R(\mathbf{C},\mathbf{R})] = \frac{(a+c)(a+b)}{a+b+c+d} \tag{3.4}$$

A partir disso, o Índice de Rand Ajustado é dado pela seguinte fórmula:

$$I_{Raj}(\mathbf{C},\mathbf{R}) = \frac{a - E[I_R(\mathbf{C},\mathbf{R})]}{\frac{((a+c)+(a+b)}{2} - E[I_R(\mathbf{C},\mathbf{R})]} \tag{3.5}$$

Os termos a, b, c, d possuem o mesmo significado quando utilizados na Subseção 3.6.1. O Índice de Rand Ajustado possui como resultados valores entre o intervalo de -1 e 1, onde o valor 1 é uma perfeita similaridade entre os dois conjuntos de dados observados. Se os pares de elementos pertencerem a uma partição em um conjunto de dados, mas pertencem a uma partição diferente no outro analisado, o valor do índice diminui (ALBUQUERQUE et al., 2016). Já os valores próximos a 0 definem a aleatoriedade entre eles, onde, quanto maior a aleatoriedade, mais negativos são os valores.

3.6.3 Índice de Jaccard

O índice de Jaccard também mede a similaridade entre duas partições, comparando os itens exclusivos contidos em cada um e os itens em comum entre eles. Também conhecido como coeficiente de similaridade de Jaccard, esta medida é definida pela quantidade de elementos da interseção entre as variáveis comparadas, dividida pela quantidade de elementos da sua união. Os valores resultantes para este índice variam entre 0 e 1, onde, quanto mais próximo de 1, mais semelhantes são os dois conjuntos (ALBUQUERQUE et al., 2016). O cálculo desta medida é

definida por:

$$S_Jac = \frac{|A \cap B|}{|A \cup B|} = \frac{P(A \cap B)}{P(A) + P(B) - P(A \cap B)} = \frac{p1p2}{p1 + p2 - p1p2} = \frac{a}{a+b+c} \tag{3.6}$$

onde $P(A)$ e $P(B)$ representam os dois conjuntos analisados. Após a operação de interseção, $p1p2$ representa o seu resultado, bem como $p1$ e $p2$ representam os resultados dos conjuntos separados. Por fim, a, b, c continuam com o mesmo significado quando utilizados no Índice de Rand, na Subseção 3.6.1.

4 O DESENVOLVIMENTO DAS ETAPAS PARA A DESCOBERTA DE CONHECIMENTO

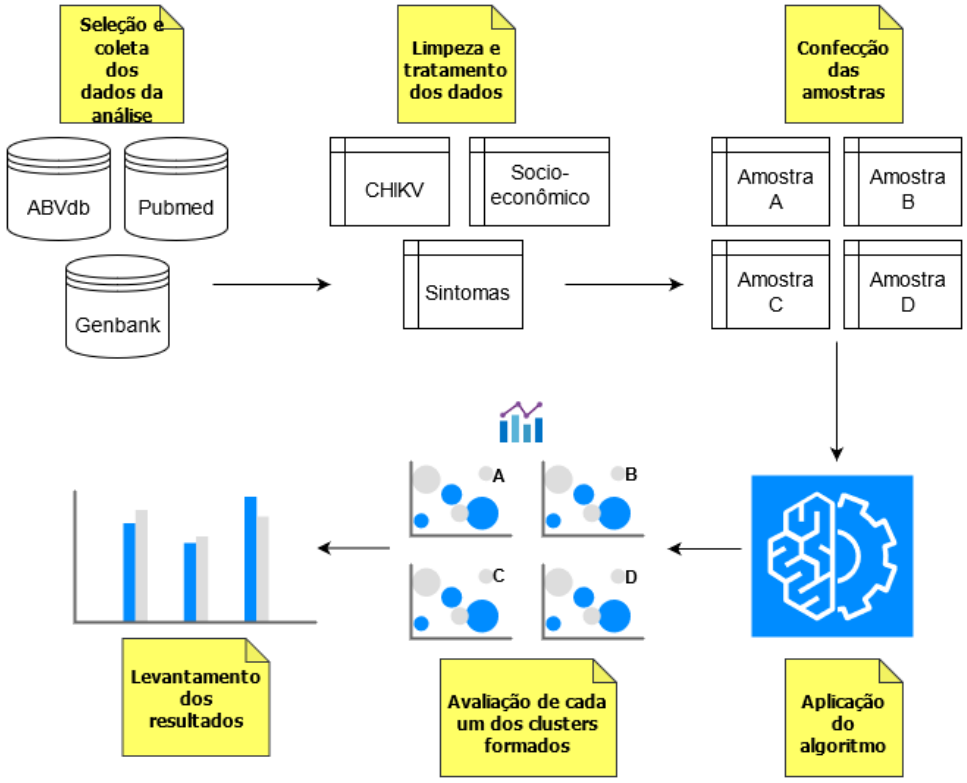
Para construir um projeto de pesquisa de *Data Science* com resultados confiáveis, é importante seguir todos os passos do KDD corretamente, de forma a evitar erros que possam prejudicar tanto o processo, como o resultado final. Este capítulo mostra como todo o processo foi desenvolvido, desde a entrada, quando os conjuntos de dados foram extraídos de suas fontes em seus formatos originais, até a saída, onde eles foram unificados, agrupados e avaliados para a busca por novos conhecimentos. A base de todo o projeto seguiu uma metodologia de pesquisa quantitativa, que será melhor detalhada a seguir.

4.1 METODOLOGIA DE PESQUISA QUANTITATIVA

A abordagem do problema a ser solucionado foi baseada na metodologia de pesquisa quantitativa Creswell et al. (2007), que consiste em fazer uso de recursos estatísticos para coletar, analisar e interpretar dados, além de mensurar os resultados obtidos. É uma abordagem propícia ao desenvolvimento do KDD, pois os resultados levantados precisam ser avaliados, de forma que se tenha a ciência que foi o melhor encontrado na análise, e com a possibilidade de comprovar que todo o processo foi realizado de maneira correta. A estatística no projeto foi fortemente empregada com o uso de tabelas e quadros para representar os resultados, o uso de métricas para avaliar o algoritmo e o desenho de gráficos para dimensionar os conjuntos de dados.

Os dados são os elementos principais, e para a obtenção de resultados corretos, é preciso adaptá-los de forma que a aplicação do algoritmo de *clustering* ocorra sem problema algum. Também, é importante que todos os dados adquiridos sejam fidedignos, e disponibilizados por órgãos que atestem a sua confiabilidade. A partir destas duas preposições, a análise realizada pode servir como base para entender os aspectos do vírus chikungunya em casos reais de infecção em seres humanos, relacionando-os com fatores socioeconômicos, das regiões onde os mesmos foram confirmados, e clínicos, referente aos sintomas e as características dos pacientes. Todo o processo implementado para estudar estes dados pode ser visto na Figura 3.

Figura 3 – Diagrama do processo de análise da Chikungunya com o uso de Data Science.



Fonte – O autor.

Explicando o processo da Figura 3, inicialmente os conjuntos de dados foram selecionados e coletados nas suas bases. Após isso, eles foram tratados e formatados, para remover possíveis inconsistências, variáveis não utilizadas e valores discrepantes. Após esta etapa, os dados foram transformados, para a criação das amostras que serão utilizadas no projeto. Cada amostra abordou atributos específicos a serem observados nos conjuntos de dados, preservando o máximo de informações possíveis. Logo após, a implementação dos algoritmos de *clustering* foi realizada, analisando cada uma destas amostras, seguindo um plano de testes para a execução e armazenamento dos resultados. Com eles em mãos, foi possível aplicar as métricas para qualificar a similaridade entre os agrupamentos realizados, e escolher o melhor em cada uma das amostras. Por fim, estes melhores resultados foram representados graficamente, para um levantamento do que foi obtido, e para identificar as relações encontradas. Todo este desenvolvimento será visto detalhadamente neste capítulo, além das ferramentas que auxiliaram durante o processo.

4.2 CONFIGURAÇÕES DO AMBIENTE DE DESENVOLVIMENTO

Algumas ferramentas computacionais foram utilizadas para auxiliar nas etapas deste projeto. A maior parte dos recursos foram provenientes de um ambiente virtual disponibilizado pelo Google Compute Engine, chamado Google Colab. Por ser uma *Integrated Development Environment* (IDE) - no português, Ambiente de Desenvolvimento Integrado - o Google Colab permite escrever e rodar códigos na linguagem Python direto do navegador web, utilizando os recursos de hardware e software do próprio Google. Ela possui todas as ferramentas de outras IDE's de Python, possibilitando importar bibliotecas e conjuntos de dados dos mais diversos formatos, criar diversos ambientes interativos (chamados de notebooks), e compartilhar na nuvem com facilidade. As informações disponibilizadas pelo Google referente às configurações do ambiente hospedado são:

- Conexão ao "de back-end do Google Compute Engine em Python 3";
- Memória Random Access Memory (RAM) de 12.72 Gigabytes (GB);
- Memória em Disco de 107.77 GB.

Uma das linguagens de programação mais indicadas para o trabalho com *Data Science* é o Python, pois possui um grande suporte da comunidade de especialistas e contém diversos recursos de análise de dados, além de visualização de resultados. Como já foi dito, o Google Colab é propício a utilizar esta linguagem para escrever e rodar códigos, e a versão utilizada pela IDE é o Python 3. Para este projeto, essa versão ofereceu suporte para o uso das bibliotecas pertinentes ao contexto do projeto, que possuem seus nomes e descrições detalhadas no Quadro 1.

Na próxima seção serão detalhados todos os procedimentos realizados referente à primeira etapa do processo de KDD, que consiste na seleção inicial dos dados em diversas fontes, que servirão para o estudo e análise de casos confirmados do vírus chikungunya.

4.3 SELEÇÃO DE DADOS

Além do *ABVdb*, outras duas fontes principais forneceram os dados necessários para o desenvolvimento do projeto. Uma delas foi o banco de dados do *United Nations Development Programme* (UNDP), no português, Programa de Desenvolvimento das Nações Unidas, que disponibiliza publicamente uma gama de informações socioeconômicas, demográficas, sustentáveis,

Quadro 1 – Bibliotecas da linguagem Python utilizadas no projeto.

Biblioteca	**Funcionalidade**
Pandas	Oferece ferramentas para análise e manipulação de conjuntos de dados dos tipos numéricos e categóricos.
Numpy	Possibilita a computação numérica com o uso da linguagem Python.
Pyplot	Permite a criação de diversos tipos de gráficos customizáveis.
Seaborn	Permite gerar gráficos estatísticos sobre um conjunto de dados.
Scikit-learn	Possui ferramentas para análise preditiva de conjuntos de dados, além de métricas avaliativas.
Kmodes	Auxilia na implementação dos algoritmos.

Fonte – O autor.

dentre outras, referente ao Relatório de Desenvolvimento Humano de diversos países ao redor do mundo. Elas são obtidas através de agências de dados internacionais com mandato, recurso e experiência para coletá-los sobre indicadores específicos e podem ser acessadas através do site: http://hdr.undp.org/en/data.

A última fonte escolhida foi o GenBank que é uma aglomerado de base de dados de sequências genéticas, disponíveis publicamente em: https://www.ncbi.nlm.nih.gov/genbank/. As utilizadas neste projeto foram a *Nucleotide* e *PubMed*. Após a busca através de um código de acesso, o *GenBank* exibe informações sobre como a sequência foi descoberta, disponibilizando artigos para um maior detalhamento, de acordo com a base de dados escolhida. Os artigos que foram selecionados para a identificação dos sintomas serviram como base para o desenvolvimento do ABVdb.

As variáveis socioeconômicas utilizadas no projeto foram selecionadas, baseadas em 3 critérios adequados para as possíveis análises:

- A variável socioeconômica deve ter alguma relação com saúde ou qualidade de vida.
- Ela deve conter registros a partir do ano de 2004, o qual foi o início da expansão do surto da chikungunya.
- A variável precisa estar disponível publicamente, de forma que, a partir de 2004, a maioria dos seus dados estejam dispostos constantemente ao longo dos anos.

Devido a estes critérios, o UNDP forneceu as variáveis necessárias, que compreenderam o período entre 2004 e 2017. Já para a base do ABVdb, foram selecionados todos

os dados do vírus chikungunya, sem qualquer outro critério. Por fim, os dados oriundos dos artigos encontrados no GenBank, como foram selecionados manualmente, já estão no formato de utilização. O Quadro 2 exibe quais as variáveis que foram selecionadas para o desenvolvimento do projeto.

Quadro 2 – Variáveis selecionadas para o projeto em cada uma das bases de dados.

Fonte de Origem	**Variáveis selecionadas**
ABVdb	Os códigos únicos dos casos confirmados de infeção pelo vírus chikungunya, os países onde os casos foram confirmados, os anos registrados, os genótipos identificados, a idade e o sexo dos indivíduos.
UNDP	Os países de registro, os anos, o IDH para cada um deles, a porcentagem da população que tem acesso a saneamento básico para cada um deles e a porcentagem do PIB gasto na área de saúde.
GenBank	Os códigos únicos dos casos confirmados de infecção pelo vírus chikungunya e os sintomas registrados.

Fonte – O autor.

Na confecção inicial do banco de dados ABVdb não foi incluso a relação entre a ocorrência da infecção pelo vírus da chikungunya e os sintomas registrados naquele caso. Estes sintomas, além de servirem para a análise final que será desenvolvida, serão adicionados em uma atualização do ABVdb para acesso ao público. Para isso, foi necessário realizar um levantamento destes sintomas confirmados decorrentes do vírus nos artigos. As primeiras informações oficiais foram obtidas através da Organização Mundial de Saúde (OMS) que dispõe de informações ao público, OMS (2017). O Quadro 3 detalha os 7 sintomas principais divulgados pelo órgão.

Para identificar o restante dos sintomas, foi preciso buscar as informações nos registros dos artigos encontrados no GenBank, que serviram como base para a construção do ABVdb. Ao todo, 48 artigos foram lidos para identificação dos sintomas do vírus, e dentre eles, em 22 foram encontrados os registros dos problemas de saúde constatados nas pessoas infectadas pelo vírus da chikungunya. Nestes artigos, para cada sintoma encontrado, foi anotado também o código de acesso referente ao caso identificado no banco. Essas informações foram verificadas ao analisar: as árvores filogenéticas documentadas em AbuBakar et al. (2007), Peyrefitte et al. (2008), Kumar et al. (2008), Ng et al. (2009), Lim et al. (2009), Parreira et al. (2014), Hisamuddin et al. (2018), Bessaud et al. (2006) e Shu et al. (2008); como detalhe no próprio corpo do texto em Parola et al. (2006), Fusco et al. (2006), Pistone et al. (2009), Lewthwaite et al. (2009), Zheng et al. (2010) e Leo et al. (2009); e em tabelas divulgadas nos textos de Pastorino et al.

Quadro 3 – Sintomas e variações da chikungunya descritos pela OMS e encontrados nos artigos bases do GenBank.

Fonte de Origem	**Sintomas**
OMS	*Fatigue, Fever, Headache, Joint Pain, Muscle Pain, Nausea e Rash*
GenBank	*Abdominal Pain, Anorexia, Arthalgia,Asthenia, Backache, Bilateral Conjunctivitis, Bleeding, Body Ache, Cardiomegaly, Chills, Conjunctivitis, Cough, Dermal Lesions, Diarrehea, Disabling Pain, Epistaxis, Erythematous Exanthema, Erythematous Rashes, Eye Pain, Facial Swelling, Fatigue, Fever, Headache, Hepatitis, High Fever, Hyperaesthesia, Hyperpigmentation, Joint Pain, Lack os Appetite, Low Trombocyte Counts, Lumbar Pain, Macopapular Rash, Macular Rash, Muscle Pain, Muscle Spasm, Myalgia, Myocarditis, Nausea, Pancytopenia, Pedal Swelling, Prostation, Pulmonary Congestion, Rash, Rheumatoid Arthritis, Seizure, Severe Arthalgia, Severe Asthenia, Severe Bilateral Arthalgia, Shortness of Breath, Tenosynovitis, Transient Macular Rash, Unwell e Wrist Pain.*

Fonte – O autor.

(2004), Yergolkar et al. (2006), Edwards et al. (2007), Chahar et al. (2009), Sam et al. (2009), Santhosh et al. (2009) e Liu et al. (2017). Ao todo, foram encontrados 54 sintomas que também são mostrados no Quadro 3.

A Tabela 1 mostra a quantidade de linhas e colunas dos *dataframes* adquiridos nas bases de dados do ABVdb e do UNDP, e do *dataframe* confeccionado com as informações do GenBank. Após esta etapa de seleção e coleta, foi preciso tratar todos os dados, de forma a viabilizar a continuidade do processo. Na próxima seção será abordado todo o processo de limpeza destes dados selecionados, o que corresponde à segunda etapa do processo de KDD descrito por Fayyad et al. (1996).

Tabela 1 – Quantidade de linhas e colunas dos *dataframes* em suas versões iniciais.

Dataframe	**Total de Linhas**	**Total de Colunas**
ABVdb	3.470	11
IDH	208	59
ASB	214	37
GSPIB	265	15
Genbank	866	2

Fonte – O autor.

4.4 PRÉ-PROCESSAMENTO DE DADOS

Para corrigir os erros nos conjuntos de dados baixados, algumas tarefas precisaram ser realizadas. A primeira parte da limpeza foi realizar algumas correções manuais nos arquivos, de forma a prepará-los adequadamente para a utilização deles nas ferramentas que auxiliaram no projeto. Após isso, o foco foi na exclusão das colunas que não foram utilizadas no projeto e por fim, os últimos ajustes foram feitos em cada um dos conjuntos de dados. Para essa etapa de pré-processamento de dados, somente foram tratados os oriundos do ABVdb e do UNDP, já que os dados do *GenBank* foram criados no formato correto, sem a necessidade de formatação. No decorrer desta seção será detalhado como foi feito todo o processo de limpeza e formatação.

O primeiro conjunto de dados a ser formatado foi o obtido através do ABVdb, representado por um arquivo único baixado no formato *XLS*, que é uma extensão para os nomes de arquivos exclusivos para planilhas Microsoft Excel. A Figura 4 representa a organização original dele, visualizado no próprio Microsoft Excel, contendo o título da planilha, suas colunas que caracterizam as variáveis, e as linhas com os valores. Antes de manipulá-lo com o Python, foi preciso excluir a primeira linha de título, pois se ela for carregada junto com os dados, atrapalhava na identificação dos cabeçalhos das colunas. Também, como a análise foi somente com os registros do vírus chikungunya confirmados em seres humanos, as linhas que contém na variável *Host* os valores referentes aos genomas de mosquitos e outras espécies foram removidos. Após isso, o arquivo foi exportado para o mesmo formato.

A biblioteca *Pandas* do Python é especializada na tarefa de manipular e analisar conjuntos de dados, com funções e estruturas específicas para auxiliar nestas tarefas. Após a sua importação, o arquivo do ABVdb pôde ser carregado corretamente no Google Colab, e atribuído à uma variável característica de conjuntos de dados, do tipo *dataframe*, para ser tratado. Com isso, a sua representação original foi preservada, armazenando linhas e colunas corretamente. O próximo passo então foi excluir as colunas que não serão utilizadas em nenhuma das próximas etapas do projeto. Após a identificação, as variáveis *Virus*, *Product*, *Size*, e *Clinic* foram removidas. A Figura 5 contém a representação do *dataframe*, exibidos na IDE, após a finalização das suas correções.

Após o tratamento dos dados referente ao vírus, os próximos conjuntos de dados formatados foram os adquiridos do UNDP. Foram baixados três arquivos diferentes para representar

Figura 4 – Visualização do conjunto de dados do arquivo baixado do ABVdb, com os 20 primeiros registros.

	A	B	C	D	E	F	G	H	I	J	K
1	Planilha Arbovirusdb										
2	**Acession**	**Virus**	**Genotype**	**Product**	**Host**	**Country**	**Date**	**Size**	**Clinic**	**Gen**	**Age**
3	AB455493	Chikungunya	East-Central-South-African	5'UTR-nsP1-nsP2-nsP3-nsP4-C-E3-E2-6K-E1-3'UTR	Hs	Sri Lanka	2006	11829	CHF	F	59
4	AB455494	Chikungunya	East-Central-South-African	5'UTR-nsP1-nsP2-nsP3-nsP4-C-E3-E2-6K-E1-3'UTR	Hs	Sri Lanka	2006	11829	CHF	F	59
5	AB857730	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
6	AB857731	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
7	AB857732	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
8	AB857733	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
9	AB857734	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
10	AB857735	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
11	AB857736	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
12	AB857737	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
13	AB857738	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
14	AB857739	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
15	AB857740	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
16	AB857741	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
17	AB857742	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
18	AB857743	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
19	AB857744	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A
20	AB857745	Chikungunya	East-Central-South-African	C-E3-E2-6K-E1-3'UTR	Hs	Thailand	2010	3747	CHF	N/A	N/A

Fonte – O autor.

Figura 5 – Exibição dos primeiros e últimos registros do *dataframe* que representa o arquivo do ABVdb já formatado.

	Acession	Genotype	Host	Country	Date	Gender	Age
0	AB455493	East-Central-South-African	Hs	Sri Lanka	2006.0	F	59.0
1	AB455494	East-Central-South-African	Hs	Sri Lanka	2006.0	F	59.0
2	AB678677	Asian and Caribbean	NaN	Indonesia	2010.0	NaN	NaN
3	AB678678	Asian and Caribbean	NaN	Indonesia	2010.0	NaN	NaN
4	AB678679	Asian and Caribbean	NaN	Indonesia	2010.0	NaN	NaN
...	...	...	...	...	...	...	...
3463	MN114342	East-Central-South-African	AEae	Thailand	2014.0	NaN	NaN
3464	MN114343	East-Central-South-African	AEae	Thailand	2014.0	NaN	NaN
3465	MN114344	East-Central-South-African	AEae	Thailand	2014.0	NaN	NaN
3466	MN114345	East-Central-South-African	AEae	Thailand	2014.0	NaN	NaN
3467	MN114346	East-Central-South-African	AEae	Thailand	2014.0	NaN	NaN

3468 rows × 7 columns

Fonte – O autor.

os valores do IDH, do Acesso ao Saneamento Básico (ASB) e do Gastos na Saúde com o PIB (GSPIB). Todos eles possuem estruturas de colunas iguais, compostas por uma inicial contendo o ranking do país em relação à variável em questão, logo após uma contendo o nome dele, e as outras em sequência contendo os anos nos quais o valor da variável está informada na linha. A única diferença é que o último foi baixado em um formato XLS, enquanto os outros dois estão em *Comma Separated Values* (CSV), no português, valores separados por vírgula. A Figura 6 exemplifica esta estrutura contida na versão original do arquivo que contém os dados do IDH.

Figura 6 – Visualização dos primeiros registros do arquivo original que representa o conjunto de dados do IDH.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Human Development Index (HDI)											
2	HDI Rank (2018),"Country","1990","","1991","","1992","","1993","","1994","","1995","","1996","","1997","","1998","","1999","",											
3	170,"Afghanistan","0.298","","0.304","","0.312","","0.308","","0.303","","0.327","","0.331","","0.335","","0.339","","0.343","","0											
4	69,"Albania","0.644","","0.625","","0.608","","0.611","","0.617","","0.629","","0.639","","0.639","","0.649","","0.660","","0.667",											
5	82,"Algeria","0.578","","0.582","","0.589","","0.593","","0.597","","0.602","","0.610","","0.619","","0.629","","0.638","","0.646",											
6	36,"Andorra","..","","..","","..","","..","","..","","..","","..","","..","","..","","..","","0.759","","0.767","","0.780","","0.820","","0.8											
7	149,"Angola","..","","..","","..","","..","","..","","..","","..","","..","","..","","0.384","","0.394","","0.404","","0.419","","0.428","",											
8	74,"Antigua and Barbuda","..","","..","","..","","..","","..","","..","","..","","..","","..","","..","","..","","..","","..","","..","","..","",											
9	48,"Argentina","0.707","","0.714","","0.719","","0.725","","0.729","","0.731","","0.738","","0.746","","0.752","","0.763","","0.77											
10	81,"Armenia","0.633","","0.629","","0.585","","0.590","","0.600","","0.604","","0.614","","0.625","","0.637","","0.644","","0.649											
11	6,"Australia","0.866","","0.867","","0.868","","0.872","","0.875","","0.883","","0.886","","0.889","","0.892","","0.895","","0.898"											
12	20,"Austria","0.795","","0.799","","0.805","","0.809","","0.813","","0.817","","0.820","","0.824","","0.828","","0.834","","0.838",											
13	87,"Azerbaijan","..","","..","","..","","..","","..","","0.612","","0.612","","0.618","","0.627","","0.634","","0.641","","0.649","","0.											
14	60,"Bahamas","..","","..","","..","","..","","..","","..","","..","","..","","..","","..","","0.787","","0.788","","0.790","","0.789","","0.											
15	45,"Bahrain","0.736","","0.755","","0.756","","0.764","","0.768","","0.775","","0.779","","0.781","","0.784","","0.786","","0.792",											
16	135,"Bangladesh","0.388","","0.395","","0.403","","0.411","","0.419","","0.427","","0.436","","0.444","","0.453","","0.462","","0											
17	56,"Barbados","0.732","","0.733","","0.733","","0.737","","0.743","","0.747","","0.751","","0.757","","0.756","","0.764","","0.771											
18	50,"Belarus","..","","..","","..","","..","","..","","0.656","","0.661","","0.667","","0.671","","0.676","","0.682","","0.689","","0.696											
19	17,"Belgium","0.806","","0.810","","0.825","","0.838","","0.845","","0.851","","0.857","","0.862","","0.866","","0.868","","0.873"											
20	103,"Belize","0.613","","0.618","","0.624","","0.627","","0.627","","0.627","","0.627","","0.630","","0.631","","0.636","","0.643",											

Fonte – O autor.

Por serem semelhantes e oriundas da mesma base de dados, estes conjuntos de dados possuem três problemas principais referente à estrutura na qual os valores de suas linhas e colunas estão dispostos. São eles:

1. Os nomes dos países são diferentes se comparados ao conjunto de dados da chikungunya. Por exemplo, enquanto um dos valores da variável *Country* no arquivo do ABVdb encontra-se como "USA", nos oriundos do UNDP o valor é "United States". Este problema atrapalharia na criação das amostras, já que o país é uma das variáveis que são levadas em consideração ao unir os *dataframes* das duas bases.
2. Existem colunas nos conjuntos de dados que não serão utilizadas. É preciso deixar o *dataframe* somente com com a estrutura país, ano e valor socioeconômico, e qualquer outra coluna que não esteja nesse escopo precisa ser removida.
3. Os valores dos anos são os próprios índices das coluna. Com esta estrutura, a criação das amostras com estas variáveis também ficaria comprometida, sendo necessário um formato mais adequado para executar esta tarefa.

Para corrigir o primeiro problema, referente aos nomes dos países nos conjuntos de dados, foi gerado um arquivo auxiliar contento a listagem de todos os países na forma como estão declarados no ABVdb, para corrigir os contidos nos três arquivos. O processo de substituição é simples, somente copiando os valores do primeiro arquivo e colando nos outros três. Por fim, eles foram salvos e o primeiro problema foi corrigido.

Com a nomenclatura dos países corretas, antes de corrigir o segundo problema foi necessário realizar algumas correções manuais nos três arquivos. São elas:

- Exclusão da linha de título nos arquivos, pelo mesmo motivo que foi excluso nos dados do ABVdb.
- Correção dos valores numéricos nulos, onde originalmente ele vieram representados por "..", que faz com que os valores não sejam reconhecidos como números. Para consertá-los, foi realizada a substituição pelo número 0.
- Os três conjuntos de dados possuem valores regionais de suas variáveis, sendo que o escopo do projeto é com os países somente. Estes valores se encontravam nas últimas linhas dos arquivos, e foram removidos.
- Os arquivos do IDH e do ASB foram convertidos para o formato *XLS*, para remover problemas de inconsistência na leitura das linhas e colunas dos conjuntos de dados. O arquivo do GSPIB não precisou desta conversão, pois já veio no formato correto.

Após estas correções manuais, foi possível resolver o segundo problema descrito de estrutura. Após o carregamento dos dados em seus respectivos *dataframes*, foram identificadas as colunas em branco, como exemplificado no *dataframe* do IDH na Figura 7, além também de uma coluna com o ranking do IDH dos países nos três conjuntos de dados. A biblioteca *pandas* foi novamente utilizada, de forma a automatizar a tarefa de excluí-las, já que não serão utilizadas em nenhuma das próximas etapas do projeto.

Figura 7 – Visualização dos primeiros e últimos registros do *dataframe* com os dados do IDH até 1995.

	HDI Rank (2018)	Country	1990	Unnamed: 3	1991	Unnamed: 5	1992	Unnamed: 7	1993	Unnamed: 9	1994	Unnamed: 11	1995
0	170	Afghanistan	0.298	NaN	0.304	NaN	0.312	NaN	0.308	NaN	0.303	NaN	0.327
1	69	Albania	0.644	NaN	0.625	NaN	0.608	NaN	0.611	NaN	0.617	NaN	0.629
2	82	Algeria	0.578	NaN	0.582	NaN	0.589	NaN	0.593	NaN	0.597	NaN	0.602
3	36	Andorra	0.000	NaN	0.000	NaN	0.000	NaN	0.000	NaN	0.000	NaN	0.000
4	149	Angola	0.000	NaN	0.000	NaN	0.000	NaN	0.000	NaN	0.000	NaN	0.000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
184	96	Venezuela	0.638	NaN	0.648	NaN	0.654	NaN	0.656	NaN	0.657	NaN	0.660
185	118	Viet Nam	0.475	NaN	0.484	NaN	0.496	NaN	0.506	NaN	0.517	NaN	0.529
186	177	Yemen	0.392	NaN	0.396	NaN	0.395	NaN	0.398	NaN	0.398	NaN	0.393
187	143	Zambia	0.424	NaN	0.421	NaN	0.420	NaN	0.422	NaN	0.418	NaN	0.419
188	150	Zimbabwe	0.498	NaN	0.500	NaN	0.485	NaN	0.480	NaN	0.478	NaN	0.472

189 rows × 59 columns

Fonte – O autor.

A Figura 8 mostra o resultado após a exclusão das colunas do IDH. O mesmo foi feito para ASB e GSPIB, e com isso, os *dataframes* ficaram prontos para a resolução do último problema. Na figura pode-se observar que, além da coluna *Country*, todos os anos estão em sequência nos cabeçalhos. Porém, esta estrutura atrapalha tanto na análise dos dados que não serão utilizados, quanto na criação das amostras para a utilização dos algoritmos. Para corrigir esta situação, foi preciso converter o formato atual, de forma que os valores dos países foram replicados, e os valores dos anos e da variável socioeconômica em questão ficasse enfileirados, de forma a se ter uma melhor representação do conjunto de dados.

Figura 8 – Exibição dos primeiros e últimos registros do *dataframe* do IDH entre os anos de 1990 e 2004, após a remoção das colunas.

	Country	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
0	Afghanistan	0.298	0.304	0.312	0.308	0.303	0.327	0.331	0.335	0.339	0.343	0.345	0.347	0.378	0.387	0.400
1	Albania	0.644	0.625	0.608	0.611	0.617	0.629	0.639	0.639	0.649	0.660	0.667	0.673	0.680	0.687	0.692
2	Algeria	0.578	0.582	0.589	0.593	0.597	0.602	0.610	0.619	0.629	0.638	0.646	0.655	0.666	0.676	0.685
3	Andorra	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.759	0.767	0.780	0.820	0.826
4	Angola	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.384	0.394	0.404	0.419	0.428	0.440
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
184	Venezuela	0.638	0.648	0.654	0.656	0.657	0.660	0.662	0.666	0.668	0.670	0.672	0.680	0.688	0.687	0.700
185	Viet Nam	0.475	0.484	0.496	0.506	0.517	0.529	0.540	0.539	0.559	0.566	0.578	0.586	0.594	0.603	0.612
186	Yemen	0.392	0.396	0.395	0.398	0.398	0.393	0.408	0.418	0.430	0.423	0.432	0.449	0.456	0.464	0.471
187	Zambia	0.424	0.421	0.420	0.422	0.418	0.419	0.419	0.420	0.419	0.424	0.428	0.436	0.445	0.455	0.464
188	Zimbabwe	0.498	0.500	0.485	0.480	0.478	0.472	0.471	0.466	0.461	0.457	0.452	0.453	0.444	0.430	0.427

189 rows × 30 columns

Fonte – O autor.

A Figura 9 ilustra o resultado final da estrutura do conjunto de dados do IDH, após todas as correções realizadas. O mesmo foi aplicado para o ASB e para o GSPIB. Por fim, o conjutno de dados oriundo do Genbank não precisou de nenhuma limpeza, pois, no processo de sua criação, a sua estrutura já ficou pronta para uso.

A Tabela 2 mostra a quantidade de linhas e colunas de cada um dos *dataframes*, nesta etapa de pré-processamento dos dados. O próximo passo foi abordar as linhas, e remover todos os dados discrepantes, pois, se utilizados, eles atrapalharão na aplicação do algoritmo. Na próxima seção será detalhada como foi realizada a identificação destes valores nos conjuntos de dados utilizados no projeto.

Figura 9 – Visualização da versão final do *dataframe* do IDH com os primeiros e últimos registros.

	Country	Date	Value
0	Afghanistan	1990	0.298
1	Afghanistan	1991	0.304
2	Afghanistan	1992	0.312
3	Afghanistan	1993	0.308
4	Afghanistan	1994	0.303
...	...	...	...
5476	Zimbabwe	2014	0.537
5477	Zimbabwe	2015	0.544
5478	Zimbabwe	2016	0.549
5479	Zimbabwe	2017	0.553
5480	Zimbabwe	2018	0.563

5481 rows × 3 columns

Fonte – O autor.

Tabela 2 – Quantidade de linhas e colunas em cada *dataframe* após o pré-processamento dos dados.

Dataframe	**Total de Linhas**	**Total de Colunas**
ABVdb	3.232	7
IDH	5.481	3
ASB	3.510	3
GSPIB	3.038	3
Genbank	866	2

Fonte – O autor.

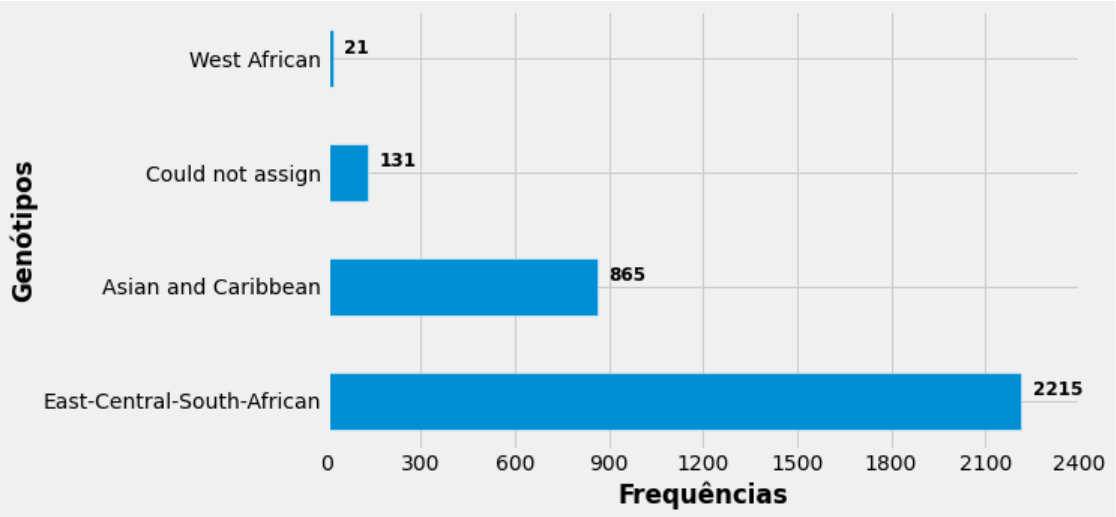
4.5 ANÁLISE EXPLORATÓRIA

Para a correta criação das amostras utilizadas no projeto, foi preciso analisar a disposição e o conteúdo dos valores nos conjuntos de dados que representam o vírus e as variáveis socioeconômicas. Para isso, uma análise exploratória foi feita, de forma a identificar quais as linhas que continham valores que atrapalhariam nas próximas etapas a serem executadas. A utilização de gráficos auxiliou nesta tarefa, que foi realizada com o auxílio das bibliotecas matplotlib e pytlot do Python no Google Colab.

O primeiro *dataframe* investigado foi o doABVdb. Na Seção 4.4, as colunas que não seriam utilizadas na análise foram removidas, e o ideal seria excluir todos os valores nulos ou vazios, para não atrapalharem nos resultados. Porém, na etapa de transformação deste projeto, quatro amostras serão criadas, e nem todas as colunas do conjunto de dados em questão serão utilizadas em todas elas. Remover cada uma das linhas que contém estes valores presentes acarretaria na perda de dados importantes para algumas destas amostras. Baseado nisso, foram analisadas quais variáveis possuiam valores que atrapalhariam na análise. Dentre as colunas

presentes, eles foram encontrados em duas: a *Genotype* e a *Gender*. Referente a esta primeira primeira, a Figura 10 mostra um gráfico de barra horizontal contendo a distribuição de frequência dos genótipos contidos no conjunto de dados.

Figura 10 – Gráfico de barra horizontal com as frequências dos genótipos no conjunto de dados do ABVdb.



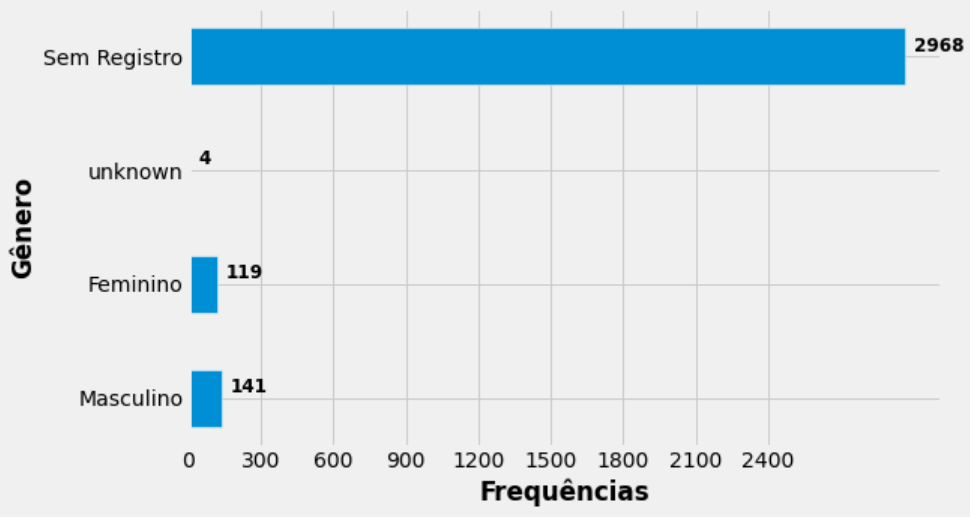
Fonte – O autor.

Pode-se observar neste gráfico que esta variável possui 131 valores iguais à "*Could not assign*", que não representam nenhum dos genótipos do vírus. Como não fazem parte do escopo do projeto, estes dados foram removidos do *dataframe*. Já referente à coluna *Gender*, foram encontradas 4 linhas que possuiam o valor "*unknown*", como pode ser visto na Figura 11. Estes valores equivalem à não categorização do gênero da pessoa infectada pelo vírus, portanto, também foram excluídos. Importante destacar também que existem 2.968 linhas sem registros de gênero dos indivíduos, mas, como este atributo não será abordado em todas as amostras, elas não foram excluídas neste momento, o que será visto na Seção 4.6.

Após o tratamento dos dados do ABVdb, foram analisados os que representam as variáveis socioeconômicas do UNDP. Estes dados também possuem valores nulos na sua distribuição de dados, que estão representados por 0. Eles poderiam ser aproveitados para a análise caso realmente significassem uma quantificação, como por exemplo, um país com um IDH muito ruim, uma população que não tenha acesso a saneamento básico, ou uma nação que gaste 0% do seu Produto Interno Bruto (PIB) com saúde. Mas como o valor 0 não possui essas representatividades, eles foram removidos. A começar pelo IDH, a Figura 12 mostra o histograma de frequência para os registros presentes neste conjunto de dados, onde cada instância

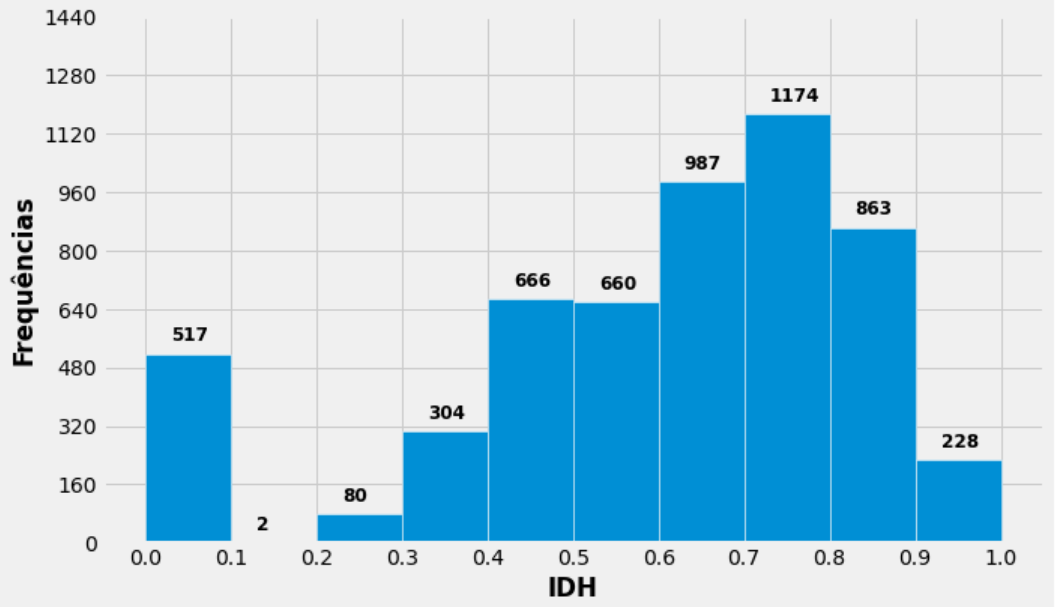
é um par que referencia o registro do país e do ano.

Figura 11 – Gráfico de barra horizontal com as frequências dos valores para o atributo gênero no conjunto de dados do ABVdb.



Fonte – O autor.

Figura 12 – Histograma de frequência dos valores do IDH.

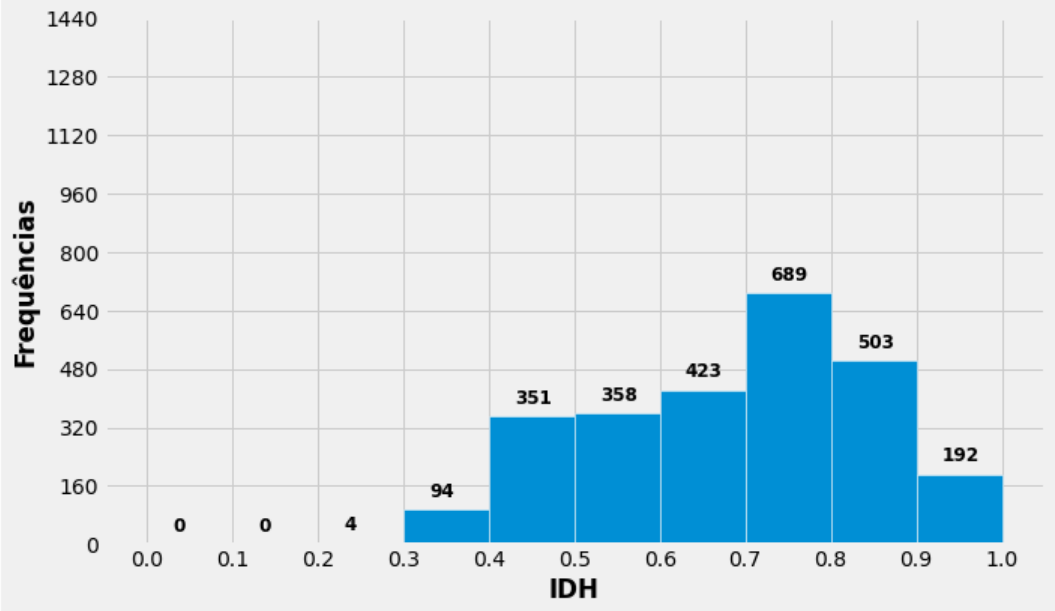


Fonte – O autor.

Os valores deste índice variam entre 0 e 1, onde quanto mais próximo de 0 indica uma avaliação ruim para um país, enquanto mais próximo de 1 indica os países com melhor resultado. No eixo X do gráfico estão os valores numéricos do índice, e no eixo Y estão as frequências de suas ocorrências no *dataframe*. As linhas removidas deste conjunto de dados foram todas as que possuiam valores que não fazem parte do escopo do projeto, entre 2004 e 2017. Além deles, os valores nulos também foram excluídos. A Figura 13 mostra como ficou a

distribuição das frequências do IDH após a remoção destas linhas.

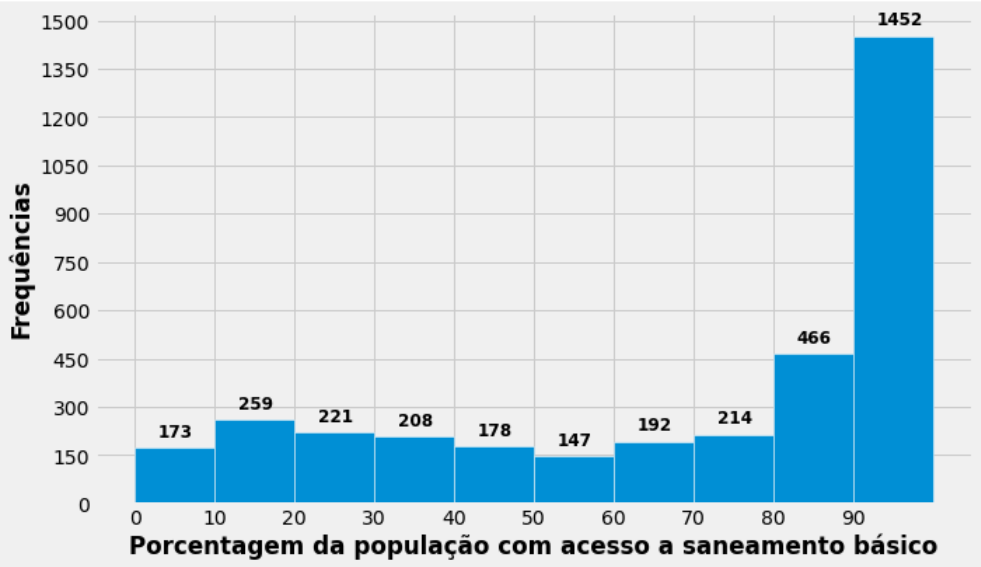
Figura 13 – Histograma de frequência dos valores do IDH após a remoção das linhas que não serão utilizadas.



Fonte – O autor.

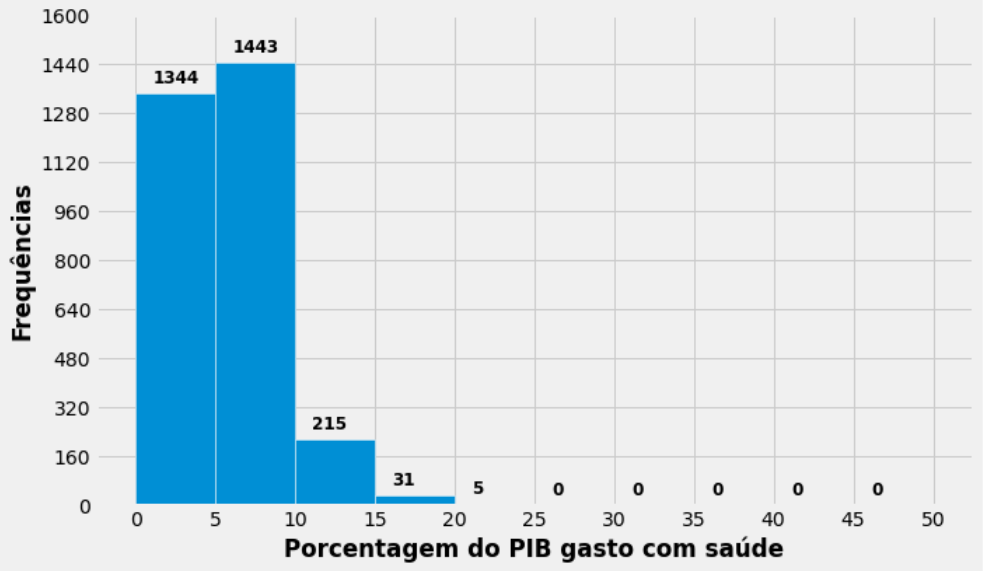
Já nas Figuras 14 e 15 são apresentados os histogramas de frequência dos valores nos conjuntos de dados do ASB e do GSPIB respectivamente. Cada instância é um par que referencia o registro da varíavel para o país e o ano. Da mesma forma como no IDH, estes *dataframes* também possuem valores nulos. É preciso também que todas as informações estejam dentro do escopo do projeto, entre os anos de 2004 e 2017.

Figura 14 – Histograma de frequência dos valores do ASB.



Fonte – O autor.

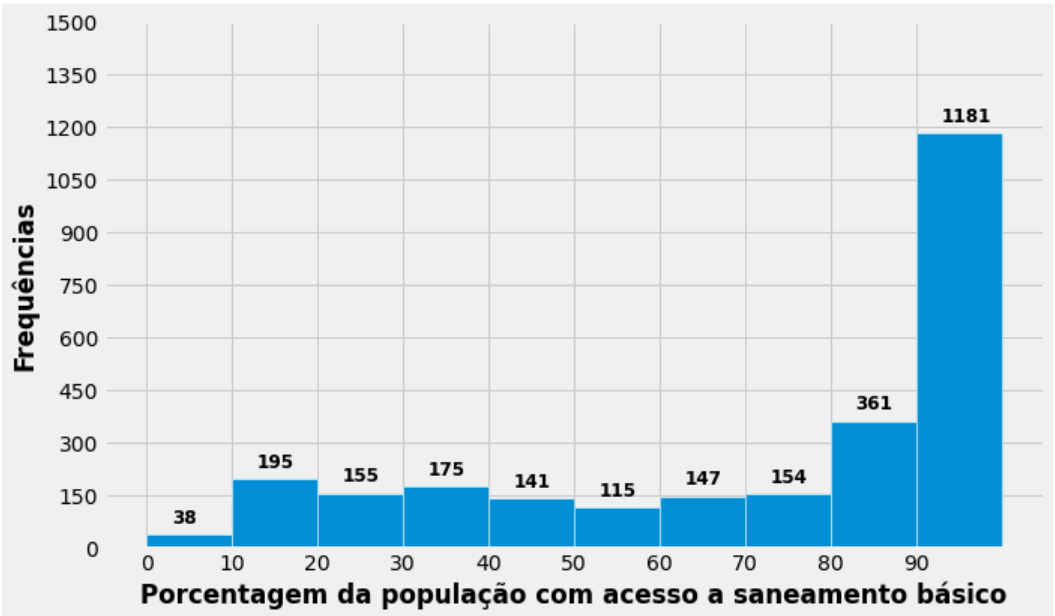
Figura 15 – Histograma de frequência dos valores do GSPIB.



Fonte – O autor.

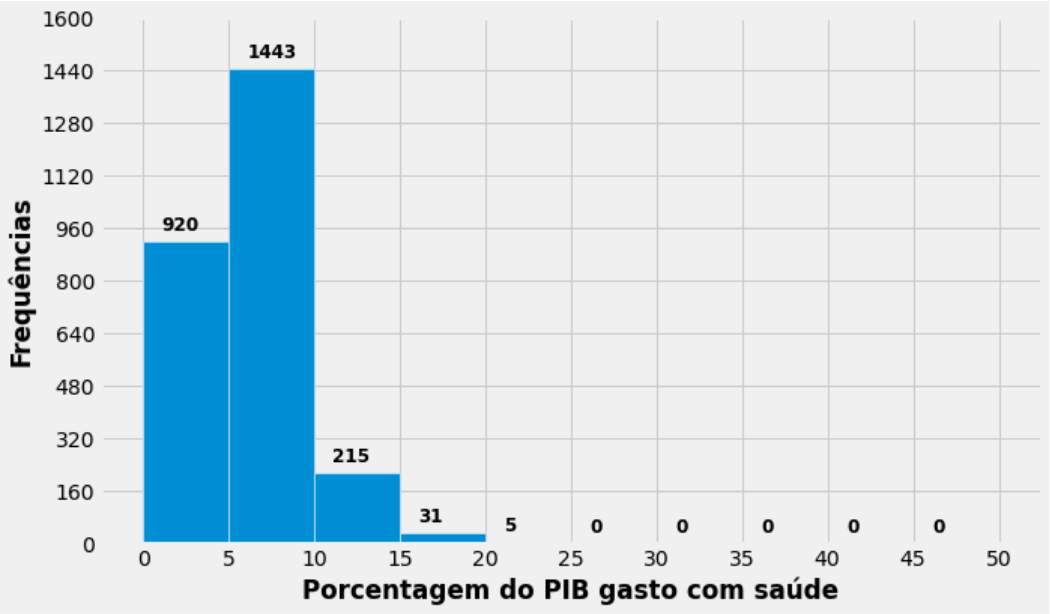
Diferentemente do IDH, os valores desta variável variam entre 0 e 100, representando a porcentagem em cada uma delas. Para o ASB, quanto mais próximo de 0, menos pessoas tem acesso ao serviço de saneamento básico, e quanto mais próximo de 100, uma maior parcela da população tem o benefício deste serviço. Já para o GSPIB, quanto mais próximo de 0, menos foi gasto com saúde utilizando o PIB, e quanto mais próximo de 100, mais foi gasto. Após a análise do gráfico, as linhas foram excluídqs dos dois conjuntos de dados. As Figuras 16 e 17 mostram os histogramas das frequências dos valores em ambos os *dataframes* após as alterações.

Figura 16 – Histograma de frequência dos valores do ASB após a remoção das linhas que não serão utilizadas.



Fonte – O autor.

Figura 17 – Histograma de frequência dos valores do GSPIB após a remoção das linhas que não serão utilizadas.



Fonte – O autor.

O *dataframe* criado com os dados do *GenBank* não precisou de uma limpeza dos seus dados, pois o seu processo de criação não apresentou nenhum valor inconsistente. Sendo assim, esta etapa foi concluída, e todos os conjuntos de dados que foram utilizados no projeto foram corrigidos com o mínimo possível de perda de informações úteis. A Tabela 3 mostra a relação entre a quantidade de linhas analisadas nesta etapa exploratória, as restantes após as correções e as que foram removidas em cada um dos conjunto de dados. Na próxima seção será visto como estes dados foram unificados para a criação das quatro amostras que serão utilizadas.

Tabela 3 – Relação entre a quantidade de linhas analisadas, as restantes e as que foram excluídas em cada *dataframe*.

Dataframe	Linhas analisadas	Linhas após correções	Linhas excluídas
ABVdb	3.232	3.097	135
IDH	5.481	2.614	2.867
ASB	3.510	2.662	848
GSPIB	3.038	2.614	424
Genbank	866	866	0

Fonte – O autor.

4.6 TRANSFORMAÇÃO DE DADOS

A última tarefa de preparação dos *dataframes* para a aplicação do algoritmo consiste na transformação dos dados. Para esta etapa, foram criadas quatro amostras, pois a intenção

foi buscar informações específicas sobre o vírus chikungunya em cada uma delas, de forma a preservar o máximo de informações possíveis. Cada amostra é o resultado da união entre os conjuntos de dados tratados, que possuem as informações específicas a serem analisadas.

As amostras foram preparadas com todos os ajustes possíveis, para que não ocorra nenhum problema na aplicação do algoritmo, e também, para que as informações geradas não sejam equivocadas, atrapalhando na interpretação. Nas próximas sub-seções será detalhado como foi o processo de adequação em cada uma delas.

4.6.1 Criação da Primeira e Segunda Amostra

A primeira amostra criada contém os dados dos *dataframes* do ABVdb, do IDH, do ASB e do GSPIB. Com a junção destes conjuntos de dados foi analisado quais possíveis padrões poderiam ser encontrados entre os genótipos dos casos registrados e as variáveis socioeconômicas escolhidas baseadas na região onde estes casos ocorreram. Para confeccioná-la, os 4 *dataframes* foram mesclados, baseado na interseção das suas colunas *Country* e *Date*. Após isso, as colunas *Acession*, *Gender*, *Host*, *Age* e *Date* foram excluídas, pois não serão utilizadas na análise desta amostra, assim também como as linhas que continham algum valor em branco devido à mesclagem.

Por fim, as colunas ASB e GSPIB estavam em formato de porcentagem, enquanto o IDH em formato decimal. Para padronizar as informações contidas nas três variáveis, todos os valores foram transformados em números decimais, com duas casas após a vírgula. Com isto, as colunas numéricas da primeira amostra ficaram com um intervalo de valores entre 0,01 e 1. A Figura 18 representa a versão final para a primeira amostra que será utilizada na análise.

As variáveis utilizadas na primeira amostra são a base de todos os conjuntos de dados a serem utilizados no algoritmo, pois elas estão presentes em todas as outras amostras. Diante deste fato, e de que as características socioeconômicas foram formatadas levando em consideração o escopo do trabalho entre o período de 2004 a 2017, todas as análises resultantes contemplarão os dados adquiridos após o início do surto do vírus chikungunya em 2004, e que foram registrados no ABVdb.

Já a segunda amostra seguiu o mesmo processo de criação da primeira, porém, levou em consideração o gênero e a idade dos indivíduos infectados. A proposta foi buscar alguma

Figura 18 – Primeiros e últimos registros do *Dataframe* que representa a primeira amostra a ser utilizada na análise.

	Genotype	Country	IDH	ASB	GSPIB
0	East-Central-South-African	Angola	0.57	0.49	0.03
1	Asian and Caribbean	Antigua and Barbuda	0.77	0.88	0.06
2	East-Central-South-African	Australia	0.90	1.00	0.08
3	East-Central-South-African	Australia	0.92	1.00	0.08
4	East-Central-South-African	Australia	0.94	1.00	0.09
...	...	...	...	...	...
2545	East-Central-South-African	Viet Nam	0.67	0.75	0.06
2546	East-Central-South-African	Viet Nam	0.67	0.75	0.06
2547	East-Central-South-African	Viet Nam	0.67	0.75	0.06
2548	East-Central-South-African	Viet Nam	0.67	0.75	0.06
2549	East-Central-South-African	Viet Nam	0.67	0.75	0.06

2550 rows × 5 columns

Fonte – O autor.

relação entre os casos confirmados do vírus chikungunya, seus genótipos, estas duas variáveis e as características socioeconômicas dos países onde os casos ocorreram. As colunas *Acession*, *Date* e *Host* foram excluídas, pois não farão parte da análise. As linhas que continham nas colunas *Gender* e *Age* valores nulos também foram excluídas, assim como as linhas com valores em branco devido à mesclagem. A Figura 19 representa a versão final para o conjunto de dados desta amostra.

Figura 19 – Visualização dos primeiros e últimos registros do *Dataframe* que representa a segunda amostra.

	Genotype	Country	Gender	Age	IDH	ASB	GSPIB
0	East-Central-South-African	Australia	M	59	0.90	1.00	0.08
1	East-Central-South-African	Brazil	F	24	0.76	0.87	0.09
2	East-Central-South-African	Brazil	F	18	0.76	0.87	0.09
3	East-Central-South-African	China	F	33	0.70	0.74	0.04
4	East-Central-South-African	China	M	7	0.70	0.74	0.04
...	...	...	...	...	...	...	...
232	Asian and Caribbean	USA	M	21	0.92	1.00	0.16
233	Asian and Caribbean	USA	M	42	0.92	1.00	0.16
234	Asian and Caribbean	USA	F	56	0.92	1.00	0.16
235	Asian and Caribbean	USA	F	45	0.92	1.00	0.17
236	Asian and Caribbean	USA	F	61	0.92	1.00	0.17

237 rows × 7 columns

Fonte – O autor.

4.6.2 Criação da Terceira e Quarta Amostra

A terceira e a quarta amostra possuem a mesma proposta de análise: buscar alguma relação entre a frequência dos casos confirmados de infecção em seres humanos pelo vírus chikungunya, baseada nos genótipos, variedade da sintomatologia identificada nos artigos do GenBank e as variáveis socioeconômicas selecionadas do UNDP. A única diferença entre ambas é que na terceira estarão somente os dados que tiveram algum registro de sintoma, enquanto na quarta levará em consideração os que não tem registro como assintomáticos. Para a confecção destas duas amostras os *dataframes* do ABVdb e do *GenBank* serão utilizados.

Na criação da terceira amostra, todos os *dataframes* selecionados foram unificados. O processo inicial foi o mesmo da primeira amostra para unir os conjuntos de dados socioeconômicos e do ABVdb, e posteriormente, os sintomas se juntaram baseado na interseção dos valores da coluna *Acession* que contém os registros únicos dos casos. Com isso, foi possível estabelecer uma relação entre os casos registrados no banco de dados do ABVdb e os sintomas deles encontrados na literatura do *GenBank*. Como em alguns casos existem registros de mais de um sintoma, após a união dos *dataframes*, a coluna *Acession* passou a ter valores replicados, para poder registrar todas as ocorrências. Por fim, os valores que não estão em ambos os conjuntos de dados foram adicionados com o sintoma "nulo", sendo estas linhas excluídas, bem como as *Acession*, *Date*, *Host*, *Gender* e *Age*, que não serão utilizadas para a análise. A Figura 20 representa a versão final do *dataframe* para a terceira amostra.

Já para a quarta amostra, o processo foi o mesmo. Porém, em vez de excluir os valores nulos para os casos com os sintomas não identificados, eles foram substituídos por *asymptomatic*, indicando que, como não existe na literatura nenhum registro sobre quais manifestações o paciente apresentou, ele foi considerado um assintomático. As mesma colunas que foram excluídas na segunda amostra também foram removidas aqui. A Figura 21 mostra como a quarta amostra está estruturada.

Após todas as alterações, as quatro amostras ficaram prontas para serem utilizadas pelo algoritmo. A Tabela 4 contém uma breve descrição das variáveis utilizadas e a quantidade de linhas e colunas geradas em cada uma delas. A seção a seguir abordará a etapa de *Data Mining* do projeto, e nela será apresentado como estes conjuntos de dados foram implementados pelo *K-prototype*, de forma a se gerar os *clusters* para poder avaliá-los e interpretá-los.

Figura 20 – Primeiros e últimos registros do *Dataframe* que representa a terceira amostra a ser utilizada na análise.

	Genotype	Country	IDH	ASB	GSPIB	Symptom
0	East-Central-South-African	China	0.68	0.70	0.04	fever
1	East-Central-South-African	China	0.68	0.70	0.04	arthralgia
2	East-Central-South-African	China	0.68	0.70	0.04	fever
3	East-Central-South-African	China	0.68	0.70	0.04	conjunctivitis
4	East-Central-South-African	China	0.68	0.70	0.04	bleeding
...	...	...	...	...	...	...
432	East-Central-South-African	Sri Lanka	0.73	0.88	0.04	high fever
433	East-Central-South-African	Sri Lanka	0.73	0.88	0.04	myalgia
434	East-Central-South-African	Sri Lanka	0.73	0.88	0.04	arthralgia
435	East-Central-South-African	Sri Lanka	0.73	0.88	0.04	rash
436	East-Central-South-African	Sri Lanka	0.73	0.88	0.04	epistaxis

437 rows × 6 columns

Fonte – O autor.

Figura 21 – Visualização dos primeiros e últimos registros do *Dataframe* que representa a quarta amostra.

	Genotype	Country	Symptom	IDH	ASB	GSPIB
0	East-Central-South-African	Sri Lanka	high fever	0.73	0.88	0.04
1	East-Central-South-African	Sri Lanka	myalgia	0.73	0.88	0.04
2	East-Central-South-African	Sri Lanka	arthralgia	0.73	0.88	0.04
3	East-Central-South-African	Sri Lanka	rash	0.73	0.88	0.04
4	East-Central-South-African	Sri Lanka	epistaxis	0.73	0.88	0.04
...	...	...	...	...	...	...
2901	East-Central-South-African	Brazil	asymptomatic	0.76	0.87	0.09
2902	East-Central-South-African	Brazil	asymptomatic	0.76	0.87	0.09
2903	East-Central-South-African	Brazil	asymptomatic	0.76	0.87	0.09
2904	East-Central-South-African	Brazil	asymptomatic	0.76	0.88	0.09
2905	East-Central-South-African	Brazil	asymptomatic	0.76	0.88	0.09

2906 rows × 6 columns

Fonte – O autor.

Tabela 4 – Estrutura das quatro amostras a serem utilizadas na análise.

Dataframe	**Variáveis Utilizadas**	**Linhas**	**Colunas**
Primeira Amostra	Genótipo, País, IDH, ASB, GSPIB	2.550	5
Segunda Amostra	Genótipo, País, IDH, ASB, GSPIB, Idade, Gênero	237	7
Terceira Amostra	Genótipo, País, IDH, ASB, GSPIB, Sintoma	437	6
Quarta Amostra	Genótipo, País, IDH, ASB, GSPIB, Sintoma	2.906	6

Fonte – O autor.

4.7 MINERAÇÃO DE DADOS

Observou-se na literatura que o *K-prototype* de Huang (1997) é um algoritmo de *Data Mining* que utiliza a técnica de *clustering* para resolver o problema de agrupar dados numéricos e categóricos. Como as quatro amostras criadas possuem esta característica, ele foi utilizado para buscar padrões em cada uma delas. A análise em cada um dos conjuntos de dados que representam estas amostras foi feita separadamente, e seguiu um roteiro para os experimentos. Devido à escassez do detalhamento da implementação do *K-prototype*, toda a programação foi auxiliada pela biblioteca *kmodes*, que possui os recursos para a execução na linguagem de programação Python.

Inicialmente, foi realizada uma rodada de testes iniciais para averiguar se a implementação do algoritmo na linguagem ocorreria sem problemas. O *K-prototype* foi executado 10 vezes em cada amostra, com uma iteração cada, para identificar possíveis erros. Uma iteração equivale à execução da sequência de passos do algoritmo, desde a alocação dos protótipos iniciais até a sua convergência, como descrito na Subseção 3.5.2. Como o intuito deste teste foi somente analisar se algum problema ocorreria na execução, adotou-se um valor aleatório para a quantidade de *clusters* a serem gerados. Como resultado, o algoritmo conseguiu agrupar sem problemas cada uma das amostras. O Apêndice A detalha como foi a implementação dele na linguagem Python.

Após a aprovação da rodada de testes, o algoritmo ficou pronto para ser utilizado em cada uma das amostras, de forma a gerar os agrupamentos para serem analisados. O primeiro passo para realizar esta tarefa foi identificar o melhor valor de k, ou seja, da quantidade de *clusters* a ser gerado em cada uma das amostras. Para isso, adotou-se a utilização de uma técnica específica, que será melhor detalhada a seguir.

4.7.1 Identificação da quantidade de *clusters*

O *K-prototype* exige que o usuário informe previamente a quantidade de *clusters* a serem gerados pelo algoritmo. Este valor é essencial para se obter um bom resultado na técnica de *clustering*. Para este procedimento ser realizado de uma maneira cientificamente segura, optou-se por buscar na literatura técnicas que encontrassem este valor. O primeiro encontrado foi o método da silhueta, porém, ao tentar implementá-lo, o Python não apresentou resultados,

alegando um problema de que não era possível calcular o coeficiente de silhueta devido as variáveis categóricas. Diante disso, ele foi descartado.

Um outro método identificado na literatura foi o do cotovelo. Ele gera um gráfico, onde, quando o valor de *k* (*cluster*) atinge um número ideal, é possível identificar uma curva semelhante a um cotovelo. Para o contexto do projeto, esta curva é gerada devido à queda mais drástica dos valores resultantes da função de custo do *K-prototype* (Subseção 3.5.1). Como a sua implementação na linguagem foi sem nenhum problema, ele foi selecionado. As bibliotecas *kmodes* e *pyploty* forneceram os recursos necessários para a sua utilização.

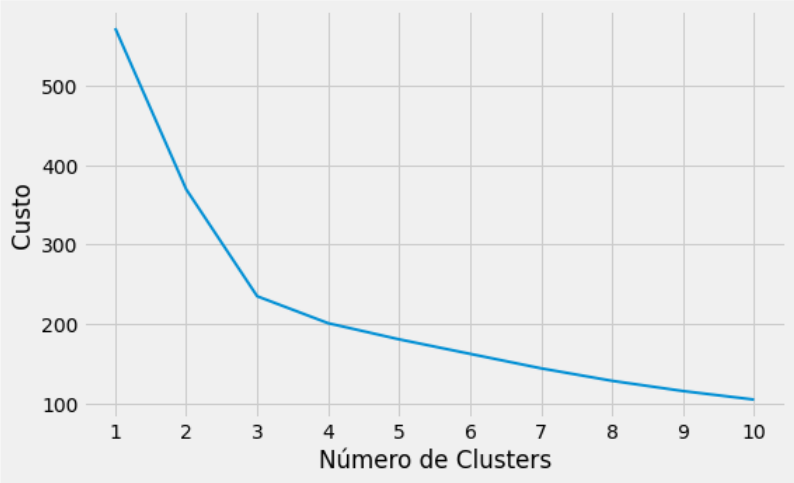
O código fonte da implementação do método é detalhado no Apêndice B. Seu funcionamento gira em torno de rodar o algoritmo *K-prototype* com valores diferentes para a quantidade de *clusters*, salvando sempre o menor custo gerado diante das iterações realizadas. Ao todo, foram realizadas 10 rodadas de execuções, onde o número da execução representou a quantidade de *clusters* a ser utilizada. Para cada rodada, houveram 1000 iterações, onde foi salvo o menor custo encontrado. Os 10 valores de custos salvos serviram para indicar os pontos em um gráfico de linhas, e identificar qual o melhor *k* para ser utilizado.

A Figura 22 representa o gráfico gerado para a primeira amostra, após a aplicação do método do cotovelo. Pode-se observar claramente que a curva é acentuada quando o número de *clusters* chega a 3. Sendo assim, este foi o valor escolhido para ser utilizado pelo *K-prototype* ao ser aplicado a esta amostra. Da mesma forma como na segunda amostra, onde a curva que representa o cotovelo pode ser identificada com o mesmo valor. A Figura 23 mostra o gráfico do método gerado para ela.

Já para a terceira amostra, o gráfico do método do cotovelo pode ser visto na Figura 24. Após a aplicação da técnica, foi possível identificar o valor 2 para ser utilizado como número de *clusters* para esta amostra. Diferentemente da quarta amostra, onde o valor identificado foi 3, como mostrado na Figura 25,

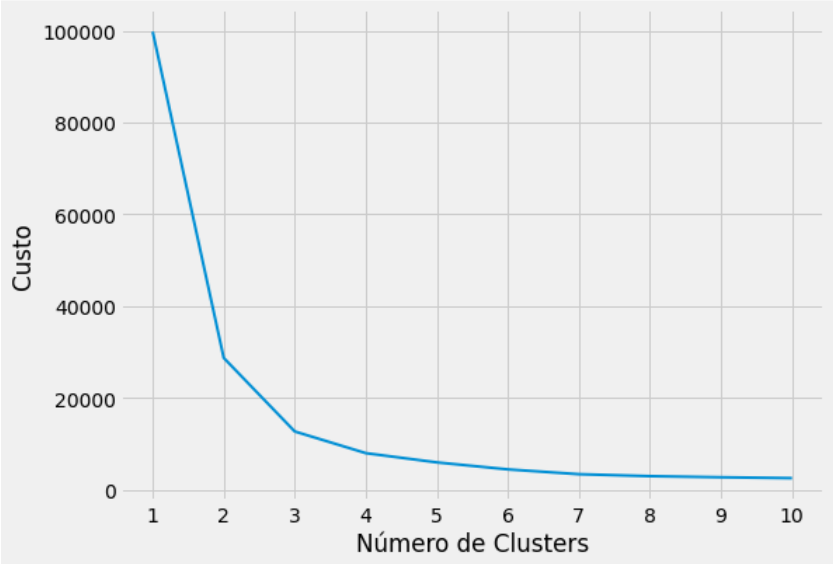
Com a correta identificação da quantidade *clusters* a serem utilizados em cada uma das amostras, foi possível começar os experimentos com o algoritmo. A próxima subseção abordará como todo o processo foi realizado.

Figura 22 – Resultado da aplicação do método do cotovelo no conjunto de dados da primeira amostra.



Fonte – O autor.

Figura 23 – Resultado da aplicação do método do cotovelo no conjunto de dados da segunda amostra.

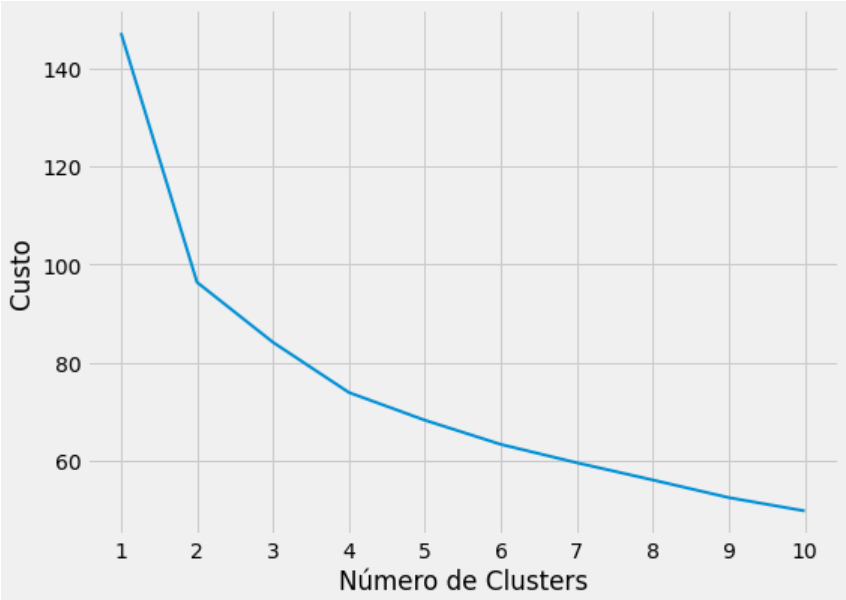


Fonte – O autor.

4.7.2 Execução dos Experimentos

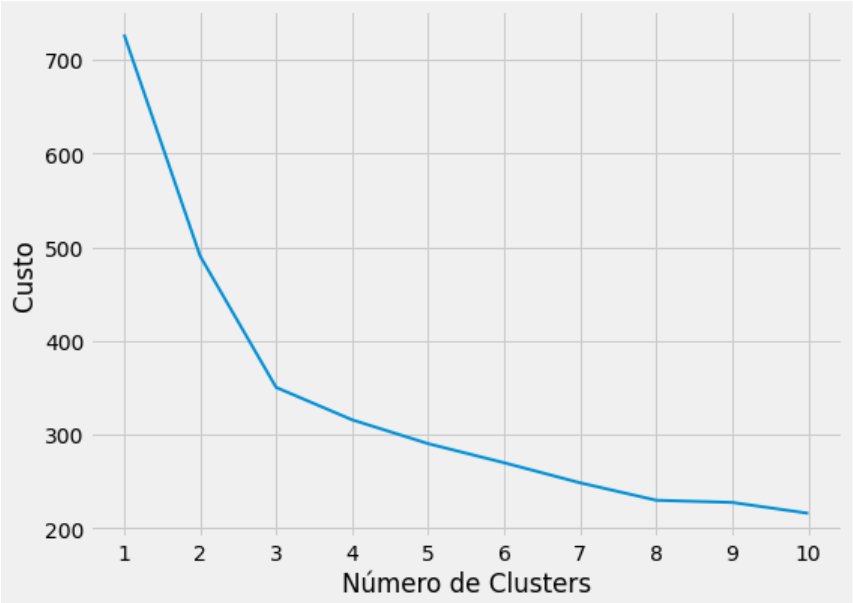
Os experimentos foram realizados de forma a implementar sistematicamente as execuções para o levantamento dos resultados que serão avaliados pelas métricas, e interpretados para a busca por conhecimento. Cada uma rodada de experimentos foi aplicada em cada uma das amostras presentes no projeto. Referente à quantidade de execuções, devido à necessidade de se obter uma segurança estatística sobre os resultados que posteriormente serão avaliados, e à disponibilidade de recursos computacionais disponíveis, optou-se por três rodadas de experimentos, onde os algoritmos foram executados 100, 500 e 1000 vezes. A Figura 26 representa o processo desenvolvido em cada rodada dos experimentos.

Figura 24 – Resultado da aplicação do método do cotovelo no conjunto de dados da terceira amostra.



Fonte – O autor.

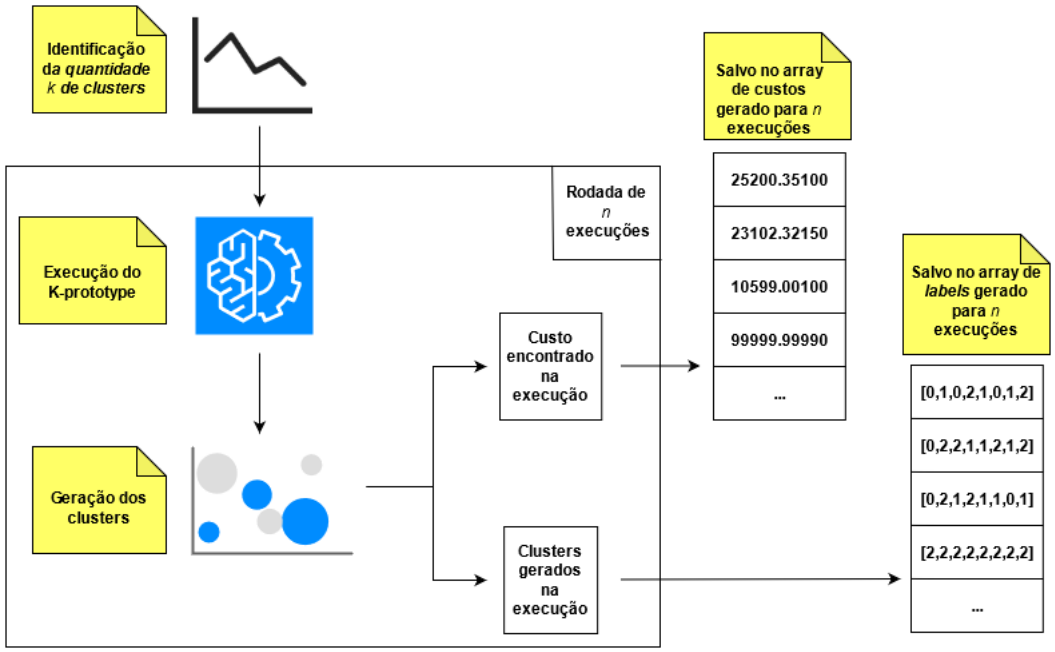
Figura 25 – Resultado da aplicação do método do cotovelo no conjunto de dados da quarta amostra.



Fonte – O autor.

A primeira etapa deste processo foi utilizar o método do cotovelo para encontrar a quantidade de *clusters* a ser utilizada para a amostra em questão, o que foi visto na Sub-seção 4.7.1. Com o valor identificado, o próximo passo foi utilizar o algoritmo *K-prototype* para agrupar os dados. Isso foi repetido uma quantidade *n* de vezes, que pode ser 100, 500 ou 1000, de acordo com a rodada de experimentos. Cada execução realizada nos experimentos foi feita com uma iteração, onde o custo encontrado foi armazenado, assim também como os *labels*, que são os rótulos que identificam a qual *cluster* pertence cada um dos objetos do conjunto de dados.

Figura 26 – Processo de execução em uma rodada de experimentos. Os valores apresentados são ilustrativos.



Fonte – O autor.

Os experimentos foram realizados sem nenhum problema identificado, e todos os valores encontrados foram armazenados. Após isso, a próxima etapa do processo do KDD consistiu em avaliar estes resultados, buscando mensurar o quão similares foram os agrupamentos gerados pelo *K-prototype*. O detalhamento de como esta avaliação foi feita será descrito na próxima seção.

4.8 AVALIAÇÃO DOS ALGORITMOS

Na busca por estudos que avaliassem de forma correta o *K-prototype*, foram encontradas na literatura algumas métricas que qualificavam os agrupamentos. Prabha e Visalakshi (2015) utilizaram medidas de validade externa para mensurar a similaridade entre conjuntos de *clusters*. Diante do contexto do projeto, duas destas métricas foram selecionadas inicialmente: o Índice de Rand e o Índice de Jaccard. Os valores resultantes variam entre 0 e 1, onde, quanto mais próximo de 1, maior a similaridade entre os *clusters* gerados pela técnica.

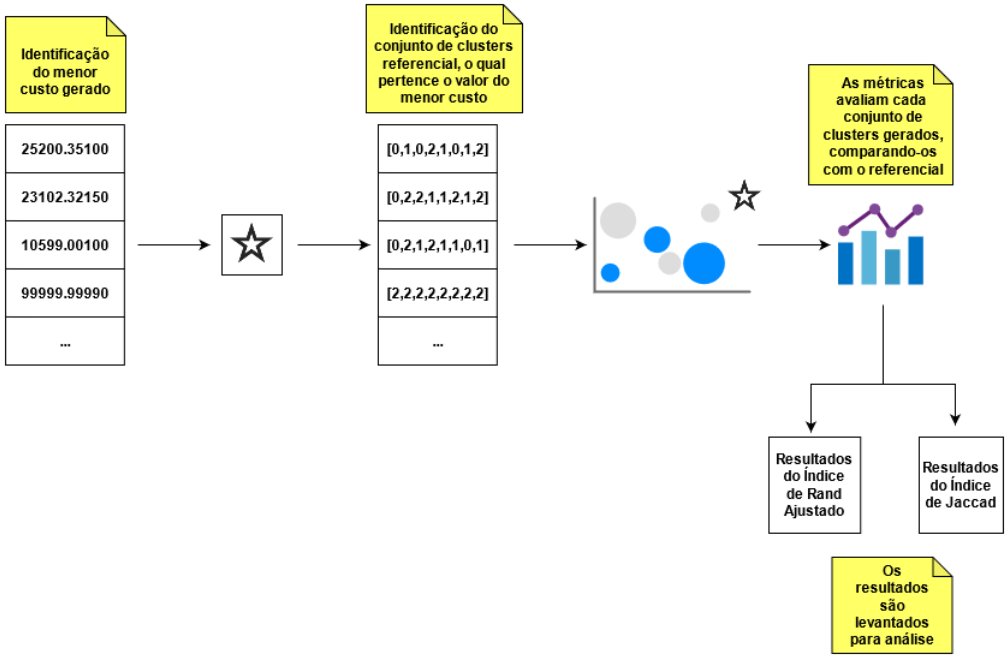
Na aplicação do Índice de Rand, se dois conjuntos de *clusters* aleatórios forem comparados, seu valor não é zero, o que não condiz com o resultado esperado da métrica. Diante disto, optou-se por substituir esta medida pelo Índice de Rand Ajustado, que leva em consideração a aleatoriedade da geração dos *clusters* para avaliá-lo, Hubert e Arabie (1985). Para

esta métrica, os valores variam entre -1 e 1, onde o valor 1 é uma combinação perfeita obtida, e o valor -1 equivale a conjunto de dados completamente aleatórios.

No fim, o Índice de Rand Ajustado e o Índice de Jaccard foram selecionados para a análise da similaridade dos agrupamentos gerados pelo algoritmo. O Apêndice C mostra o código fonte para a implementação de ambos na linguagem Python, que foi auxiliada pela biblioteca *Scikit-learn*. A aplicação destas métricas também seguiu um roteiro, como pode ser visto na Figura 27.

O processo foi realizado separadamente para cada amostra e em cada rodada de experimentos. Para utilizar as métricas, é preciso ter um conjunto de *clusters* referencial para servir como base de comparação com todos os outros gerados pelo algoritmo. Diante disso, foi necessário identificar qual dentre todos os agrupamentos gerados em cada rodada de experimentos serviria de referência para comparação. Seguindo a lógica de que, o *K-prototype* busca minimizar o valor resultante da sua função de custo, optou-se por selecionar o referencial baseado no conjunto de *cluster* que encontrou o menor custo na aplicação do algoritmo em cada experimento.

Figura 27 – Processo de aplicação das métricas para avaliar o algoritmo. Os valores apresentados são ilustrativos.



Fonte – O autor.

Como nas execuções, a cada iteração foi salvo o custo e os *labels* dos *clusters*, a

posição deles nas suas respectivas listas é a mesma. Sendo assim, para encontrar o conjunto de *clusters* que será utilizado como referencial, bastou realizar uma busca na lista de custos, e armazenar o índice da posição onde o menor valor foi encontrado. Com este índice aplicado na lista dos *labels*, o referencial foi identificado. O próximo passo então foi avaliar os resultados do agrupamento. Esta avaliação foi realizada aplicando as métricas na comparação entre o conjunto de *clusters* referencial e todos os outros gerados na rodada de experimentos em questão. Por fim, os valores resultantes da avaliação foram armazenados em duas listas: uma referente ao Índice de Rand Ajustado, e a outra ao Índice de Jaccard.

Após a aplicação das métricas em todos os conjuntos de *clusters* gerados em cada rodada de experimentos, foi possível avaliar a similaridade dos agrupamentos performados pelo *K-prototype*, e verificar se o algoritmo é instável a partir do aumento da quantidade de execuções. O próximo capítulo discutirá todos os resultados obtidos, as percepções a partir dos valores entregues pelas medidas, e os padrões identificados após a aplicação do *clustering*.

5 RESULTADOS

O levantamentos dos resultados levantados no projeto foi dividida em duas partes. A primeira, é referente a avaliação dos agrupamentos realizados pelo algoritmo através das métricas. Já a segunda, é referente a interpretação para a busca pelo conhecimento, ao identificar se existem relações ao analisar a melhor solução encontrada pelo algoritmo. Os resultados das métricas refletiram na interpretação final das amostras. Este capítulo abordará a condução do processo de análise dos resultados da técnica de *clustering*.

5.1 AVALIAÇÃO DOS RESULTADOS

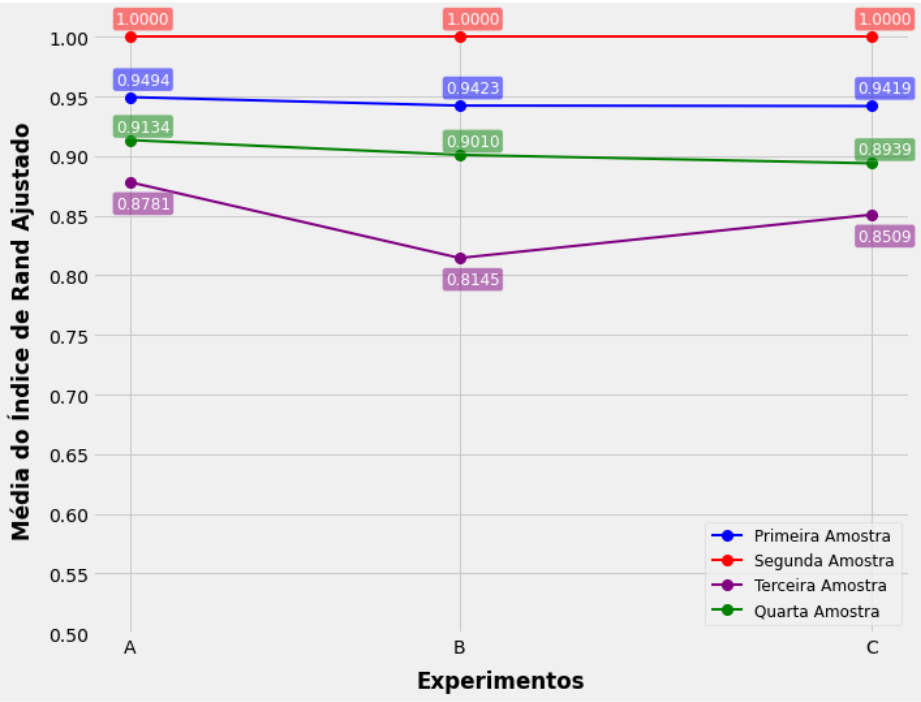
Como visto na Seção 4.8, o Índice de Rand Ajustado e o Índice de Jaccard foram as métricas utilizadas para avaliar a similaridade dos *clusters* gerados na etapa de *Data Mining*. A avaliação foi feita sobre as três rodadas dos experimentos, onde cada uma teve um aumento significativo na quantidade de agrupamentos performados pelos algoritmos, como visto na seção 4.7.2. Com a progressividade do aumento das execuções e os resultados das métricas em cada experimento, foi possível observar o desempenho do *K-prototype* ao agrupar os dados de cada uma das amostras.

As informações a respeito da análise das métricas foram dispostas em gráficos. Referente ao seu conteúdo, foram representados os valores da média aritmética dos resultados obtidos nos agrupamentos com o Índice de Rand Ajustado e o Índice de Jaccard para cada uma das amostras. A avaliação da métrica foi realizada em cada agrupamento individualmente, ou seja, foram gerados 100 valores para o experimento A, 500 valores para o experimento B e 1000 valores para o experimento C. Optou-se por representá-los com a média aritmética de todo o conjunto de agrupamentos, pois é uma medida estatística que dá uma noção geral de como a métrica avaliou a amostra em questão. Nas sub-seções a seguir, todos os resultados referente a avaliação dos experimentos foram levantados e discutidos, com as percepções encontradas a respeito das métricas.

5.1.1 Avaliação dos resultados para o Índice de Rand Ajustado

Os resultados levantados com o Índice de Rand Ajustado podem ser visualizados na Figura 28. Neste gráfico, é possível observar as variações das médias aritméticas resultantes desta métrica com o aumento das execuções (representadas por cada experimento). A segunda amostra foi a que apresentou os melhores resultados dentre as quatro, com o valor máximo em todos os experimentos, ou seja, todos os conjuntos de *clusters* formados pelo algoritmos foram similares. Por outro lado, a terceira amostra, que foi a única que agrupou em torno de 2 *clusters*, teve os piores resultados dentre todas as outras, além de não mostrar um padrão no comportamento dos resultados em cada experimento. Já a primeira e a quarta amostra tiveram decréscimos a cada experimento realizado, mostrando que, com mais execuções, menos similares foram os agrupamentos destas amostras.

Figura 28 – Resultados do Índice de Rand Ajustado para cada uma das amostras em cada um dos experimentos.

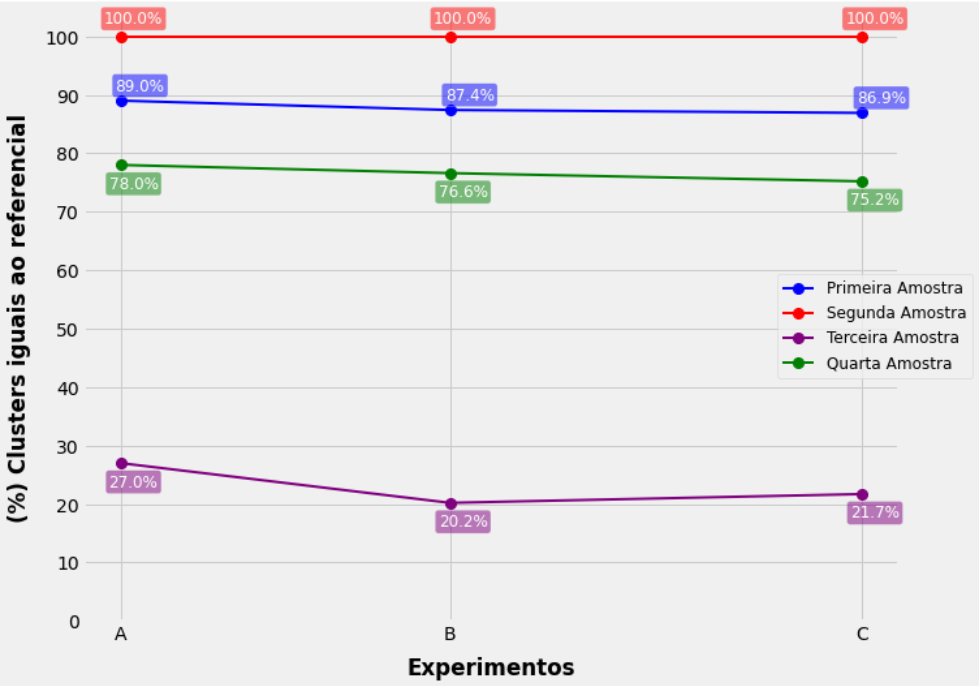


Fonte – O autor.

O maior decréscimo nas notas pode ser observado na terceira amostra. No experimento A, o valor resultante foi de 0.8781, enquanto no experimento B, foi de 0.8145, variando negativamente em -0.0636. Já o maior acréscimo também foi identificado na terceira amostra, onde no experimento C o valor da média foi 0.8509, e se comparado com o experimento B, houve uma variação positiva de 0.0364.

É interessante observar também o quão influente foi a frequência com que aparece o valor máximo de similaridade nestas médias aritméticas dos resultados do Índice de Rand Ajustado. Ele é representado pela quantidade de conjunto de *clusters* que são exatamente iguais ao conjunto de *clusters* referencial, e obtiveram o valor 1. A Figura 29 mostra a porcentagem destes valores para cada uma das amostras em cada um dos experimentos.

Figura 29 – Porcentagem dos *clusters* gerados semelhantes ao *cluster* referencial.



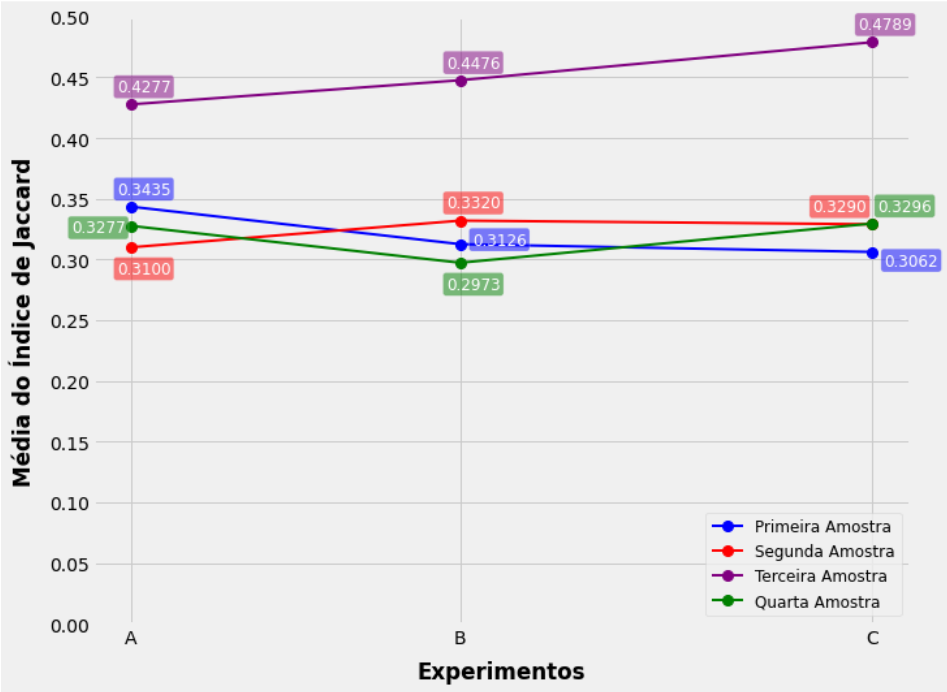
Fonte – O autor.

O comportamento das linhas para este gráfico é semelhante a do gráfico da Figura 28. Todos os *clusters* da segunda amostra são iguais ao referencial, e, na primeira e na quarta amostra, houveram decréscimos com o aumento do número de execuções, além de apresentarem uma grande maioria de agrupamentos semelhantes. Porém, um fato interessante pode ser analisado na terceira amostra. A porcentagem de *clusters* iguais ao referencial teve um valor abaixo de 30% em todos os três experimentos, porém, na média dos valores do Índice de Rand Ajustado para esta amostra, a nota ficou acima de 0.80. Isso ocorreu devido ao fato de que, a maioria dos agrupamentos desta amostra tiveram uma nota de 0.909890, com 67% de ocorrências deste valor no experimento A, 67,6% no experimento B e 69,9% no experimento C. Essa nota significa que, os *clusters* gerados não são completamente iguais ao referencial, mas são bastante similares.

5.1.2 Avaliação dos resultados para o Índice de Jaccard

Já para o Índice de Jaccard, os resultados foram bem diferentes se comparados ao Índice de Rand Ajustado. Como pode ser visto na Figura 30, as médias aritméticas resultantes da avaliação desta métrica não ultrapassaram a nota 0.5 em nenhum dos 3 experimentos. Todas as amostras que foram agrupadas com um valor de *k* igual a 3 ficaram com a média das notas entre 0.29 e 0.35, e dentre elas, somente a primeira amostra apresentou um padrão a cada experimento, com um decréscimo dos seus resultados. Já a terceira amostra, que foi agrupada em torno de 2 *clusters*, apresentou um crescimento na média dos resultados desta métrica com o aumento do número de execuções, e obteve o melhor resultado dentre as quatro.

Figura 30 – Resultados do Índice de Jaccard para cada uma das amostras em cada um dos experimentos.



Fonte – O autor.

A maior variação positiva foi observada na quarta amostra, onde no experimento B apresentou um valor da média de 0.2973, enquanto no experimento C um valor de 0.3296, variando 0.0323. Já a maior variação negativa se deu na primeira amostra, onde no experimento A apresentou um valor de 0.3435 e no B, 0.3126, variando em -0.0309.

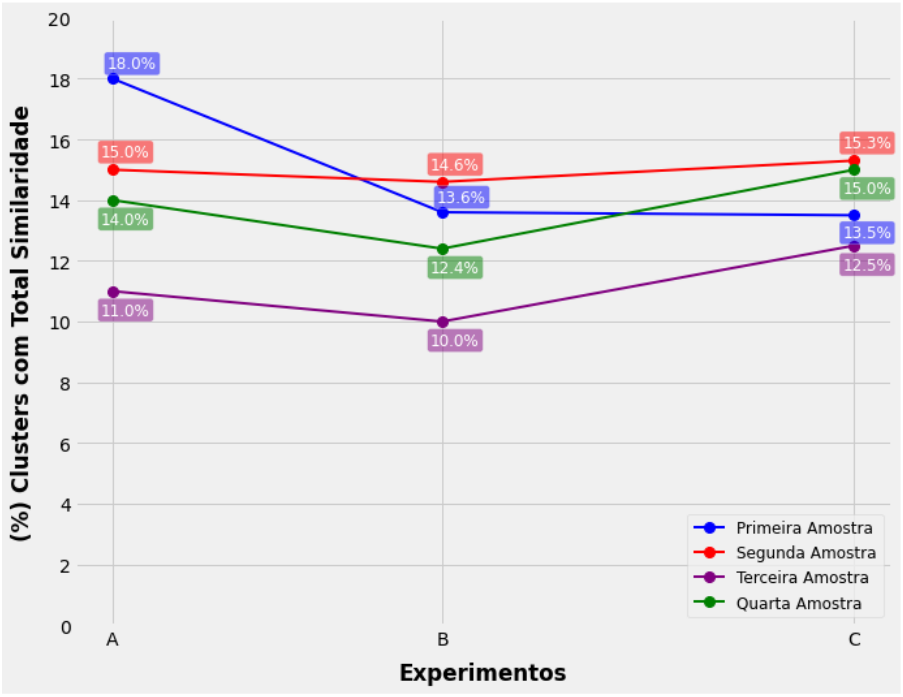
Estes valores resultantes do Índice de Jaccard não dão uma exata precisão do quão similares são os agrupamentos do algoritmo *K-prototype*. Diferente do Índice de Rand Ajustado, esta métrica leva em consideração a qual *label* os dados agrupados pertencem. Exemplificando,

se um conjunto de *clusters* avaliado for comparado com o conjunto de *clusters* referencial, ambos tiverem os dados agrupados nos mesmos *clusters*, porém, cada *cluster* tem um *label* diferente, o valor desta métrica não vai ser 1. Como o algoritmo *K-prototype* inicializa a sua execução com diferentes centróides, essa aleatoriedade afeta diretamente os resultados desta métrica.

5.1.3 Avaliação Final das métricas

Buscou-se analisar também a porcentagem dos conjuntos de *clusters* gerados que, em ambas as métricas, obtiveram o valor máximo de similaridade. A Figura 31 mostra um gráfico que representa essa distribuição em cada um dos experimentos. Como pode ser visto, a primeira amostra continuou com o mesmo comportamento observado em cada métrica individualmente, onde, com o aumento do número de execuções, foi observado um decréscimo dos seus resultados. As outras amostras não apresentaram um padrão específico de crescimento ou diminuição da porcentagem encontrada em cada experimento. Já a terceira amostra, apresentou os piores resultados dentre as quatro em todos os experimentos.

Figura 31 – Porcentagem dos *clusters* que em ambas as métricas obtiveram o valor 1.



Fonte – O autor.

Dentre todas as amostras, a segunda foi a que menos teve uma variação relevante para este resultado, variando somente -0.4% na comparação das execuções dos experimentos A e B, e 0.7% nos experimentos B e C. A maior variação negativa foi a da primeira amostra, onde

no experimento A apresentou 18% e no B, 13,6%, variando -4,4%. Já a maior variação positiva dos *clusters* com total similaridade foi da quarta amostra, onde no experimento B apresentou 12,4% e no experimento A 15%, representando um crescimento de 2,6%.

Com o levantamento dos resultados de ambas as métricas separadamente, foi possível observar que o algoritmo *K-prototype*, de uma forma geral, performou bem os agrupamentos, no sentido da similaridade entre os resultados da técnica de *clustering*. Não foi encontrado na literatura nenhum parâmetro para indicar se os resultados foram aprovados ou reprovados, mas, partindo do princípio que os valores do Índice de Rand Ajustado variam entre -1 e 1, e que o menor valor encontrado para a média foi de 0.8145, como visto na Figura 28, pode-se afirmar que a amostra com o pior resultado dentre as quatro, ainda apresentou um valor muito bom. Os valores desta métrica foram as mais seguras para verificar o quão similares foram os agrupamentos.

Já referente ao Índice de Jaccard, a sensibilidade com o qual a sua avaliação é feita referente aos *labels* dos *clusters* interferiu diretamente nos resultados finais. Como o algoritmo *K-prototype*, por padrão, inicializa os seus protótipos iniciais aleatoriamente, sem a definição de qual *labels* eles pertencem, o resultado final da métrica é afetado.

Após a avaliação e a análise dos resultados das métricas, os resultados puderam ser interpretados. A próxima seção irá detalhar quais *insights* puderam ser encontrados no levantamento de cada uma das amostras, e como os protótipos gerados foram interpretados.

5.2 INTERPRETAÇÃO DOS RESULTADOS

Para descrever as características de cada *cluster*, foi preciso inicialmente identificar qual conjunto de *clusters* seria escolhido para análise. Tendo em vista que o conjunto referencial foi encontrado baseado no menor custo gerado pelo algoritmo em cada um dos experimentos, e nos resultados das métricas em cada experimento, o conjunto de *clusters* escolhido para ser interpretado foi o referencial encontrado no maior número de execuções, ou seja, no experimento C.

De forma a auxiliar na interpretação dos resultados, foi utilizada uma biblioteca do Python chamada SHAP (SHapley Additive exPlanations), cujo objetivo é revelar informações específicas sobre os modelos de *machine learning*. Ao utilizar na técnica de *clustering*, foi

possível identificar através de um gráfico, quais variáveis foram mais relevantes durante o processo no qual o algoritmo agrupou os dados, de acordo com uma média de impacto gerada pela biblioteca. Sua documentação está disponível em https://github.com/slundberg/shap. Isso foi de grande valia para o projeto, pois aspectos importantes sobre os resultados finais puderam ser identificados.

Cada conjunto de dados foi analisado individualmente, a partir da identificação das variáveis mais relevantes para os resultados obtidos. Buscou-se também, a partir dos dados dos protótipos levantados, verificar se existem relações entre os genótipos e as outras variáveis, de forma a identificar padrões ou tendências entre as características mais singulares das amostras, como os dados socioeconômicos na primeira, a idade e o gênero na segunda, e os sintomas na terceira e na quarta.

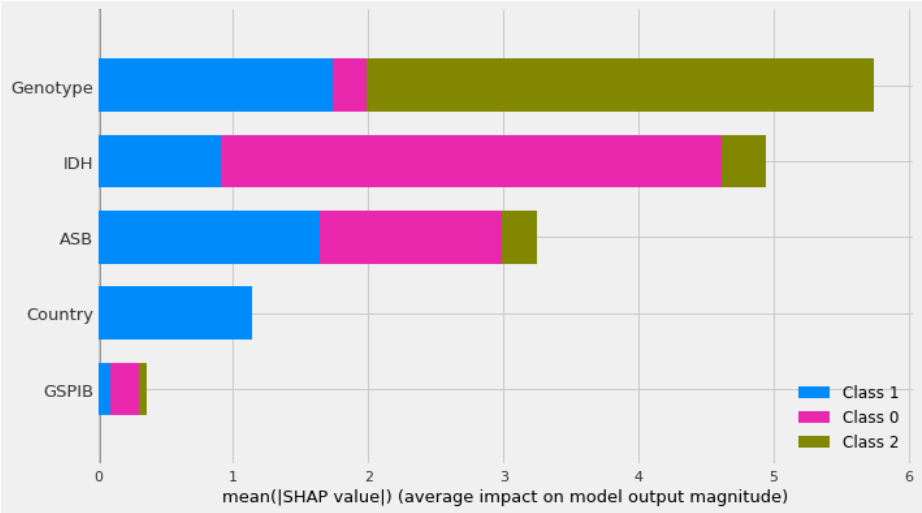
A estatística foi fortemente utilizada para levantar as informações encontradas nos *clusters*. Para as variáveis numéricas, a média, os limites inferior e superior, a frequência absoluta e a frequência relativa percentual puderam dar uma noção de como os dados estão distribuídos, e se existe alguma exclusividade ou predominância. Já para os dados categóricos, foi utilizada a moda, para se ter estas percepções. O uso de gráficos também foram utilizados para compreender as informações nos conjuntos de dados. As bibliotecas *Numpy*, *Plotly* e *Seaborn* ofereceram os recursos necessários para gerar e representar os resultados.

5.2.1 Interpretação dos resultados da primeira amostra.

A Figura 32 mostra um gráfico de barras horizontal, contendo as informações acerca das variáveis mais significativas para a criação dos *clusters*. Como pode ser observado, o genótipo, o IDH e o ASB se destacaram mais dentre as outras no agrupamento. O país e o GSPIB também influenciam nos resultados, porém, possuem menos impacto com o uso do algoritmo *K-prototype*. É importante destacar que nenhuma variável foi irrelevante para o processo.

Os dados dos protótipos gerados para a primeira amostra podem ser visualizados no Tabela 5. Ao associar as informações junto ao gráfico da Figura 32, os *clusters* 1, 2 e 3 equivalem respectivamente às classes 0, 1 e 2. Com a análise dos protótipos e do gráfico, algumas percepções puderam ser levantadas. O genótipo teve forte impacto no segundo e no terceiro *cluster*, onde todos os registros encontrados foram, respectivamente, do ECSA e do Asiático. Já para o primeiro, 85,36% dos países com IDH entre 0.43 e 0.67 estão contidos neste *cluster*.

Figura 32 – Gráfico de barras contendo em ordem, de cima para baixo, as variáveis mais significativas para o agrupamento da primeira amostra.



Fonte – O autor.

Tabela 5 – Dados dos protótipos encontrados com o *K-prototype* na primeira amostra

	Cluster 1	**Cluster 2**	**Cluster 3**
Objetos Totais	1.213	687	650
Genótipo	ECSA	ECSA	Asian and Caribbean
País	Índia	Tailândia	Nicarágua
IDH	0.57	0.76	0.74
ASB	0.41	0.94	0.83
GSPIB	0.03	0.05	0.08

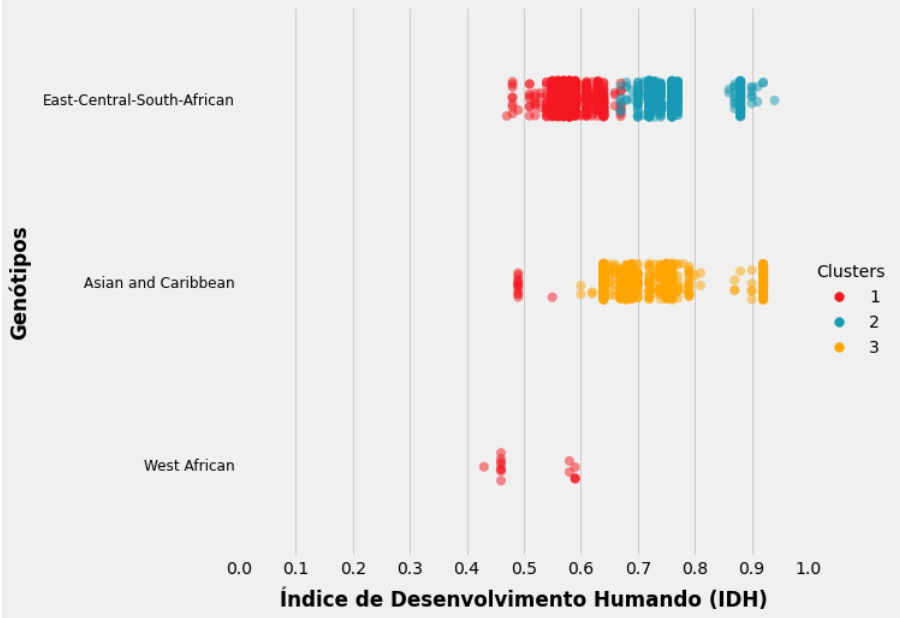
Fonte – O autor.

A partir destas percepções, buscou-se encontrar alguma nova informação a respeito dos genótipos de acordo com os resultados levantados nos protótipos e nas variáveis mais relevantes. Ao explorar a relação entre os genótipos e o IDH, como mostrado na Figura 33, o *cluster* 1, mesmo com a grande influência desta variável numérica, os genótipos não foram tão impactantes, fazendo com que o Asiático, o ECSA e o West African fossem agrupados nele. Já ao explorar o terceiro *cluster*, que foi o mais influenciado pelos genótipos, não foi identificado nenhum padrão exclusivo é possível verificar que todos os casos registrados contidos neste *cluster* estão concentrados em países com IDH superior a 0.6. Como ele equivale a 97,89% (636 registros) do total de casos de pacientes infectados pelo genótipo Asiático, é possível observar uma tendência a partir deste aspecto.

O mesmo pode ser visto ao explorar o ASB. Como visto na Figura 34, o limite inferior e superior para o terceiro *cluster* é de 0.45 e 1.00. Diante disso, é possível observar também uma tendência, na qual, os casos de pacientes infectados com o genótipo Asiático foram em países cuja porcentagem da população com acesso a saneamento básico ultrapassa 45%. Já

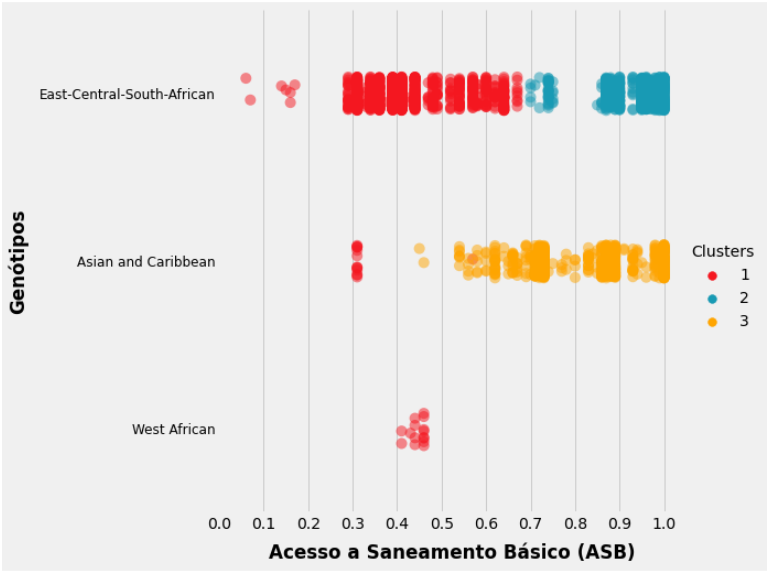
referente à variável ASB, não foi possível encontrar nenhuma relação que pudesse ser visível.

Figura 33 – Gráfico de distribuição de dados das variáveis genótipo e IDH nos *clusters* da primeira amostra.



Fonte – O autor.

Figura 34 – Gráfico de distribuição de dados das variáveis genótipo e ASB nos *clusters* da primeira amostra.

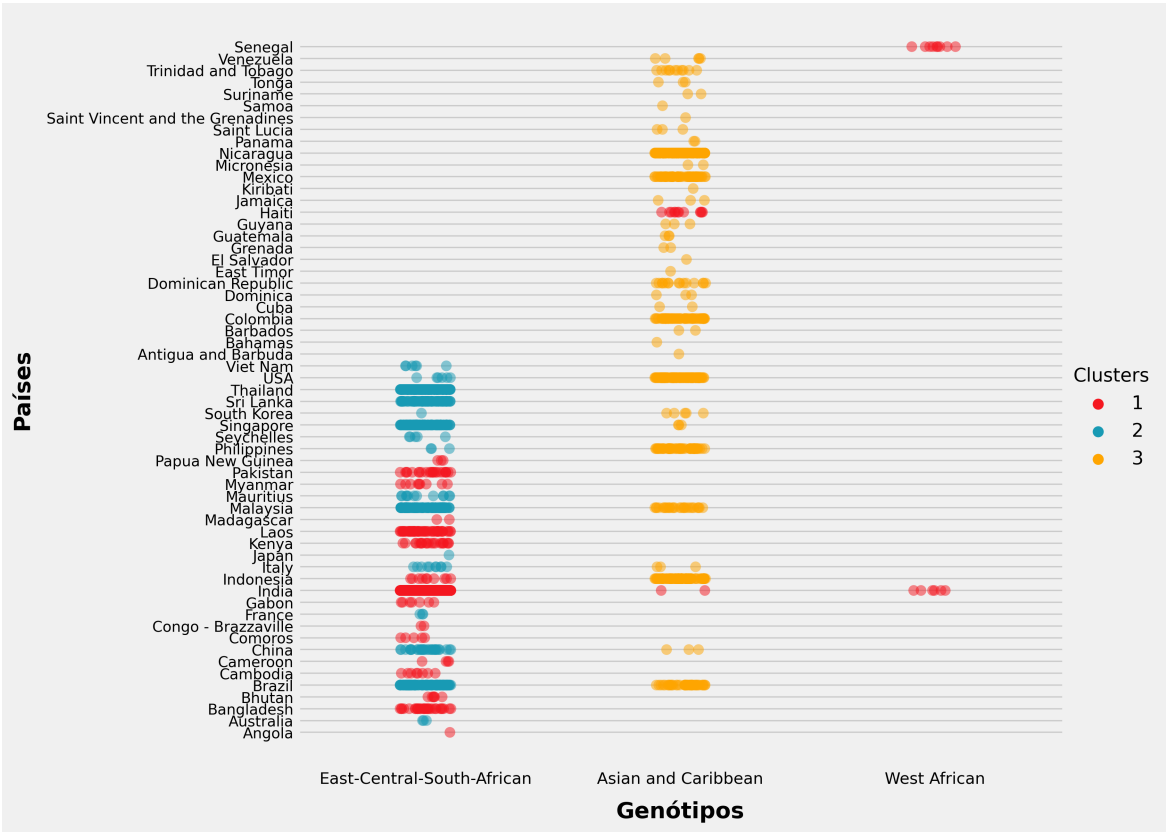


Fonte – O autor.

Já na relação entre os genótipos e os países, é possível observar nos gráfico da Figura 35 que existem alguns países que que aparecem exclusividade no segundo e no terceiro *cluster*. Ao analisar o gráfico, e ter esta percepção, é possível identificar que, como esta variável é fortemente influenciada pelos genótipos, existem alguns países onde muitos casos registrados

de infecção pelo vírus chikungunya foram somente do ECSA ou do Asiático. Isso pode ser visto com ocorrências exclusivas no *cluster* 2, na Tailândia com 32,75% de infectados (225 pessoas) e na Sri Lanka com 10,63% de infectados (73 pessoas), países localizados no continente asiático. No *cluster* 3, os registros foram na Nicarágua com 22,92% de infectados (149 pessoas), na Colômbia com 7,85% de infectados (51 pessoas) e no México com 7,38% de infectados (48 pessoas), países localizados no continente americano.

Figura 35 – Gráfico de distribuição de dados das variáveis genótipo e País nos *clusters* da primeira amostra.

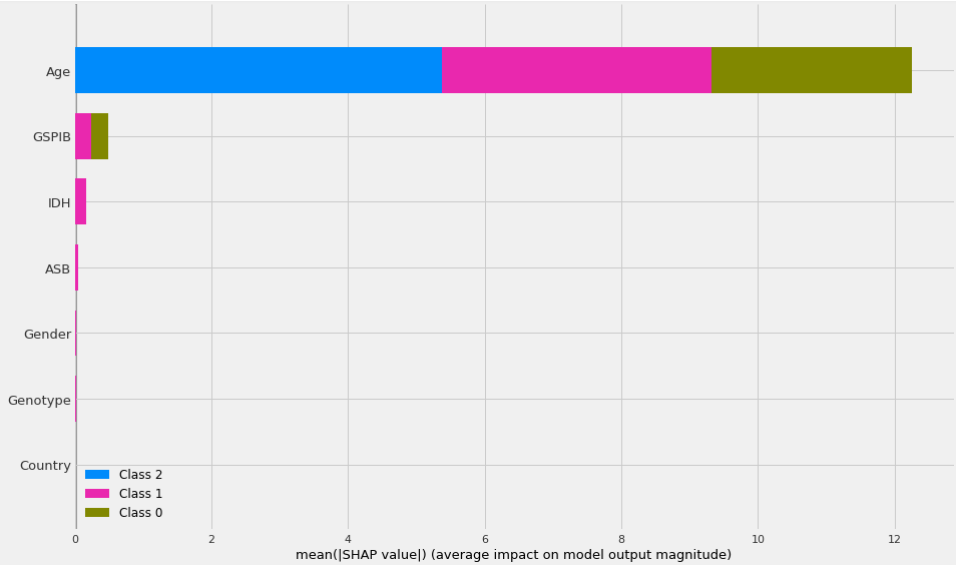


Fonte – O autor.

5.2.2 Interpretação dos resultados da segunda amostra.

Referente à segunda amostra, a Figura 36 mostra um gráfico de barras que indica quais das variáveis utilizadas teve mais influência no agrupamento performado pelo *K-prototype*. Como pode ser visto, a idade possui a importância mais significativa dentre todas. O GSPIB e o IDH vem logo após, apresentando uma bem menor, se comparado à idade. Já o gênero e o genótipo possuem pouca influência na criação dos *clusters*. Por fim, o país foi irrelevante para o processo.

Figura 36 – Gráfico de barras contendo em ordem, de cima para baixo, as variáveis mais significativas para o agrupamento da segunda amostra.



Fonte – O autor.

Já no Tabela 6 é possível visualizar as informações acerca dos protótipos resultantes. Cada *cluster* ficou com intervalos de idade exclusivos, que foi a variável numérica mais significativa para os agrupamentos desta amostra. No primeiro *cluster*, o valor da média para esta variável foi de 60 anos, com o limite inferior em 48 e o superior em 88. Já no segundo, o valor da média foi de 34 anos, com os limites em 23 e 46. No terceiro *cluster*, a média identificada foi de 9 anos, com os limites em 1 e 21.

Tabela 6 – Dados dos protótipos encontrados com o *K-prototype* na segunda amostra

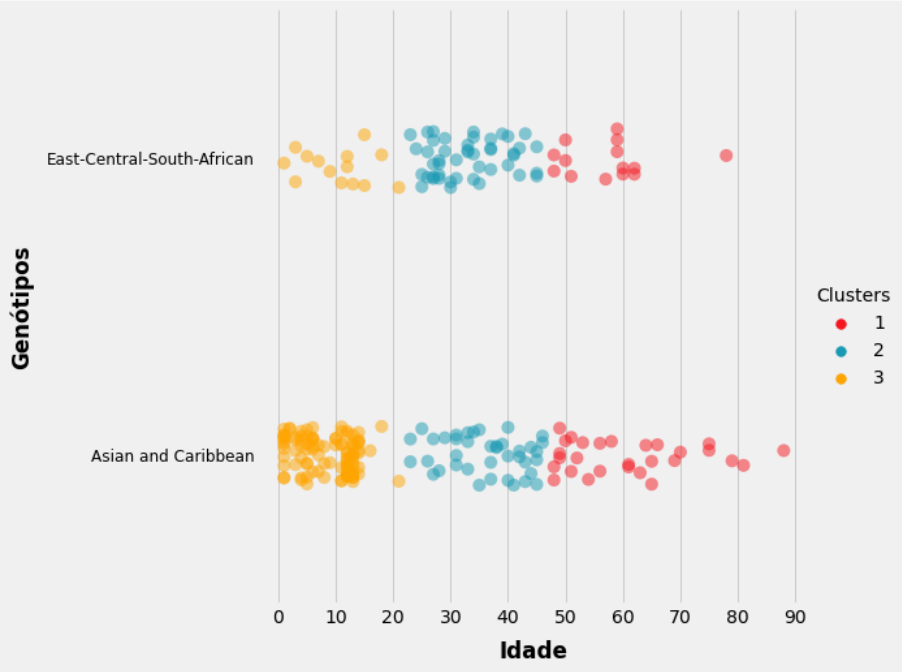
	Cluster 1	**Cluster 2**	**Cluster 3**
Objetos Totais	42	85	110
Genótipo	Asian and Caribbean	ECSA	Asian and Caribbean
País	USA	Índia	Nicarágua
IDH	0.83	0.71	0.65
ASB	0.88	0.65	0.72
GSPIB	0.12	0.08	0.08
Idade	60	34	9
Gênero	Feminino	Feminino	Masculino

Fonte – O autor.

Baseado nas faixas de idade e nas informações a respeito dos protótipos das variáveis categóricas, foi possível identificar algumas tendências relacionadas a esta amostra. Existe uma predominância dos genótipos e do gênero em cada um dos *clusters*. Referente a esta primeira variável, só existem registros do ECSA e do Asiático. O gráfico de dispersão da Figura 37 mostra como estão concentrados os valores da relação entre os genótipos e as idades. No *cluster* 1 a

predominância foi do Asiático, com 66,67% dos casos (28 registros) no total. Já no segundo, o ECSA apareceu em mais casos, totalizando 54,12% (46 registros) do total. No terceiro *cluster*, a predominância foi do Asiático, com 87,27% (96 registros) dos casos.

Figura 37 – Gráfico de distribuição de dados das variáveis genótipo e idade nos *clusters* da segunda amostra.



Fonte – O autor.

Já referente a variável gênero, a Figura 38 mostra a relação dela com a idade. No primeiro *cluster* a maior quantidade registrada é de mulheres, com 61,90% dos casos (26 registros). As mulheres também aparecem mais vezes no segundo, com 54,12% dos casos (46 registros). Já os homens possuem mais frequência no terceiro *cluster*, com 66,36% dos casos (73 registros).

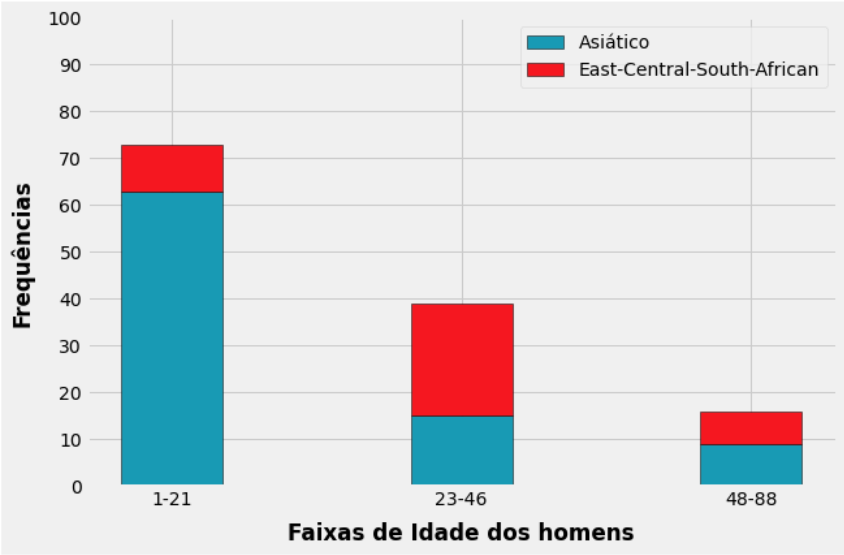
A partir da concentração dos valores nos gráficos das Figuras 37 e 38, foi possível explorar as três variáveis, de forma a se buscar tendências mais detalhadas a respeito da infecção do vírus chikungunya em homens e mulheres de faixas de idade diferentes. A começar pelos homens, a Figura 39 mostra um gráfico de barras que representa as frequências dos casos de infecção pelo ECSA e pelo Asiático nos 3 intervalos de idade.

A partir deste gráfico, foi possível observar que ocorreram mais infecções pelo vírus chikungunya nos homens mais jovens, entre 1 e 21 anos. Dentro desta mesma faixa de idade, a maioria dos casos foram de infecção pelo genótipo Asiático, totalizando 86,30% das pessoas do

Figura 38 – Gráfico de distribuição de dados das variáveis gênero e idade nos *clusters* da segunda amostra.

Fonte – O autor.

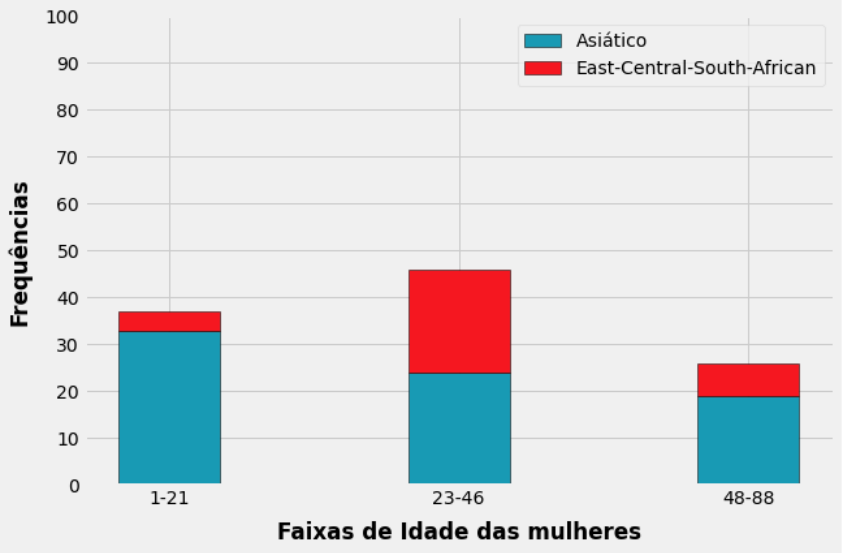
Figura 39 – Gráfico de barras com a frequência de infecção dos genótipos da chikungunya em homens de diferentes faixas de idade.



Fonte – O autor.

sexo masculino (63 registros). Outra predominância bem visível foi vista em homens entre 23 e 46 anos, que contraíram o genótipo ECSA em 61,54% dos casos (24 registros). Já para a última faixa de idade, a quantidade de casos para cada genótipo ficou próxima, com um total de 56,25% (9 registros) para o Asiático e 43,75% (7 registros) para o ECSA. Já para o sexo feminino, a Figura 40 representa o gráfico de barras das frequências dos casos de infecção pelos 2 genótipos, também nos 3 intervalos de idade.

Figura 40 – Gráfico de barras com a frequência de infecção dos genótipos da chikungunya em mulheres de diferentes faixas de idade.



Fonte – O autor.

O vírus chikungunya atingiu mais mulheres entre 23 e 46 anos, como pode ser visto no gráfico. O genótipo Asiático dominou todas as 3 faixas de idade, e isto foi mais perceptível em pessoas do sexo feminino mais jovens, entre 1 e 21 anos, com um total de 89,19% dos casos (33 registros), e também naquelas com idade mais avançada, entre 48 e 88 anos, com um total de 73,08% dos casos (19 registros). Por fim, para as mulheres entre 23 e 46 anos, a quantidade de casos para cada genótipo ficou bem próxima, com 52,17% dos casos (24 registros) para o Asiático e 47,83% dos casos (22 registros) para o ECSA.

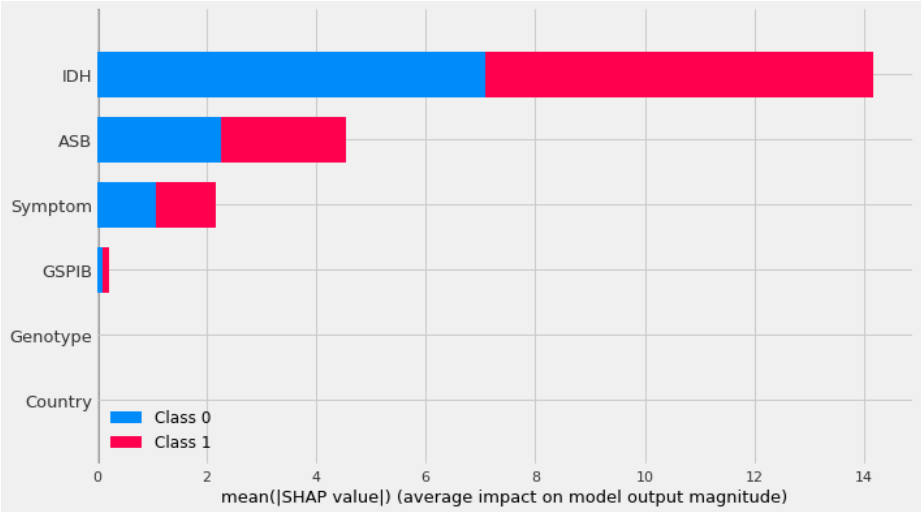
5.2.3 Interpretação dos resultados da terceira amostra.

As variáveis que mais influenciaram no agrupamento dos dados da terceira amostra podem ser visualizadas através do gráfico exposto na Figura 41. Pode-se observar que o IDH foi o que possuiu uma importância mais significativa. Logo após, aparecem o ASB e os sintomas, apresentando uma influência bem menor que o IDH. Já o genótipo que é a principal variável do estudo e o país se mostraram irrelevantes para os *clusters* gerados nesta amostra.

Já na Tabela 7 pode ser visto os valores dos protótipos resultantes de cada *cluster*. A partir destas informações, e diante das variáveis mais relevantes, foi possível observar que no primeiro *cluster* estão contidos 95,83% dos casos em países que apresentaram o IDH entre 0.49 e 0.68. Já no segundo, estão contidos 52,27% dos casos em países com o valor do IDH entre 0.56 e 0.88. Porém, este valor não reflete a porcentagem real dessa variável, devido a um fato

que compromete a análise para as variáveis numéricas deste conjunto de dados.

Figura 41 – Gráfico de barras contendo em ordem, de cima para baixo, as variáveis mais significativas para o agrupamento da terceira amostra.



Fonte – O autor.

Tabela 7 – Dados dos protótipos encontrados com o *K-prototype* na terceira amostra

	Cluster 1	**Cluster 2**
Objetos Totais	276	161
Genótipo	ECSA	ECSA
País	Índia	Malásia
Sintoma	Febre	Irritação na Pele
IDH	0.56	0.78
ASB	0.37	0.95
GSPIB	0.04	0.03

Fonte – O autor.

A natureza com a qual os dados da terceira amostra estão dispostos atrapalha diretamente nas possíveis análises que viriam a ser feitas, de acordo como o algoritmo *K-prototype* agrupou este conjunto de dados. No processo de criação, foi preciso replicar as linhas de cada caso para relacionar com os sintomas identificados no paciente. Como esta variável apresentou uma pequena relevância para o agrupamento, ocorreu que, em algumas situações, um mesmo caso foi replicado em *clusters* diferentes, juntamente com os valores de genótipo, país, IDH, ASB e GSPIB. Foi observado que este problema ocorreu com os dados onde apareceram os valores da moda dos protótipos dos sintomas: a febre no *cluster* 1 e a irritação na pele no *cluster* 2. A Figura 42 exemplifica esta situação para o caso de acesso GU199351.

Figura 42 – Primeiros registros do conjunto de dados da terceira amostra, com as informações dos acessos e dos *clusters*.

Acession	Genotype	Country	IDH	ASB	GSPIB	Symptom	Clusters
GU199350	East-Central-South-African	China	0.68	0.7	0.04	fever	0
GU199351	East-Central-South-African	China	0.68	0.7	0.04	arthralgia	1
GU199351	East-Central-South-African	China	0.68	0.7	0.04	fever	0
GU199352	East-Central-South-African	China	0.68	0.7	0.04	conjunctivitis	1
GU199352	East-Central-South-African	China	0.68	0.7	0.04	bleeding	1
GU199352	East-Central-South-African	China	0.68	0.7	0.04	fever	0
GU199353	East-Central-South-African	China	0.68	0.7	0.04	high fever	1
GU199353	East-Central-South-African	China	0.68	0.7	0.04	conjunctivitis	1

Fonte – O autor.

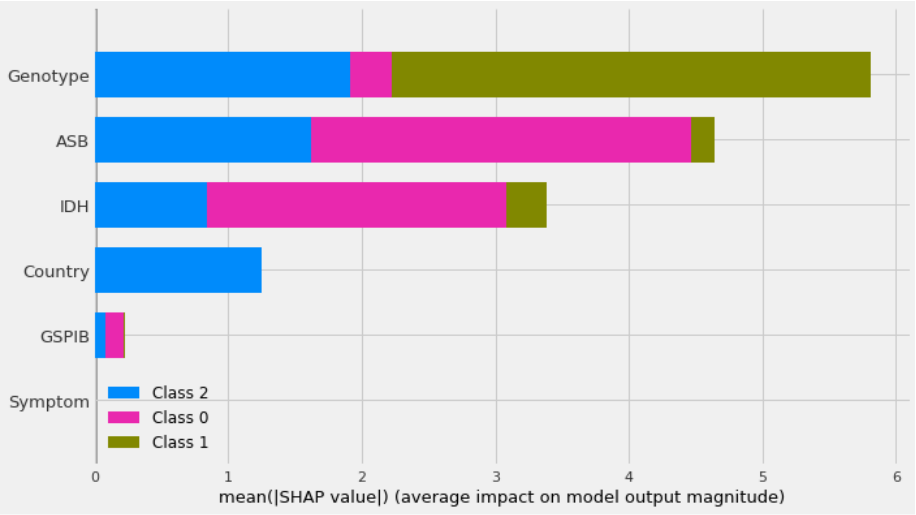
Esse problema possivelmente tem relação com o valor de k escolhido para a amostra. A coluna dos sintomas possui um total de 34 valores únicos para os sintomas, e, por ser uma variedade grande para uma variável com dados do tipo categórico, ela não deveria ter relevância alguma no agrupamento, e os casos iriam ficar no mesmo *cluster* após a aplicação da técnica (isso foi observado na amostra seguinte, que possui a mesma natureza). Devido ao fato da técnica gerar somente 2 *clusters*, é possível que isso tenha influenciado no aparecimento deste problema. Diante dessa situação, a interpretação para este conjunto de dados em específico foi inconclusiva.

5.2.4 Interpretação dos resultados da quarta amostra.

Na quarta amostra, existe a mesma situação da replicação das linhas para a adequação dos sintomas. Porém, como pode ser visto no gráfico da Figura 43, os sintomas não tiveram impacto algum durante o agrupamento, onde cada caso replicado ficou no mesmo *cluster*, sendo possível levantar as informações acerca deste conjunto de dados. O genótipo foi a variável que mais influenciou na geração dos *clusters*, seguido pelo ASB e pelo IDH. Já o país e o GSPIB tiveram uma relevância menor, se comparados com as outras variáveis.

O Tabela 8 mostra os protótipos encontrados para esta amostra após a execução do *K-prototype*. A partir da análise dele e da Figura 43, é possível observar que o primeiro *cluster* é fortemente influenciado pela variável ASB, contendo casos registrados nos países onde a taxa da população que possui acesso ao saneamento básico tiveram registros entre 6 e 67%. Já os outros *clusters* tiveram nos genótipos o seu fator mais preponderante, onde no segundo, todos os registros são do Asiático, enquanto no terceiro, do ECSA.

Figura 43 – Gráfico de barras contendo em ordem, de cima para baixo, as variáveis mais significativas para o agrupamento da quarta amostra



Fonte – O autor.

Tabela 8 – Dados dos protótipos encontrados com o *K-prototype* na quarta amostra

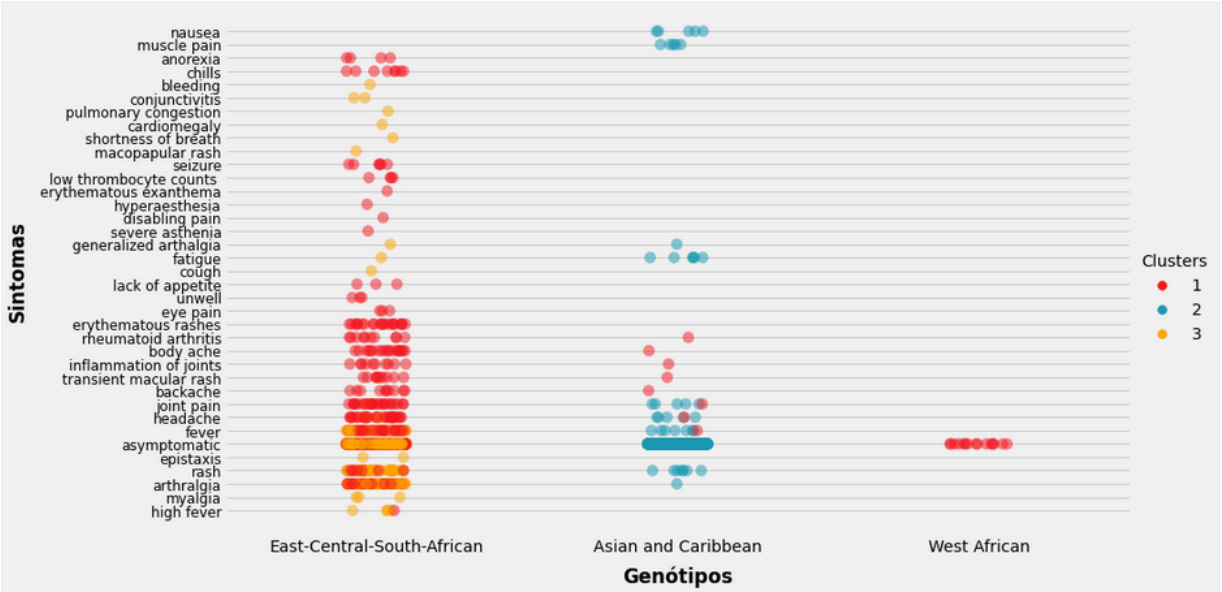
	Cluster 1	**Cluster 2**	**Cluster 3**
Objetos Totais	1439	686	781
Genótipo	ECSA	Asiático	ECSA
País	Índia	Nicarágua	Tailândia
Sintoma	Assintomático	Assintomático	Assintomático
IDH	0.57	0.74	0.77
ASB	0.41	0.83	0.94
GSPIB	0.04	0.08	0.04

Fonte – O autor.

Nenhuma relação ou padrão foi verificado na relação entre os genótipos e os sintomas com a aplicação da técnica, o que pode ser visto no gráfico da Figura 44. Houve a tentativa de analisar quais sintomas foram exclusivos nos *clusters* 2 e 3, já que o ECSA e o Asiático são dominantes em cada um. Porém, as informações encontradas foram irrelevantes para o resultado, analisando o total de casos contidos em cada *cluster*. Como por exemplo, no terceiro, foi visto que somente 1 pessoa apresentou sintomas como anorexia, dificuldade respiratória, sangramentos, congestão pulmonar, dentre outras doenças, o que é muito pouco para um total de 781 registros de sintomas e 689 registros de casos únicos. O mesmo foi observado no segundo *cluster*, com 5 pessoas apresentando náuseas e dores musculares.

Porém, algumas informações puderam ser observadas ao analisar o gráfico da Figura 44, e que independem dos resultados do *clustering*. Uma delas, é que existe uma tendência ao genótipo ECSA desenvolver uma variedade maior de sintomas, se comparado aos outros dois. Outra percepção importante é que todos os registros do genótipo West African são de pessoas

Figura 44 – Gráfico de distribuição de dados das variáveis genótipo e sintomas nos *clusters* da quarta amostra.



Fonte – O autor.

assintomáticas, ou seja, não apresentaram nenhuma manifestação dos sintomas. Para confirmar estas informações ao longo do tempo, seria interessante observar se, com a inclusão de novos dados do vírus chikungunya no ABVdb, estas tendências se manterão.

O comportamento das variáveis ASB, IDH, GSPIB e país são semelhantes ao que foi visto na primeira amostra, devido a irrelevância dos sintomas. A única diferença é que no agrupamento da amostra, e nos resultados do *cluster* 1, o ASB é mais relevante que o IDH. O fato dos valores para os sintomas estarem bem variados, e que o algoritmo *K-prototype* leva em consideração a moda na sua medida de similaridade para variáveis categóricas, fez com que o *clustering* não identificasse nenhum padrão desta variável com as outras. Na terceira amostra, essa situação foi um problema, e atrapalhou na interpretação dos resultados. Já na quarta, ele não impacta, e não produz os resultados esperados para a amostra. Uma possível solução seria categorizar esta variável, de forma que essa variedade diminua drasticamente, e mesmo assim represente os sintomas que o paciente apresentou quando infectado pelo vírus chikungunya.

6 CONSIDERAÇÕES FINAIS

Este projeto apresentou uma metodologia em torno de *Data Science* para levantar informações úteis a respeito de aspectos epidemiológicos e socioeconômicos em torno de casos confirmados do vírus chikungunya. A partir da aplicação, avaliação e interpretação do algoritmo *K-prototype*, foi possível encontrar algumas relações e tendências a respeito das variáveis utilizadas.

No que diz respeito às variáveis utilizadas, existe uma escassez na obtenção de dados públicos e consistentes para realizar análises mais concretas. Dentre diversas características socioeconômicas que poderiam ser observadas no trabalho, somente o IDH, a porcentagem da população com acesso a saneamento básico e a porcentagem do PIB gasto em saúde tiveram conjuntos de dados consistentes, e puderam ser utilizadas no projeto. Isso limitou a variedade de relações que poderiam ser encontradas com a análise.

Referente às métricas utilizadas, o Índice de Rand Ajustado se mostrou mais fidedigno a mostrar o quão similares foram os resultados dos agrupamentos realizados, se comparado ao Índice de Jaccard. Como os protótipos iniciais do *K-prototype* são gerados aleatoriamente, os resultados desta primeira métrica não são afetados por este fator. Ao analisar a média dos valores resultantes dela, foi possível observar que o algoritmo, na maioria das vezes, agrupou de forma bem similar os dados em cada uma das amostras.

Sobre os resultados, foi possível destacar informações muito importantes sobre os genótipos do vírus chikungunya, relacionados à idade dos pacientes. Sobre as variáveis socioeconômicas, foram observadas tendências entre elas, seus países de ocorrência e os genótipos. Já os resultados referente aos sintomas não puderam ser obtidos, pois, devido a sua grande variedade e ele ser do tipo categórico, se mostrou irrelevante para a técnica de *clustering*.

Sobre os trabalhos futuros, de forma a dar relevância para os sintomas e buscar por possíveis relações em torno desta variável, seria importante a categorização da variedade destes valores com o auxílio de um especialista. Caso o estudo deseje ser feito da forma como ele está distribuído, a utilização de uma outra técnica de *Data Mining* seria pertinente para se obter melhores resultados. Por fim, a utilização de outros tipos de variáveis, além das

socioeconômicas, também seria interessante, de forma a buscar novas relações com os três genótipos da chikungunya.

Os resultados levantados com este projeto podem auxiliar em futuros estudos a respeito dos genótipos da chikungunya, principalmente o ECSA e o Asiático. Também, seria interessante observar se com novas inserções de dados sobre o vírus no ABVdb, as tendências identificadas irão permanecer. Por fim, o objetivo foi cumprido, ao buscar padrões entre os genótipos do vírus chikungunya e as variáveis em torno dos casos.

REFERÊNCIAS

ABUBAKAR, S. et al. Reemergence of Endemic Chikungunya, Malaysia. *Emerging Infectious Diseases*, v. 13, n. 1, p. 147–149, 2007.

AGARWAL, N.; KOTI, S.; SARAN, S.; KUMAR, A. S. Data mining techniques for predicting dengue outbreak in geospatial domain using weather parameters for New Delhi, India. *Current Science*, v. 114, n. 11, p. 2281–2291, 2018.

ALBUQUERQUE, M.; SILVA, E.; BARROS, K.; JUNIOR, S. F. Comparação entre coeficientes similaridade um aplicação em ciências florestais. 12 2016.

BARATA, R. B. Como e por que as desigualdades sociais fazem mal à saúde. *Editora FIOCRUZ*, 2009.

BESSAUD, M. et al. Chikungunya Virus Strains, Reunion Island Outbreak. *Emerging Infectious Diseases*, v. 12, n. 10, p. 1604–1605, 2006.

BURT, F. et al. Chikungunya virus: an update on the biology and pathogenesis of this emerging pathogen. *The Lancet. Infectious diseases*, v. 17 4, p. e107–e117, 2017.

CASSIANO, K. M. Análise de séries temporais usando análise espectral singular (ssa) e clusterização de suas componentes baseada em densidade. *Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio*, 2015.

CAVIQUE, L. Big data e data science. *Boletim da APDIO*, n. 51, p. 11–14, 2014.

CHAHAR, H. S. et al. Co-infections with Chikungunya Virus and Dengue Virus in Delhi, India. *Emerging Infectious Diseases*, v. 15, n. 7, p. 1077–1080, 2009.

CONCEIÇÃO, F. R. da S. Análise comparativa do desempenho de técnicas de data mining com atributos mistos sob uma base de dados epdemiológica. *Universidade Estadual da Bahia*, 2020.

CRESWELL, J. W.; ROCHA, L. d. O. d.; SILVA, M. I. da Costa e. *Projeto de pesquisa métodos qualitativo, quantitativo e misto*. Porto Alegre: Artmed, 2007. 161-178 p. ISBN 978-85-363-0892-0.

DEY, L.; AHMAD, A. A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, p. 503–527, 2007.

EDNA, H. V. Algoritmo de clustering basado en entropía para descubrir grupos en atributos de tipo mixto. *Centro de Investigacion y de Estudios Avazados del Instituto Politecnico Nacional Departamento de Ingenieria Electrica Sección de Computación*, 2006.

EDWARDS, C. J. et al. Molecular diagnosis and analysis of Chikungunya virus. *Journal of Clinical Virology*, v. 39, n. 4, p. 271–275, 2007.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, v. 17, n. 3, 1996.

FUSCO, F. M. et al. di ritorno dall'Oceano Indiano e rischio di introduzione nel territorio italiano. p. 8, 2006.

HAN, J.; KAMBER, M. Data mining: Concepts and techniques. 2001.

HISAMUDDIN, M. et al. Co-circulation of Chikungunya and Dengue viruses in Dengue endemic region of New Delhi, India during 2016. *Epidemiology and Infection*, v. 146, n. 13, p. 1642–1653, 2018.

HUANG, Z. Clustering large data sets with mixed numeric and categorical values. *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, p. 21–34, 1997.

HUANG, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, v. 2, n. 3, p. 283–304, 1998.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of Classification*, v. 2, p. 193–218, 1985.

KUMAR, N. P.; JOSEPH, R.; KAMARAJ, T.; JAMBULINGAM, P. A226V mutation in virus during the 2007 chikungunya outbreak in Kerala, India. *Journal of General Virology*, v. 89, n. 8, p. 1945–1948, 2008.

LEO, Y. S. et al. Chikungunya Outbreak, Singapore, 2008. *Emerging Infectious Diseases*, v. 15, n. 5, p. 836–837, maio 2009.

LEWTHWAITE, P. et al. Chikungunya Virus and Central Nervous System Infections in Children, India. *Emerging Infectious Diseases*, v. 15, n. 2, p. 329–331, 2009.

LIM, C.-K. et al. Chikungunya Virus Isolated from a Returnee to Japan from Sri Lanka: Isolation of Two Sub-Strains with Different Characteristics. *The American Journal of Tropical Medicine and Hygiene*, v. 81, n. 5, p. 865–868, 2009.

LIU, S.-Q. et al. Detection, isolation, and characterization of chikungunya viruses associated with the Pakistan outbreak of 2016–2017. *Virologica Sinica*, v. 32, n. 6, p. 511–519, 2017.

LOPES, N.; NOZAWA, C.; LINHARES, R. E. C. Características gerais e epidemiologia dos arbovírus emergentes no Brasil. *Revista Pan-Amazônica de Saúde*, v. 5, n. 3, p. 55–64, 2014. ISSN 2176-6223.

MANNILAT, H. Data mining: machine learning, statistics, and databases. *Department of Computer Science*, 1996.

MATHULAMUTHU, S. S.; ASIRVADAM, V. S.; DASS, S. C.; GILL, B. S. Predicting dengue cases by aggregation of climate variable using manifold learning. *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, p. 535–540, 2017.

MUZAKKI, M.; NHITA, F. The spreading prediction of Dengue Hemorrhagic Fever (DHF) in Bandung regency using K-means clustering and support vector machine algorithm. *Institute of Electrical and Electronics Engineers Inc.*, p. 453–458, 2018.

NG, L.-C. et al. Entomologic and Virologic Investigation of Chikungunya, Singapore. *Emerging Infectious Diseases*, v. 15, n. 8, p. 1243–1249, 2009.

OMS. *Chikungunya*. 2017. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/chikungunya>. Acesso em: 06-12-2019.

PAROLA, P. et al. Novel Chikungunya Virus Variant in Travelers Returning from Indian Ocean Islands. *Emerging Infectious Diseases*, v. 12, n. 10, p. 1493–1499, 2006.

PARREIRA, R. et al. Dengue virus serotype 4 and chikungunya virus coinfection in a traveller returning from Luanda, Angola, January 2014. *Eurosurveillance*, v. 19, n. 10, 2014.

PASTORINO, B. et al. Epidemic resurgence of Chikungunya virus in democratic Republic of the Congo: Identification of a new central African strain. *Journal of Medical Virology*, v. 74, n. 2, p. 277–282, 2004.

PEYREFITTE, C. N. et al. Circulation of Chikungunya virus in Gabon, 2006–2007. *Journal of Medical Virology*, v. 80, n. 3, p. 430–433, 2008.

PISTONE, T. et al. An imported case of Chikungunya fever from Madagascar: Use of the sentinel traveller for detecting emerging arboviral infections in tropical and European countries. *Travel Medicine and Infectious Disease*, v. 7, n. 1, p. 52–54, 2009.

PRABHA, K.; VISALAKSHI, N. Particle swarm optimization based k-prototype clustering algorithm. In: . [S.l.: s.n.], 2015.

RAND, W. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, v. 66, p. 846–850, 1971.

RESTOVIC, M. I. V. Um banco de dados de epidemiologia molecular para os vírus dengue, zika e chikungunya. *Fundação Oswaldo Cruz*, 2018.

SAM, I.-C. et al. Chikungunya virus of Asian and Central/East African genotypes in Malaysia. *Journal of Clinical Virology*, v. 46, n. 2, p. 180–183, 2009.

SANTHOSH, S. et al. Appearance of EI: A226V mutant Chikungunya virus in Coastal Karnataka, India during 2008 outbreak. *Virology Journal*, v. 6, n. 1, p. 172, 2009.

SAÚDE, M. da. *Boletim Epidemiológico 22 - Monitoramento dos casos de arboviroses urbanas transmitidas pelo Aedes (dengue, chikungunya e Zika), Semanas Epidemiológicas 1 a 34*. 2019. Disponível em: <http://portalarquivos2.saude.gov.br/images/pdf/2019/setembro/11/BE-arbovirose-22.pdf0>. Acesso em: 06-12-2019.

SAÚDE, M. da. *Quase mil cidades podem ter surto de dengue, zika e chikungunya no país*. 2019. Disponível em: <http://www.saude.gov.br/noticias/agencia-saude/45407-quase-mil-cidades-podem-ter-surto-de-dengue-zika-e-chikungunya-no-pais>. Acesso em: 2019-10-20.

SHU, P.-Y. et al. Two Imported Chikungunya Cases, Taiwan. *Emerging Infectious Diseases*, v. 14, n. 8, 2008.

SIRIYASATIEN, P.; CHADSUTHI, S.; JAMPACHAISRI, K.; KESORN, K. Dengue epidemics prediction: A survey of the state-of-the-art based on data science processes. *IEEE Access*, v. 6, p. 53757–53795, 2018.

THAKUR, P.; KAUR, S. An intelligent system for predicting and preventing chikungunya virus. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, p. 3483–3492, 2017.

VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, v. 3, n. 4, p. 209–235, 2010. ISSN 1932-1864.

YERGOLKAR, P. et al. Chikungunya Outbreaks Caused by African Genotype, India. *Emerging Infectious Diseases*, v. 12, n. 10, p. 1580–1583, 2006.

YOGAPRIYA, P.; GEETHA, P. Dengue disease detection using k-means, hierarchical, kohonen-som clustering. *International Journal of Innovative Technology and Exploring Engineering*, v. 8, n. 10, p. 904–907, 2019.

ZHENG, K. et al. Genetic analysis of chikungunya viruses imported to mainland China in 2008. *Virology Journal*, v. 7, n. 1, p. 8, 2010.

APÊNDICES

APÊNDICE A – Implementação do algoritmo *K-prototype* na linguagem Python

Para a correta implementação do algoritmo K-prototype utilizado no projeto, uma função foi criada para auxiliar na obtenção das informações a respeito dos *clusters* gerados, que serão utilizadas para a análise. A função *execKPrototype()* recebe como parâmetros o conjunto de dados que será agrupado, a quantidade de *clusters* que serão gerados, os índices das colunas categóricas e a quantidade de execuções do algoritmo. A Figura 45 exibe o código fonte da função. O seu funcionamento é descrito pelo passo a passo a seguir:

Figura 45 – Recorte do código da função execKPrototype() implementada na linguagem Python.

```
def execKPrototype(matrix,nClusters,colunas_c,exec):
  for i in range(0, exec):
    try:
      kprototype = KPrototypes(n_clusters = nClusters, init = 'Huang',verbose=1, n_init=1)
      kprototype.fit_predict(matrix, categorical=colunas_c)
      clusters.append(kprototype.labels_)
      costs.append(kprototype.cost_)

    except:
        break
```

Fonte – O autor

1. É criado um laço de repetição, que é repetido uma quantidade de vezes igual ao valor da quantidade de execuções do algoritmo passado por parâmetro na função. Cada iteração do laço equivale a uma execução do algoritmo. O laço vai tentar executar o agrupamento, mas caso ocorra algum problema durante o processo ele é interrompido.
2. É atribuída a uma variável o construtor da classe que representa o *K-prototype*. Para ele, é passado como parâmetros a quantidade de *clusters* a serem gerados, o método de inicialização como sendo o de 'Huang', o modo de verbosidade para detalhamento das informações na execução do algoritmo e o número de vezes que ele vai ser executado com diferentes centróides (a iteração), com o valor 1.
3. Após isso, os *clusters* são gerados atráves da função *fit_predict()*, que recebe como parâmetros o conjunto de dados a ser agrupado e a indicação das colunas que são categóricas.
4. Após a execução do agrupamento, os valores dos *labels* que indicam a qual *cluster* os dados pertencem e o custo encontrado na execução são salvos em variáveis globais que armazenam estes valores em cada iteração.
5. Os passos 2, 3, e 4 são repetidos, até que se chegue a quantidade desejada de execuções.

APÊNDICE B – Implementação do método do cotovelo na linguagem Python

A implementação do método do cotovelo seguiu baseada na repetitiva execução do algoritmo, alterando a quantidade de *clusters* gerados, e armazenando o custo encontrado para gerar o gráfico. Foi criada uma função auxiliar chamada *elbowKPrototype()*, que recebe como parâmetros o conjunto de dados a ser analisado e as posições das colunas categóricas. A Figura 46 exibe o código fonte da função *elbowKPrototype()*, e o seu funcionamento é descrito pela sequência de passos a seguir:

Figura 46 – Recorte do código da função *elbowKPrototype()* implementada na linguagem Python.

```
def elbowKPrototype(mx,catColumns):
  cost = []
  for cluster in range(1, 11):
    try:
      kprototype = KPrototypes(n_clusters = cluster, init = 'Huang',verbose=1, n_init=1000)
      kprototype.fit_predict(mx, categorical=catColumns)
      cost.append(kprototype.cost_)

    except:
        break

  return cost
```

Fonte – O autor

1. É inicializada uma lista vazia para armazenar os custos encontrados a cada execução do algoritmo.
2. É criado um laço de repetição, que é repetido em um intervalo de 1 a 10, no qual cada valor representa a quantidade de *clusters* a serem gerados.
3. É atribuída a uma variável o construtor da classe que representa o *K-prototype*. É passado como parâmetros a quantidade de *clusters* a serem gerados para aquela execução, o método de inicialização como sendo o de 'Huang', o modo de verbosidade para detalhamento das informações na execução do algoritmo e o número de vezes que ele vai ser executado com diferentes centróides (a iteração), com o valor 1000.
4. Após isso, os *clusters* são gerados através da função *fit_predict()*, que recebe como parâmetros o conjunto de dados a ser agrupado e a indicação das colunas que são categóricas. O valor do melhor custo encontrado após as 1000 iterações é adicionado à lista de custos.
5. Os passos 3 e 4 são repetidos, até que se chegue na décima execução.
6. Por fim, a lista de custos é retornada.

APÊNDICE C – Implementação das métricas

As métricas foram calculadas em torno dos agrupamentos realizados pelos algoritmos na etapa de *Data Mining* do projeto. Para levantar os resultados, foram criadas duas funções, uma para aplicar o Índice de Rand Ajustado e a outra para aplicar o Índice de Jaccard. A começar pela primeira, a função *randIndex()* recebe como parâmetros uma lista contendo os agrupamentos realizados, e o agrupamento referencial, que servirá como base para a avaliação da métrica. A Figura 47 detalha o código fonte da função. O funcionamento da função é descrita pelo passo a passo a seguir:

Figura 47 – Recorte do código da função *randIndex()* implementada na linguagem Python.

```
def randIndex(clusters,best_result):
  rand = []
  for item in range(len(clusters)):
    score = adjusted_rand_score(best_result, clusters[item])
    rand.append(score)

  return rand
```

Fonte – O autor

1. É inicializada uma lista vazia para armazenar cada resultado da avaliação do índice.
2. É criado um laço de repetição, que percorre do primeiro ao último item contido na lista dos agrupamentos realizados.
3. A função *adjusted_rand_score()* é chamada para calcular o valor do Índice de Rand Ajustado, e recebe como parâmetros o agrupamento referencial e o agrupamento da iteração atual do laço. A nota da avaliação é armazenada em uma variável.
4. Esta variável é salva na lista que armazena todos os resultados da avaliação.
5. Os passos 3 e 4 são repetidos, até que se chegue no último agrupamento a ser avaliado.
6. Por fim, a lista de resultados da métrica é retornada.

Já para o Índice de Jaccard, a função *jaccardIndex()* foi criada, e o seu funcionamento é semelhante a *randIndex()*, com uma única mudança exclusiva no passo 3. Para este índice, é chamada a função *jaccard_score()*, que recebe três parâmetros: o agrupamento referencial, o agrupamento da iteração atual do laço e o tipo de média a ser retornada como resultado. A Figura 48 exibe o código fonte desta função.

Figura 48 – Recorte do código da função *jaccardIndex()* implementada na linguagem Python.

```python
def jaccardIndex(clusters, best_result):
  jaccard = []
  for item in range(len(clusters)):
    score = jaccard_score(best_result, clusters[item],average='macro')
    jaccard.append(score)

  return jaccard
```

Fonte – O autor