



**UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
CENTRO CIÊNCIAS E TECNOLOGIAS
CURSO DE LICENCIATURA EM COMPUTAÇÃO**

KLEYTON KLAUS GUEDES DE SOUZA

**O USO DE *BIG DATA ANALYTICS* NA ANÁLISE EPIDEMIOLÓGICA DA
TUBERCULOSE NA PARAÍBA, NO PERÍODO DE 2001 A 2015**

CAMPINA GRANDE – PB

2018

KLEYTON KLAUS GUEDES DE SOUZA

**O USO DE *BIG DATA ANALYTICS* NA ANÁLISE EPIDEMIOLÓGICA DA
TUBERCULOSE NA PARAÍBA, NO PERÍODO DE 2001 A 2015**

Trabalho de Conclusão de Curso apresentado a Universidade Estadual da Paraíba, como requisito para obtenção do título de Licenciado em Computação.

Área de concentração: *Big Data*; Banco de Dados.

Orientador: Prof. Dr. Vladimir Costa de Alencar.

Coorientadora: Prof.^a Esp. Micheline da Silveira Mendes.

Colaborador: Jessé de Oliveira.

**CAMPINA GRANDE – PB
2018**

É expressamente proibido a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano do trabalho.

S729u Souza, Kleyton Klaus Guedes de.

O uso de *big data analytics* na análise epidemiológica da tuberculose na Paraíba, no período de 2001 a 2015 [manuscrito] : / Kleyton Klaus Guedes de Souza, Jessé de Oliveira. - 2018.

55 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Computação) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2018.

"Orientação : Prof. Dr. Vladimir Costa de Alencar, Departamento de Computação - CCT."

"Coorientação: Profa. Esp. Micheline da Silveira Mendes, SSJP - Secretaria de Saúde de João Pessoa"

1. Big data analytics. 2. Mineração de dados. 3. Perfis epidemiológicos.

21. ed. CDD 005.13

KLEYTON KLAUS GUEDES DE SOUZA

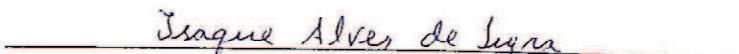
**O USO DE BIG DATA ANALYTICS NA ANÁLISE
EPIDEMIOLÓGICA DA TUBERCULOSE NA PARAÍBA, NO
PERÍODO DE 2001 A 2015**

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Licenciatura plena em Computação da Universidade Estadual da Paraíba, em cumprimento à exigência para obtenção do grau de Licenciado em Computação.

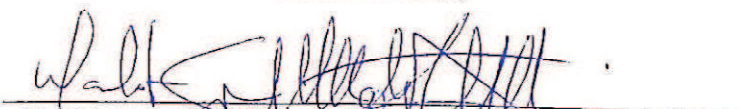
Aprovada em 27 de Fevereiro de 2017.



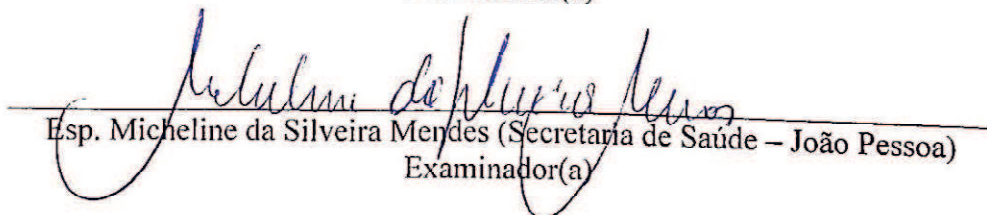
Prof. Dr. Vladimir Costa De Alencar
Orientador(a)



Me. Isaque Alves de Lyra (UEPB)
Examinador(a)



Dr. Marcelo José Siqueira Coutinho de Almeida (IFPB)
Examinador(a)



Esp. Micheline da Silveira Mendes (Secretaria de Saúde – João Pessoa)
Examinador(a)

À Aisha e Snow por serem companheiros fiéis
sempre, DEDICO.

AGRADECIMENTOS

À Halyni Gomes, minha querida esposa, por sua paciência, apoio e atenção.

À professora Micheline da Silveira pelas leituras sugeridas ao longo dessa orientação e pelo apoio nos momentos difíceis.

Ao professor Vladimir Costa pela força motivadora ao longo dessa orientação, por acreditar no sucesso de um aspirante a licenciado e pela dedicação e empolgação ao qual me guiou.

À minha querida “mainha”, Nely Guedes, por acreditar em mim desde a mais tenra idade, mostrando-me o que minha juventude me incapacitava de ver.

Ao meu pai Edvaldo Azevedo e familiares, pelas ausências cotidianas.

Ao meu avô Severino (*in memoriam*), e à minha avó Anália (*in memoriam*), que embora fisicamente ausentes, sempre acreditaram no meu sucesso acadêmico e me viram como um “doutorzinho” desde que aprendi a ler.

Aos professores do Curso de Licenciatura em Computação da UEPB. Em especial, Mizael, Fred, Kátia, Lúcia Serafim e Antônio Carlos (AC) que contribuíram, ao longo destas idas e vindas ao curso, por meio do amor e conhecimento demonstrados em suas disciplinas, e por compartilharem suas experiências de vida.

Aos colegas de classe pelos momentos de amizade e apoio.

"Big Data é básico na gênese de tudo o que é uma tendência hoje: mídia social, celular, nuvem e jogos online."

Chris Lynch

RESUMO

Neste trabalho foram traçados perfis epidemiológicos da tuberculose com as técnicas de *Big Data Analytics* para uma análise descritiva e exploratória em um *dataset* formado com os dados públicos dos municípios da Paraíba, entre os anos de 2001 a 2015. Inicialmente, o *dataset* foi gerado com as informações dos indicadores da tuberculose, consolidados nas bases de dados que compõe o SAGE e dados socioeconômicos contidos no IBGE. O *dataset* passou por etapas de seleção, pré-processamento e avaliação, onde foram extraídas regras de associação com o processo de mineração de dados, fazendo uso do algoritmo apriori. As regras geradas deram origem a perfis epidemiológicos que foram analisados sobre a perspectiva dos indicadores epidemiológicos contidos na bibliografia referente à tuberculose, juntamente com análise de uma especialista em epidemiologia. Foram geradas 43 regras, das quais, as que apresentaram maior confiança foram analisadas para traçar os perfis epidemiológicos da tuberculose no estado. Os resultados obtidos inferem que a tuberculose na Paraíba está intimamente ligada com a baixa renda e falta de saneamento básico. Existe também a possibilidade de inferir uma incidência relativamente alta de casos de tuberculose, mesmo em municípios que possuem acesso à saúde básica e que apresentam IDH-M alto e/ou dentro da média do IDH do Brasil, o que evidencia uma aparente falta de conhecimento a respeito da doença e má distribuição de renda nesses municípios. As regras, portanto, estão de acordo com os relatos na bibliografia sobre a tuberculose.

Palavras-Chave: *Big Data Analytics*; mineração de dados; apriori; tuberculose; perfis; Paraíba.

ABSTRACT

In this work, epidemiological profiles of tuberculosis with the techniques of *Big Data Analytics* were drawn for a descriptive and exploratory analysis in a dataset formed with the public data of the municipalities of Paraíba, between the years of 2001 to 2015. Initially, the dataset was generated with the information from the tuberculosis indicators, consolidated in the databases that make up the SAGE and socioeconomic data contained in IBGE. The dataset went through selection, pre-processing and evaluation stages, where association rules were extracted with the data mining process, making use of the apriori algorithm. The rules generated gave rise to epidemiological profiles that were analyzed from the perspective of the epidemiological indicators contained in the bibliography referring to tuberculosis, together with the analysis of a specialist in epidemiology. A total of 43 rules were generated, of which the ones with the greatest confidence were analyzed to trace the epidemiological profiles of tuberculosis in the state. The results obtained infer that tuberculosis in Paraíba is closely linked with low income and lack of basic sanitation. There is also the possibility of inferring a relatively high incidence of tuberculosis cases, even in municipalities that have access to basic health and that present high HDI-M and / or within the HDI average of Brazil, which evidences an apparent lack of knowledge regarding the disease and poor distribution of income in these municipalities. The rules, therefore, are in line with the reports in the literature on tuberculosis.

Keywords: *Big Data Analytics*; data mining; apriori; tuberculosis; profiles; Paraíba.

LISTA DE ILUSTRAÇÕES

Figura 1: <i>Figura</i> Representando o Processo de KDD	21
Figura 2: Seleção do <i>Itemset</i> “tx_incidencia_tuberculose” para Salvar em CSV	38
Figura 3: Script Codificado em <i>Python</i> 3 - Gerador de Distribuição de Frequências	39
Figura 4: Histograma Gerado no Spyder 3 – <i>Itemset</i> tx_incidencia_tuberculose.csv”	39
Figura 5: Distribuição de Frequência com a Regra de Sturges, com as Classes para o <i>Itemset</i> “tx_incidencia_tuberculose.csv”	40
Figura 6: Função SQL que Retorna o Nome do <i>Itemset</i> da “tx_incidencia_bacilifera” com seu Intervalo de Classe	40
Figura 7: Seleção de Valores de Todos o <i>Itemset</i> para criar o <i>dataset</i> final com Valores Discretizados	41
Figura 8: Resultado para Alguns <i>Itemset</i> do <i>Dataset</i> Final Discretizados	41
Figura 9: Script em <i>Python</i> com a Biblioteca Pandas e Apyori para Minerar o <i>Dataset</i>	42

LISTA DE GRÁFICOS

Gráfico 1: Demonstrativo de Crescimento – Confiança x Interesse em Relação às Regras Geradas.....	43
---	----

LISTA DE TABELAS

Tabela 1: Exemplo de <i>Dataset</i> Lista de Comprar.....	23
Tabela 2: Conjunto de Atributos para Classificação Associativa Provindos do SAGE do Período de 2001 a 2015	31
Tabela 3: Conjunto de atributos para Classificação Associativa Provindos do IBGE (Censo Demográfico 2010).....	32
Tabela 4: Planilha com a Média dos Indicadores de Tuberculose dos Municípios da Paraíba de 2001 a 2015 e Indicadores Socioeconômicos de 2010	35
Tabela 5: Tabela de frequências de pH	37
Tabela 6: Abreviatura dos Nomes dos <i>Itemset</i>	37

LISTA DE QUADROS

Quadro 1: Resultado das 23 Melhores Regras Geradas	45
---	----

LISTA DE ABREVIATURAS E SIGLAS

BD	<i>Big Data</i>
BDA	<i>Big Data Analytics</i>
BHD	<i>Big Health Data</i>
CSV	<i>Comma-separated values</i>
DATASUS	Departamento de Informática do SUS
GAL	Gerenciador de Ambiente Laboratorial
IBGE	Instituto de Geografia e Estatística
KDD	Knowledge Discovery in Databases
MS	Ministério da Saúde
MSF	Médicos Sem Fronteiras
OMS	Organização Mundial da Saúde
SAGE	Sala de Apoio à Gestão a Saúde do Ministério da Saúde
SAI	Sistema de Informações Ambulatoriais
SIAB	Sistema de Informação da Atenção Básica
SIH	Sistema de Internações Hospitalares
SIM	Sistema de Informação sobre Mortalidade
SINAM	Sistema Nacional de Atendimento Médico
SUS	Sistema Único de Saúde
TB	Tuberculose
TBM	Tuberculose Muti-resistente
WHO	World Health Organization

SUMÁRIO

1. INTRODUÇÃO	15
2. REVISÃO DE LITERATURA	17
2.1 Definições	17
2.1.1 Tuberculose	17
2.1.1.1 Formas de Registro de Dados	18
2.1.1.2 Repositório de Dados e Estatísticas	19
2.1.2 Big Data Analytics	19
2.1.2.1 O Big Data Analytics e a Gestão de Informações	20
2.1.2.1.1 Mineração de Dados (Knowledge Discovery in Databases - KDD)...	21
2.1.2.2 Regras de Associação e Algoritmos	22
2.1.2.3 Linguagens de Programação e Bibliotecas.....	25
2.2 Aplicado BDA na Área da Saúde	25
2.2.1 O Cientista de Dados na Área da Saúde	26
2.2.2 Perfil Epidemiológico da TB e Big Data Analytics	28
3. METODOLOGIA	29
3.1 Tipo de Estudo	29
3.2 Local do Estudo	29
3.3 População do Estudo	30
3.4 Coleta e Análise de Dados.....	30
3.4.1 Mineração de Dados do <i>Dataset</i> Tuberculose na Paraíba	32
3.4.1.1 Dados.....	33
3.4.1.2 Pré-processamento e Transformação.....	33
3.4.1.2.1 Retirando Municípios com Incidência Menor que Dez Casos	36
3.4.1.2.2 Aplicando Distribuição de Frequência nos Valores dos Itemset.....	36
3.4.1.3 Mineração de Dados	42
4. RESULTADOS	43
4.1 Resumo dos Resultados	44
4.2 Análise Comparativa entre Medidas de Interesse.....	46
5. CONCLUSÃO	48
REFERÊNCIAS	49

1. INTRODUÇÃO

Existe uma grande quantidade de dados sobre a Tuberculose (TB) no mundo e no Brasil em diversos repositórios de organizações de saúde que demonstram o impacto socioeconômico e cultural causado por esta epidemia, porém, poucos tentam traçar especificamente o perfil epidemiológico da TB nos municípios da Paraíba utilizando *Big Data Analytics*.

Os bancos de dados disponíveis sobre a TB e a população afetada por ela são inúmeros, fazendo-se saber, existem os da iniciativa *The Paradigm Shift*, os da Organização Mundial da Saúde (OMS), o da Sala de Apoio à Gestão à Saúde do Ministério da Saúde (SAGE), os do Sistema Nacional de Atendimento Médico (SINAM) e os do Departamento de Informática do SUS (DATASUS), ligados também ao Ministério da Saúde (MS), que possibilitam, através de seus dados, analisarem e verificarem o quão grave é esta epidemia.

Segundo a WHO (2017), Brasil, Rússia, Índia, China e África do Sul possuem mais da metade da carga global dos casos de TB e dois terços da carga global de Tuberculose Multirresistente (TBM).

Segundo Araújo (2012, p.15), a tuberculose matou em 2010 cerca 1,1 milhão de pessoas no mundo. Se levarmos em conta, só a capital da Paraíba, João Pessoa, Coutinho *et al.* (2012, p. 37) aponta que nos anos de 2007 a 2010, obteve-se uma média de 457 casos por ano, o que deu uma taxa média de incidência de 65,2/100 mil habitantes.

Desse modo, o conhecimento acerca da situação epidemiológica da TB no estado da Paraíba pode contribuir com ações de planejamento político, técnico e científico no seu enfrentamento.

O *Big Data Analytics* (BDA) e as técnicas envolvidas neste, permitem gerar perfis epidemiológicos com os dados de cada município do Estado, através de regras de associação, que revelem a correlação entre os conjuntos de itens (atributos) que compõem os dados sobre a TB nos municípios da Paraíba para confrontá-las com o comportamento característico da TB contido na bibliografia consultada, almejando obter-se um melhor entendimento dos casos na Paraíba e conseqüentemente maneiras de utilizar o BDA para ajudar aperfeiçoar as políticas públicas voltadas para o controle e

eliminação da TB, além, de servir como referência para a análise, entendimento e aplicação do DBA em outros estados ou em outros dados epidemiológicos.

2. REVISÃO DE LITERATURA

Ao longo da história a tuberculose (TB) se configura de várias maneiras. Silva (2014, p.11) explica que em países da Europa Ocidental e da América do Norte ela caracteriza-se como uma doença reemergente, enquanto no Brasil é uma doença marginalizada, que causa vulnerabilidade social, principalmente em áreas de concentração de pobreza, considerada um grave problema de saúde pública (SILVA, 2014, p.11). A TB, não pode ser ignorada e seus dados podem ser melhor entendidos aplicando as técnicas de *Big Data Analytics*.

O BDA pode abrir novas perspectivas sobre o estudo da TB no mundo, no Brasil e em seus Estados, sendo, o foco desse trabalho os municípios da Paraíba, a fim de analisar o comportamento dos dados epidemiológicos da TB que possuímos nos repositórios mantidos pelo MS e dados socioeconômicos do IBGE em seus diversos sistemas computacionais, que podem colaborar na compreensão de como esta doença manifesta-se e quais as implicações na população paraibana.

2.1. Definições

2.1.1. Tuberculose

A TB é acometida por uma bactéria chamada *Mycobacterium tuberculosis* “é transmitido de forma indireta através de gotículas contendo os bacilos, estes são liberados no ar por pessoas infectadas, através da tosse, fala ou espirro” (FIGUEIREDO *et al.*, 2013, p.4). Apesar de tratar-se de uma patologia de cura relativamente simples e tratamento gratuito, ofertado pelos serviços de saúde, a TB não está sobre controle e instituições como a OMS, Médicos Sem Fronteiras (MSF) e MS estipulam metas para os próximos anos a fim de controlar e eliminar a TB (SILVA, 2014, p.11).

Segundo Ribeiro *et al.* (2016) a TB “possui estreita relação com *status* socioeconômico, pobreza e desemprego”. Sá LD *et al.* (2008, p. 3918) completa alertando que a TB “permanece como um dos problemas mundiais da saúde pública, matando pelo menos 6 mil pessoas/ano no Brasil”, assim, Daronco *et al.* (2012) conclui

que a TB é uma doença que perdura há séculos, pois, não há um controle efetivo em muitos países, como é o caso do Brasil.

O MS (2017, p.6) afirma que “acontecem aproximadamente 69 mil novos casos e 4.500 óbitos a cada ano como causa básica a tuberculose”. A TB não pode ser negligenciada e merece atenção, o próprio MS aponta para gravidade da TB, colocando-a na Lista Nacional de Agravos de Notificação Compulsória (BRASIL, 2017, p.6).

A TB produz elevada morbimortalidade, tendo o Brasil forte contribuição para os indicadores mundiais, o que amplia a importância de conhecer o motivo e fatores que são determinantes para o aumento dos casos de TB. Conhecer de forma mais profunda as causas e correlações epidemiológicas e socioeconômicas que levam aos agravos causados pela TB é um elemento fundamental para o desenvolvimento de políticas públicas adequadas na área da saúde e investimentos significativos nos sistemas de saúde, evitando mais mortes. (Paradigm Shift, 2015).

2.1.1.1. Formas de Registro de Dados

Os dados de diagnósticos e acompanhamento dos casos de TB no Brasil são registrados nas unidades de saúde utilizando várias formas de coleta.

O MS recomenda o uso de instrumentos para registro, notificação e acompanhamento dos casos no Manual de Recomendações para o Controle da Tuberculose no Brasil, que informa como principais instrumentos de coleta de dados, os Livros de Registro de Sintomáticos Respiratórios, de exames laboratoriais para o diagnóstico da TB e de casos diagnosticados e tratados na Unidade de Saúde, também, os formulários padronizados que possuam informações de internações hospitalares por TB (guia de internação) e declaração de óbito. Contudo, o principal formulário de coletas é a ficha de notificação e investigação de tuberculose do Sistema de Informação de Agravos de Notificação (SINAN). (MS, 2011, p.192).

2.1.1.2. Repositório de Dados e Estatísticas

Os dados sobre a tuberculose de forma digital estão disponíveis em vários sistemas, como: o de internações do SIH/SUS; o do atendimento ambulatorial do SIA/SUS, o de ações da atenção básica (SIAB), no de mortalidade (SIM) e no de notificações (SINAN), essas informações geram um grande volume de dados, que podem ser analisados e entendidos (BRASIL, 2011, p.192).

Segundo Araújo (2012, p.16), levando em conta apenas os casos de TB da região nordeste, “o coeficiente de incidência em 2010 foi de 36,9/100.000 habitantes, o estado da Paraíba foi responsável pela notificação de 1.060 novos casos de TB, apresentando um coeficiente de incidência de 28,1/100.000” habitantes, comparado com os dados mais recentes do portal da SAGE (2017), a região nordeste, em 2015, obteve um coeficiente de 30,95/100.000 habitantes e o estado da Paraíba para este mesmo período 25,30/100.000 habitantes. Apesar da Paraíba apresentar um coeficiente um pouco menor que a média apresentada na região, ainda, mostra um coeficiente de incidência bastante elevado e relevante, que corroboram para o aumento dos índices no país e no mundo.

Estes valores são consideravelmente volumosos e preocupantes para Paraíba, se levar em consideração apenas o ano de 2010, este se encontrava na 14^a posição entre os estados com maior número de casos. (COUTINHO, 2012, p. 36 *apud* BRASIL, 2011).

2.1.2. *Big Data Analytics*

O *Big Data Analytics* (BDA) permite analisar o volume de dados que cresce de maneira astronômica a cada segundo, dos vários formatos de dados disponíveis, como textos, sons, imagens e vídeos (SOUZA, *et al.*, 2015). Um estudo intitulado “*A Universe of Opportunities and Challenges*”, desenvolvido pela consultoria EMC, escrito por Gantz (2012), “aponta que de 2006 a 2010, o volume de dados digitais gerados cresceu de 166 *Exabytes* para 988 *Exabytes*, com perspectiva que o crescimento chegue a 40 *Zettabytes* (ou 40 trilhões de *Gigabytes*)”, Ávila (2017) chama a esse

volume de dados de *Big Data* (BD) e possuem como características sua variedade, sua velocidade e o grande volume que possuem.

2.1.2.1. O *Big Data Analytics* e a Gestão de Informações

Para ultrapassar as limitações de análise de dados tradicionais usa-se o BDA, Santos e Ferreira (DEV MEDIA, 2017) reforçam o *Big Data* como uma “nova” maneira de armazenar, gerenciar e analisar grandes volumes de dados de diversas fontes, a uma velocidade considerável, utilizando sistemas computacionais; o portal Hoppen (2015) conclui que através de técnicas de mineração de dados chega-se a descoberta de comportamentos, tendências ou realização de previsões; já Taurion (2016) considera o termo *Big Data* como definido de forma inadequada, pois “passa a impressão de um imenso e estático volume de dados, quando na verdade o valor real dos dados está em seu tratamento e análise, principalmente através de algoritmos descritivos e/ou preditivos”. O BDA está mais ligado à gestão de informações, assim, “tratam-se de diversas estratégias de tecnologia para deixar a percepção mais rica, profunda e precisa no que se refere a clientes, analisando padrões e correlações, ganhando por fim vantagem [...]” (VIANNA; DUTRA e FRAZZON, 2016, p.196).

O BDA pode ser melhor entendido como uma ferramenta de apoio estratégico para fortalecer a tomada de decisão (ORACLE, 2016), objetivando aperfeiçoar processos de trabalho e adquirir informações valiosos acerca das tendências de mercado (BIGDATABUSINESS, 2016), para propiciar a compreensão de comportamentos e expectativas de um determinado público, o BDA utiliza todas as técnicas envolvidas na mineração de dados, que Fayyad (1996) define como a transformação de dados sem valor aparente em informação (conhecimento). De qualquer forma, a mineração de dados geralmente envolve as etapas de:

- a. Seleção de dados;
- b. Pré-processamento de dados;
- c. Transformação de dados;
- d. Mineração de dados;

- e. Avaliação, e
- f. Conhecimento.

Através das etapas citadas, com a ajuda de um especialista para correta avaliação das regras geradas, chega-se ao conhecimento que estava implícito nos dados. (FAYYAD, 1996).

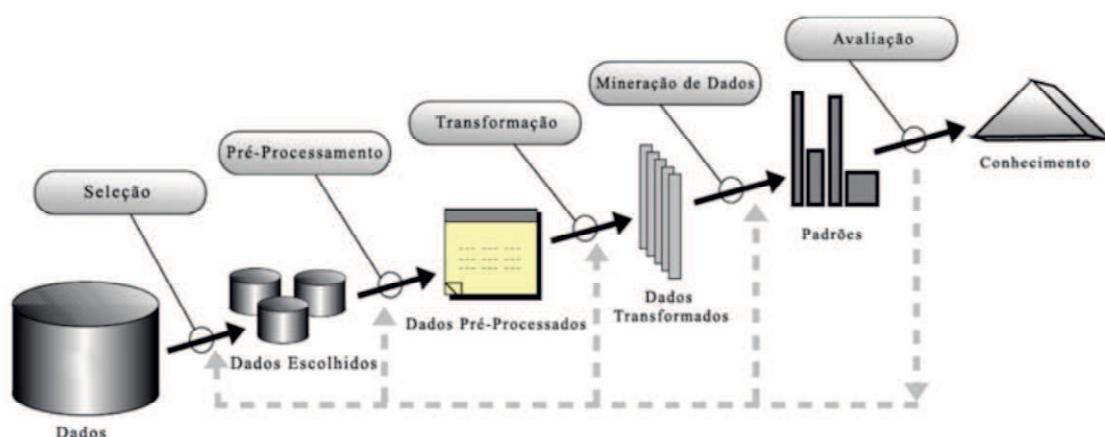
2.1.2.1.1. Mineração de Dados (*Knowledge Discovery in Databases - KDD*)

A mineração de dados como define Fayyad (1996) pode ser entendido com um processo do *KDD* (*Knowledge Discovery in Databases* ou Descoberta de Conhecimento nas Bases de Dados) que surge para sanar os problemas gerados pela chamada, era da informação, que impossibilita analisar manualmente todo o volume de dados gerados cotidianamente.

A mineração de dados e a descoberta de conhecimentos em bases de dados podem ser considerados a mesma coisa como afirmam Rezende (2005) e respectivamente Han e Kamber (2006). Também podem ser entendido como parte do processo, como afirma Cios (*et al.*, 2007).

Todavia, geralmente apresentam as etapas da imagem abaixo:

Figura 1: Figura Representando o Processo de KDD



Fonte: CAMILO; SILVA, 2009, p.3 *apud* FAYYAD, 1996.

2.1.2.2. Regras de Associação e Algoritmos

As regras de associação, segundo Ferraz (2008, p.15), é um dos tipos de regras possíveis, e são do tipo “se/então”, ou seja, se X ocorre, então, Y tende a ocorrer, com X e Y sendo conjuntos disjuntos de itens de dados.

Segundo, Agrawal; Imielinski e Swami (1993), as regras de associação podem ser expressas também como uma regra $X \Rightarrow Y$, na qual o conjunto X é o antecedente (*Left Hand Side*) e o conjunto Y é o conseqüente (*Right Hand Side*).

Para Baranauskas (2018), se todos os itens de X forem encontrados em uma transação, existe a possibilidade de também encontrar os itens de Y. Neste sentido, Verde (2016), aponta a frequência da ocorrência desses itens como padrões, por exemplo, padrões de compra.

Os algoritmos entram na fase de mineração de dados, de forma concisa, pode-se dizer que “data mining é um processo iterativo”. (IBM, 2018). Suas etapas agem umas sobre as outras, com a participação de especialistas para analisar as regras geradas, de forma menos estrita, a mineração de dados pode ser vista como:

[...] um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados. (CABENA *et al.*, 1998).

Para a extração de conhecimento em bases de dados, existem várias formas de mineração de dados, algumas delas são, usando regras de associação, redes neurais, algoritmos genéticos e lógica nebulosa, análise de agrupamentos, árvores de decisão, dentre outras. (CAMILO, 2009, p.19).

Para este trabalho foi evidenciado o uso de algoritmos associativos. Um dos mais famosos é o apriori, que utiliza a estratégia de itens frequentes. Esse algoritmo foi introduzido por (AGRAWAL; SRIKANT, 1994) e apontado pela IEEE na *International Conference on Data Mining (ICDM)*, segundo, Silva *et al.* (2013 *apud* WU *et al.*, 2007) “como o mais promissor algoritmo de geração de regras de associação e umas das abordagens mais populares em mineração de dados, é um algoritmo seminal para encontrar conjuntos de itens frequentes”. O apriori, voltado para área de saúde foi utilizado por Librelotto e Mozzaquatro (2013), que o evidencia como o melhor para detecção de indicadores da qualidade de vida e saúde. De outra forma, foi usado para

mineração de dados, afim, de aprimorar o atendimento médico em um cenário de plano de saúde. (PACHIAROTTI, 2012).

O algoritmo apriori, funciona fazendo uma busca em profundidade nos dados, gerando conjuntos de itens candidatos (reconhecidos como o padrão), a varredura termina com o último elemento da base de dados, os padrões menos frequentes não são exibidos. (SILVA *et al.*, 2013, p.5).

As regras geradas utilizam algumas métricas de interesse e relevância, dentre elas suporte e confiança determinadas no trabalho de Agrawal; Imielinski e Swami (1993). Outras métricas foram pesquisadas por Silva *et al.* (2013 *apud* Geng e Hamilton 2006) a fim de encontrar estratégias para domínios e exigências específicos. Neste trabalho as métricas utilizadas foram:

- **Suporte (*support*):** $P(AB)$. O suporte de uma regra é definido como sendo a fração de itens I que satisfazem o conjunto A e B da regra. Se o suporte não é grande o suficiente, isso significa que a regra não deve ser levada em consideração ou deve ser analisado em outro momento;
- **Confiança (*confidence*):** $P(A/B)$. É uma métrica da força do suporte às regras e corresponde a significância estatística. A probabilidade de encontrar B da regra nas transações sobre a condição que essas transações também contenham A ;
- **Interesse (*lift*):** $P(B|A) / P(B)$ ou $P(AB) / P(A)P(B)$. Serve para achar dependências, ela mostra o quanto mais frequente torna-se B quando A ocorre. Varia entre 0 e ∞ .

Na **Tabela 1** é possível ver um exemplo simples, de um *dataset* (coleção de dados tabulado) que representa uma lista de compras e em seguida, um exemplo com os cálculos da confiança e do suporte para uma regra do *dataset*.

Tabela 1: Exemplo de *Dataset* Lista de Comprar

TID	Lista de Itens
T1	Pão, Leite
T2	Pão, Fralda, Cerveja, Ovos
T3	Leite, Fralda, Cerveja, Coca
T4	Pão, Leite, Fralda, Cerveja
T5	Pão, Leite, Fralda, Coca

Fonte: Regras de Associação (OLIVEIRA)

Exemplo:

Regra: {Leite, Fralda} \Rightarrow {Cerveja}

$$Suporte = \frac{Frequencia(Leite, Fralda, Cerveja)}{|T|} = \frac{2}{5}$$

$$Confianca = \frac{Frequencia(Leite, Fralda, Cerveja)}{Frequencia(Leite, Fralda)} = \frac{2}{4}$$

Obter conjuntos de itens com frequência maior ou igual à especificada pelo usuário como sendo o suporte mínimo, não é trivial, devido à grande possibilidade de análises combinatórias ocorridas ao gerar os subconjuntos de itens. Quando os itens frequentes são obtidos, é fácil gerar regras de associação com confiança maior ou igual à especificada pelo usuário como valor mínimo (WU *et al.*, 2007). Estas regras são utilizadas para diminuir o número de causas a serem comparadas com cada transação do banco de dados, sendo que, todas as causas geradas que contenham algum subconjunto que não seja frequente são eliminadas (LANGE, 2007).

Esse algoritmo possui algumas variações elaboradas por desenvolvedores com a finalidade de aperfeiçoar os algoritmos de associação, com características específicas e respectivas vantagens em relação ao apriori, são algoritmos desenvolvidos por:

- Casanova (2005): usa o Algoritmo da Confiança Inversa junto com a Lógica Nebulosa para gerar regras mais precisas;
- Zaki e Gouda: usa o ECLAT (*Equivalence CLASS Transformation*) com *FP-growth* com a promessa de melhores resultados;
- Palancar (2008): usa o *CBMine* (*Compressed Binary Mine*) que, segundo os testes, apresenta melhores resultados que os algoritmos tradicionais.

2.1.2.3. Linguagens de Programação e Bibliotecas

No universo do BDA existem várias linguagens de programação e ferramentas para desenvolvimento das principais técnicas de mineração de dados e análise de dados. A busca por um consenso de qual destas deve ser utilizada é complicada, as principais linguagens/ferramentas são: *Python*, *Stata*, *SPSS*, *R*, *JMP*, *MATLAB* e *Julia* (CIO, 2017). Essa gama de opções acaba por dificultar o compartilhamento de resultados e o desenvolvimento de novas análises. Todavia, duas linguagens de programação obtiveram um grande apoio dos cientistas de dados na última década, a *R* e a *Python* (CHIAVEGATTO FILHO, 2015, p. 328). A revista CIO (2017) afirma que as vantagens dessas linguagens são o grande apoio da comunidade e o fato de ser *Open Source*, que são *softwares* que podem “ser usado, copiado, estudado, modificado e redistribuído sem restrição” (Free Software Foundation, 2017), apesar de não serem necessários grandes investimentos em termos de softwares (PEREIRA, 2016, p. 28), para programar nessas linguagens é preciso uma equipe bem capacitada e especializada, demandando tempo e custos elevados. (DAVENPORT, 2014, p.115).

2.2. Aplicando BDA na Área da Saúde

O BDA “tem se tornado cada vez mais importante tanto no meio acadêmico como em empresas ao longo das últimas duas décadas” (MELLO, 2015, p. 13 *apud* CHEN; CHIANG; STOREY, 2012), pois, com essas técnicas podem-se obter resultados invisíveis a análises relacionais comuns. Chiavegatto Filho (2015, p. 328) anuncia que

“é possível encontrar fortes indícios de que a próxima grande fronteira da epidemiologia seja a análise de grandes bancos de dados (*Big Data*)” e Hernández-Leal, Duque-Méndez e Moreno-Cadavid (2017) reforçam que com o BD (tradução nossa) surge um novo tipo de profissional, o cientista de dados, que são as pessoas capacitadas para analisar e interpretar os dados computacionais.

A mineração de dados, portanto representa para as corporações a possibilidade de tomar decisões mais precisas e, sobretudo, antecipadas. (BIGDATABUSINESS, 2016). Em um cenário que envolve a área saúde, (tradução minha) “grandes análises de dados têm o potencial de melhorar o atendimento, salvar vidas e reduzir os custos.”(RAGHUPATHI; RAGHUPATHI, 2014, p.1).

Na área de saúde, o uso do BD, também pode ser convencionado como *Big Health Data* (BHD). O BHD já obteve sucessos expressivos que obtém vários sucessos, como o sistema *HealthMap* que foi criado por pesquisadores, epidemiologista e desenvolvedores de *software*, no Boston *Children's Hospital* em 2006 (Healthmap, 2017). Souza (*et al.*, 2015) elucida que o sistema *HealthMap*:

[...] foi capaz de prever o surto de Ebola na África Ocidental nove dias antes de ser anunciado formalmente pela Organização Mundial da Saúde (OMS). Este sistema identificou a existência de uma “febre hemorrágica desconhecida” no sudeste da Guiné por meio de cruzamento de dados de milhares de sites diferentes, como sites governamentais, redes sociais, sites especializados em doenças infecciosas e outras fontes. (SOUZA *et al.*, 2015).

As análises envolvendo a comunicação entre vários centros de pesquisa aumentam a cada dia, juntamente com a necessidade por decisões mais acertadas, para tanto, são necessárias novos tipos de profissionais a fim de analisar dados complexos e desestruturados para extrair informações úteis (FELDMAN; SANGER, 2007).

2.2.1. O Cientista de Dados na Área da Saúde

No Brasil, algumas das oportunidades de *Big Health Data* (BHD), que pode ser entendido como o Big Data usado na área da saúde, que neste trabalho é chamado simplesmente de *Big Data Analytics* (BDA), inclui a análise dos bancos de dados

mantidos pelos sistemas do Ministério da Saúde, como a Sala de Apoio à Gestão Estratégica (SAGE), o Sistema de Informação de Agravos de Notificação (Sinan) e a esta lista, Souza, *et al.* (2015) ainda incluem o “Sistema de Informações sobre Nascidos Vivos (Sinasc), o Sistema de Informações sobre Mortalidade (SIM), o cartão SUS, entre outros”.

Apesar do BDA ser uma grande ferramenta na área da saúde, esta ainda é uma área emergente e promissora, que tem como foco de crescimento “a medicina de precisão, os prontuários eletrônicos do paciente e a internet das coisas”. (CHIAVEGATTO FILHO, 2015). Porém, para que essa área de atuação cresça, são necessários esforços conjuntos, nacionais e internacionais, para definir linguagens de programação adequadas e nomenclaturas padronizadas para que os dados entre os sistemas se interconectem sem dificuldade.

Os dados expostos até agora mostram que utilizar as técnicas de DBA para traçar perfis epidemiológicos para saúde é necessário (CAMARGO; PINHEIRO; COELI, 2015). No Brasil, “o volume de dados deve ir de 212 bilhões de gigabytes em 2013 para 1.600 bilhões de gigabytes em 2020” (EXAME, 2017), e dados sobre a saúde fazem parte desse crescimento.

O DATASUS evidencia a importância da análise desses dados, possuindo seus próprios projetos de *Business Intelligence* (BI), que podem ser vistos como “as tecnologias que são utilizadas para coletar, acessar e analisar dados e informações de apoio à tomada de decisão.” (BALTZAN; PHILLIPS, 2012, p. 234). O que permite às instituições de saúde, de qualquer porte buscar e interpretar informações armazenadas para apoio às decisões dentro do ciclo de vida do cidadão (DATASUS, 2017).

As epidemiologias são um dos focos nesses projetos, como exemplo, o projeto *Business Intelligence GAL* (Gerenciador de Ambiente Laboratorial) que disponibiliza informações epidemiológicas obtidas no sistema GAL, este “objetiva-se a proporcionar o gerenciamento das rotinas, o acompanhamento das etapas para realização dos exames/ensaios e a obtenção de relatórios produção/ epidemiológicos/ analíticos nas redes estaduais de laboratórios de saúde pública” (DATASUS, 2018), sob a forma de painéis de indicadores (*dashboards*) e relatórios, ainda encontra-se desenvolvimento. (Ministério da Saúde, 2017).

2.2.2. Perfil Epidemiológico da TB e *Big Data Analytics*

Na Paraíba se faz importante o reforço de estudos epidemiológicos sobre a TB, tendo em vista possuírem um grande número de casos anuais e dados a seu respeito. O BDA então é uma forte ferramenta para verificar até onde os casos de TB e seus agravos se enquadram no perfil epidemiológico típico descrito na bibliografia sobre a TB. Os perfis gerados podem melhorar a forma como é vista esta epidemia nos municípios da Paraíba, a fim de reforçar os esforços mundiais na tomada de decisões referentes às políticas públicas para obtenção do controle e eliminação da TB, servindo também para evidenciar o quanto o BDA pode ser eficiente para analisar os casos de TB nos municípios paraibanos.

3. METODOLOGIA

3.1. Tipo de Estudo

Trata-se de um estudo descritivo, com abordagem exploratória, do tipo série temporal, envolvendo a coleta de dados secundários de domínio público, no período de 2001 a 2015, referente às ocorrências relacionada à tuberculose nos municípios da Paraíba, bem como dos dados sobre as condições socioeconômicas destes municípios.

3.2. Local do Estudo

O estudo foi realizado no estado da Paraíba (PB), localizado no litoral oriental da Região Nordeste do território Brasileiro, o qual possui uma área territorial de 56.469 Km², dividida em 223 municípios (IBGE, 2010). “Oito municípios têm suas populações entre 30.000 e 60.000 habitantes (3,58%); três municípios apresentam população que varia de 60.000 até 100.000 habitantes (1,3%); e cinco municípios têm população superior a 100.000 habitantes (2,24%). Como explica Souza (2014, p. 36 *apud* IBGE, 2010), “isso significa uma concentração percentual de 92,8% de municípios com população até 29.000 habitantes” e o COSEMSPB (2017) a divide em 16 Regiões de Saúde, assim divididos para fins administrativos e definidos no último Plano Diretor de Regionalização (PDR) que foi atualizado em 2008. Dessa forma, a Paraíba foi dividida em regiões/macro/micro/módulos assistenciais, com populações pertencentes definidas, subsídios técnico-operacionais para elaboração da Programação Pactuada e Integrada (PPI) e as redes de referências articuladas e resolutivas, dentre outros (COMISSÃO INTERGESTORES BIPARTITE, 2008).

3.3. População do Estudo

A população do estudo constitui todos os municípios com casos notificados de tuberculose, diagnosticados no período de 2001 a 2015, segundo município de residência. Assim, serão incluídos todos os municípios onde ocorreram casos novos diagnosticados no período proposto, de ambos os sexos, independente do local de diagnóstico; e serão excluídos os municípios que não apresentaram casos de tuberculose por mais de 10 anos, consecutivos ou não.

3.4. Coleta e Análise de Dados

A coleta foi realizada a partir de dados secundários de domínio público, disponíveis na Sala de Apoio à Gestão Estratégica (SAGE) do Ministério da Saúde (MS) e no portal do Instituto Brasileiro de Geografia Estatística (IBGE).

Para criar o *dataset* com os dados provenientes dos casos de TB dos municípios da Paraíba, onde foram aplicadas as técnicas de *Big Data Analytics* a fim de minerar os dados para extrair conhecimento sobre os mesmos. Foram utilizados os índices disponíveis publicamente no portal da Sala de Apoio à Gestão Estratégica (SAGE) do Ministério da Saúde (MS) e os índices do Instituto Brasileiro de Geografia Estatística (IBGE).

A SAGE foi considerada a melhor fonte nacional pública para os indicadores epidemiológicos e operacionais sobre a tuberculose e acesso a serviços de saúde, e o IBGE como melhor possuidor de indicadores socioeconômicos.

O estudo exploratório seguiu 3 etapas:

Etapa 1 – seleção de atributos. Foram escolhidos os que possuem as características e representem agravos da tuberculose, cuja listagem foi obtida a partir de revisão bibliográfica, que aponta como principais vítimas da TB as populações com menor poder aquisitivo, com falta de saneamento e rede esgoto adequados, onde existem aglomerados de pessoas, com pouca formação e portadores de HIV. Os grupos

de atributos foram submetidos à validação da especialista da área da saúde e epidemiologia, que considerou relevante a busca por estes tipos de atributos;

Etapa 2 – estruturação da base de dados. A partir dos atributos escolhidos, foi elaborada uma série de planilhas eletrônicas com os dados agrupados e tabulados no formato *Comma-separated values* (CSV), com o auxílio do software de Planilhas Eletrônicas: *-WPS Office Spreadsheets*. Os atributos escolhidos para base de dados inicial encontram-se na **Tabela 2** e **Tabela 3**; como os dados foram provenientes de duas bases de dados distintas, precisaram ser normalizados em uma base de dados única; os dados da Tabela 2 compõem as médias dos dados de cada atributo em relação aos anos que possuem dados diferentes de “0” e não “vazios”, para que a média não divergisse muito em relação à realidade da amostra, isso, porque não é possível saber se nos anos que estão como dados “0” ou “vazios” são por falta de ocorrências ou falta de envio dos dados para consolidação no SAGE. Foi necessário fazer à média, porque o período analisado é de 2001 a 2015, para, poder alinhando-se com os dados da Tabela 3 que são referentes apenas ao ano de 2010, o último Censo do IBGE;

Etapa 3 – aplicação do processo de mineração de dados. Esta etapa será detalhada de forma mais abrangente no item 3.4.1 a seguir e, de maneira sucinta, foi desenvolvida seguindo os passos básicos para mineração de dados, o pré-processamento por meio das tarefas da seleção e transformação dos dados.

Tabela 2: Conjunto de Atributos para Classificação Associativa Provindos do SAGE do Período de 2001 a 2015

Atributos
Taxa de Incidência de Tuberculose por 100.000 hab/ano
1. Tx. Incidência ;
2. Tx. Mortalidade;
3. Tx. Incidência Bacilífera ;
Percentual de casos segundo raça/cor
4. Casos Brancos;
5. Casos Amarelos;
6. Casos Indígenas;

-
7. Casos Pardos;
 8. Casos Pretos;
 9. Casos Ignorados;

Percentual de óbitos segundo raça/cor

10. Óbitos Brancos;
11. Óbitos Amarelos;
12. Óbitos Indígenas;
13. Óbitos Pardos;
14. Óbitos Pretos;
15. Óbitos Ignorados;

Indicadores operacionais

16. Casos Bacilíferos Curados;
 17. Casos de Retratamento com Cultura;
 18. Casos com teste HIV realizado;
 19. Acesso a Saúde Básica;
-

Tabela 3: Conjunto de atributos para Classificação Associativa Provindos do IBGE (Censo DemoGráfico 2010)

Atributos

Dados socioeconômicos

20. Renda Percapta 1/2 Salário;
 21. Taxa de escolarização de 6 a 14 anos de idade;
 22. Índice de Desenvolvimento Humano Municipal (IDH-M);
 23. Esgotamento Sanitário Adequado;
 24. Média de Moradores em Domicílios Particulares Ocupados;
 25. Índice GINI (mede o grau de concentração de renda).
-

3.4.1. Mineração de Dados do *Dataset* Tuberculose na Paraíba

Para minerar os dados foram utilizadas a suíte Anaconda *Python* 3 (Anaconda 3), que possui a IDE *Spyder* 3, um ambiente de desenvolvimento *Open Source* para

programação científica na linguagem *Python*, que integra as bibliotecas *NumPy*, *Pandas*, *SciPy*, *Matplotlib* e *IPython*, o *MySQL Community Server*, que é um banco de dados gratuito e com código aberto, também o *MySQL Workbench*, que oferece aos administradores de bancos de dados e desenvolvedores um ambiente que integra várias ferramentas (design e modelagem de banco de dados, desenvolvimento SQL, administração de banco de dados e migração de banco de dados) e, por fim, uma codificação em *Python 3* do algoritmo *apriori*, chamado *Apyori*, a fim de resolver o problema de associação preditiva, no *dataset* dos casos de tuberculose da Paraíba, para reconhecer os padrões comportamentais associativos e obter perfis epidemiológicos com probabilidades diversas, dentro do conjunto de dados de aprendizado utilizado.

3.4.1.1. Dados

Na **etapa 2**, foi construída uma Planilha com os dados do SAGE e do IBGE referentes aos 223 municípios da Paraíba; este, com 27 colunas que representam seus atributos, neste trabalho, chamados de *Itemset*. Todavia, as colunas Período e Municípios não são relevantes para mineração de dados e posteriormente foi ignorados, assim, como a linha com os cabeçalhos dos *Itemset*. A planilha gerada possui a estrutura da Tabela 4. É notório que existem muitas células sem dados, que os valores são contínuos e que precisaram passar na fase de pré-processamento, levando em conta, que a Tabela 4 é *figurativa*, apresentando apenas dados de 18 municípios.

Esta planilha foi importada para o MySQL, a base de dados recebeu o nome “tuberculose” e o dados foram armazenados em uma tabela com o nome de “tb_paraiba”.

3.4.1.2. Pré-processamento e Transformação

Devido às origens diversas dos dados do SAGE, que tem seu envio dependente da administração municipal, para que sejam consolidados, é corriqueiro que os dados estejam despreparados para aplicação direta das técnicas de mineração de dados.

Foram utilizadas técnicas para preencher os espaços vazios, afim, de transformar os valores contínuos em valores discretos, também, de antemão, foram excluídos os municípios que não atendem ao critério apresentado no título 3.3 deste trabalho, ou seja, de apresentar pelo menos nove casos de tuberculose entre os anos de 2001 a 2015.

Os arquivos CSV dos *dataset* gerados para a mineração podem ser acessados no seguinte endereço, <<https://github.com/tuberculoseparaiba/datasets>>.

Tabela 4: Planilha com a Média dos Indicadores de Tuberculose dos Municípios da Paraíba de 2001 a 2015 e Indicadores Socioeconômicos de 2010

Período	Município	Tx. Incidência	Tx. Mortalidade	Tx. Incidência Bacilifera	Casos Baciliferos Curados	Casos de Retratamento com Cultura	Casos com teste HIV realizado	Casos Brancos	Casos Amarelos	Casos Ignorados	Casos Indígena	Casos Parda	Casos Negros	Óbitos Brancos	Óbitos Amarelos	Óbitos Ignorados	Óbitos Indígena	Óbitos Parda	Óbitos Negros	Acesso a Saúde Básica	Renda Percpt a 1/2 Salário	Tx. Escolarização de 6 a 14 Anos de Idade	IDH-M	Esgotamento Sanitário	Moradores Domicílios Particulares Ocupados	GINI	
2001 - 2015	Água Branca	18,881	11,310	15,763	100,000	83,333	75,000					93,750						100,000		99,624	53,100	97,600	0,572	3,690	0,233	0,545	
	Aguiar	32,702	18,775	22,053	87,500		100,000	66,667				85,714						100,000		96,898	50,100	98,900	0,597	17,800	0,033	0,654	
	Ataíde	22,933	5,241	16,370	78,887		50,530	23,053	18,850	30,555		59,783	15,670	75,000	50,000				87,500	75,000	99,778	50,800	97,400	0,582	49,400	0,036	0,551
	Ataíde	15,061	8,276	10,324	86,665	100,000	47,220	71,970			100,000	56,481	33,333	50,000				75,000	75,000	97,858	51,800	97,900	0,576	38,100	0,036	0,537	
	Ataíde	23,501	10,593	14,956	86,363		75,499	33,333			60,000	77,167	37,500	50,000				83,333		96,058	49,900	98,700	0,595	9,100	0,037	0,531	
	Alcantil	28,598	19,580	22,948	100,000	100,000	83,333	66,667				85,714						100,000		99,784	49,500	95,700	0,578	4,700	0,033	0,481	
	Ataíde	62,738	42,410	59,710	100,000		72,220	100,000				92,857	50,000	100,000						99,740	52,600	98,500	0,548	45,700	0,035	0,473	
	Alhandra	38,577	6,393	21,798	70,501	100,000	60,872	31,329	19,168	11,110		71,680	23,546					100,000		97,778	49,300	96,300	0,582	9,300	0,036	0,465	
	Tacima	22,019	11,280	23,122	85,415	50,000	72,220	52,777			100,000	90,278						100,000		99,526	56,600	97,000	0,551	0,117	3,97	0,493	
	Taperoá	20,275	6,977	18,478	87,575	100,000	52,378	67,619				74,168	34,165	100,000				100,000		96,908	50,100	96,800	0,578	0,553	0,035	0,492	
	Tavares				100,000		83,333	100,000				100,000	100,000								81,654	51,500	98,300	0,586	0,253	3,71	0,532
	Teixeira	13,535	10,663	12,124	87,500		41,667	83,333				92,857						100,000		99,953	51,400	97,300	0,605	0,320	0,035	0,561	
	Tenório	102,223	39,760	116,026	88,635		77,273	72,725				86,363	9,090	100,000						99,659	49,900	95,700	0,581	0,023	3,68	0,429	
	Umbuzeiro	21,784	14,930	21,982	93,333		55,553	62,500	100,000	50,000		75,000	100,000	66,667				66,667		99,646	54,100	99,400	0,584	0,298	0,036	0,549	
	Várzea							100,000					100,000					100,000		99,460	35,700	99,500	0,707	0,687	0,032	0,408	
	Vieirópolis	41,738	50,105	31,098	83,333		100,000	66,000				75,000	20,000	75,000				25,000		97,801	53,100	98,800	0,571	0,019	3,56	0,453	
	Vista Serrana	36,572	31,850	41,410	75,000		100,000	100,000				100,000		100,000						99,060	50,800	92,900	0,566	0,351	3,80	0,476	
	Zabelê	65,620	50,745	47,580			100,000	100,000				100,000		100,000						99,929	48,800	97,300	0,623	0,513	0,031	0,436	

Fontes: SAGE e IBGE

3.4.1.2.1. Retirando Municípios com Incidência Menor que Dez Casos

Os dados existentes no SAGE são por vezes muitos dispersos, pois são enviados pelas prefeituras dos municípios. Os *itemset* de alguns municípios estão sem informações para alguns anos, outros não apresentaram dados. Nesses municípios, não é possível ter certeza se a ausência de dados significa que não houve incidência de tuberculose ou apenas não foram registrados. Assim, para refinar a base de dados que deu origem ao *dataset* inicial, foram eliminadas as cidades, através do critério pré-determinado para a amostra, os municípios que não possuem dados sobre a incidência de TB por no mínimo nove anos, dos quinze anos analisados; isso, visando melhorar o resultado gerado na fase de mineração. Dessa forma, dos 223 municípios, foram analisados 146.

3.4.1.2.2. Aplicando Distribuição de Frequência nos Valores dos *Itemset*

Todos os valores do SAGE e IBGE utilizados no *dataset* inicial são do tipo contínuo, ficando praticamente impossível gerar alguma regra legível, para isso, utilizou-se para todos os valores uma abreviatura que identifique o nome do *itemset* (atributo) a qual o valor pertence, estes, listados na Tabela 6, junto, com intervalos de classes gerados através da técnica estatística de distribuição de frequência, usando o método de *Sturges*, que é considerado o mais consagrado para determinação do número de classes para uma distribuição de frequências e, conseqüentemente, o histograma, fornecendo valores baixos para “n”, que cresce muito devagar. Indicado para valores elevados de “n” para torná-los valores discretos. (FOGO, 2018). Na Tabela 5 vemos um exemplo de distribuição de frequência:

Tabela 5: Tabela de frequências de pH

Classe - pH	n_i	f_i	F_{ac}
4,12 [--- 4,40	9	0,346	0,346
4,40 [--- 4,68	9	0,346	0,692
4,68 [--- 4,96	2	0,077	0,769
4,96 [--- 5,24	1	0,038	0,807
5,24 [--- 5,52	2	0,077	0,884
5,52 [--- 5,80	3	0,115	0,999
Total	26	0,999*	-

* o valor 0.999 ocorreu devido ao arredondamento na precisão considerada (3 casas).

Fonte: Departamento de Estatística – UFSCar (FOGO, 2018)

Tabela 6: Abreviatura dos Nomes dos *Itemset*

<i>Itemset</i>	Abreviatura	Descrição
tx_incidencia_tuberculose	Tx_Inc_TB	Taxa de incidência de tuberculose para cada 100.000 habitantes por ano
tx_incidencia_bacilífera	Tx_Inc_Bacilifera	Taxa incidência de tuberculose Bacilífera para cada 100.000 habitantes por ano
bacilíferos_curados	Bacilifera_Curado	Percentual de Casos Bacilíferos Curados
tx_mortalidade	Tx_Mortalidade_TB	Taxa de mortalidade para cada 100.000 habitantes por ano
retratamento_realizaram_cultura	Retratamento_Cultura	Percentual de casos de retratamento que realizaram cultura
teste_hiv_realizado	Teste_HIV	Percentual de Casos com teste HIV realizado
casos_brancos	Casos_Branco	Percentual de casos de tuberculose divididos por etnia
casos_negros	Casos_Negros	
casos_amarelos	Casos_Amarelos	
casos_parda	Casos_Pardos	
casos_indigena	Casos_Indigena	
casos_ignorado	Casos_Ignorados	
obitos_branco	Obitos_Branco	Percentual de óbitos de tuberculose divididos por etnia
obitos_negros	Obitos_Negros	
obitos_amarelos	Obitos_Amarelos	
obitos_parda	Obitos_Pardos	
obitos_indigena	Obitos_Indigena	

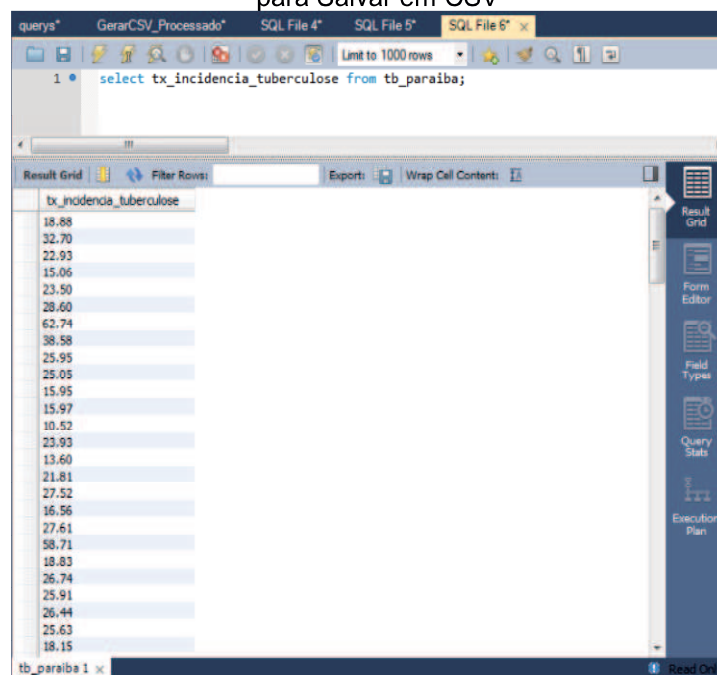
obitos_ignorado	Obitos_Ignorados	
renda_percapta_meio_salario	Percapta_Meio_Salario	Percentual habitantes com renda per capita de meio salário mínimo
esgotamento_sanitario	Esgotamento_Sanitario	Percentual de habitantes com acesso a esgotamento sanitário
media_moradores_domicilios	Moradores_Domicilios	Percentual de moradores por domicílios particulares ocupados
tx_escolarizacao_6_14_anos	Tx_Escolarizacao_6a14	Percentual de habitantes escolarizados entre 6 e 14 anos
acesso_saude_basica	Acesso_Saude_Basica	Percentual de habitantes com acesso à saúde básica
idh_m	IDH_M	Índice de Desenvolvimento Humano Municipal
Gini	GINI	Índice GINI

Fonte: Autor

Os procedimentos adotados nesta etapa foram cruciais e se deu da seguinte maneira:

1º Utilizando *MySQL Workbench*, foi feita a seleção dos dados de cada *itemset*, do *dataset inicial* (veja *Figura 2*) e salvos em CSV;

Figura 2: Seleção do *Itemset* “tx_incidencia_tuberculose” para Salvar em CSV



Fonte: Autor

2º Utilizando a IDE *Spyder 3*, foi codificado um script em *Python 3* (veja *Figura 3*), usando as biblioteca *matplotlib* e *numpy*, para, ler os CSV's gerados com os dados de cada *itemset* do *dataset* inicial e exibir seus histogramas (veja *Figura 4*) e sua Distribuição de Frequência segundo a Regra de Sturges (veja *Figura 5*);

Figura 3: Script Codificado em *Python 3* - Gerador de Distribuição de Frequências

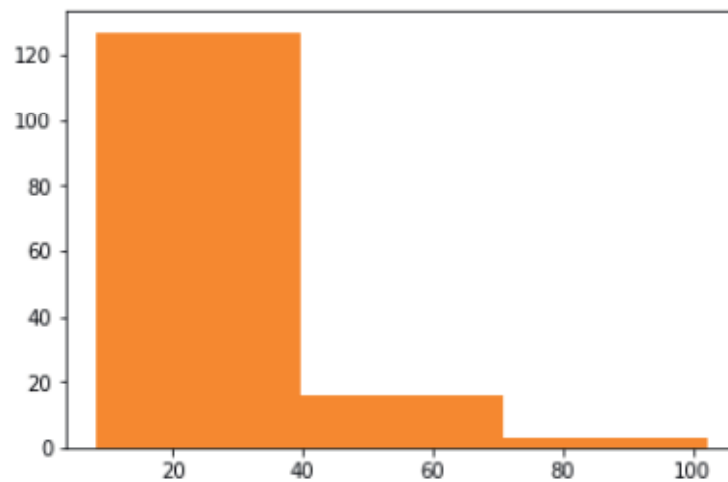
```

1 # -*- coding: utf-8 -*-
2 """
3 Criado Segunda-feira, 05 de fevereiro de 2018
4
5 @author: Kleyton Klaus Guedes de Souza
6 =====
7 Gerador de Distribuição de Frequências
8 ++++++
9 """
10
11 import matplotlib.pyplot as plt
12 from numpy import genfromtxt
13
14 dados = genfromtxt('gini.csv')
15 histograma1 = plt.hist(dados, bins="sturges")
16 |

```

Fonte: Autor

Figura 4: Histograma Gerado no *Spyder 3* –
Itemset "tx_incidencia_tuberculose.csv"



Fonte: Autor

Figura 5: Distribuição de Frequência com a Regra de Sturges, com as Classes para o *Itemset* “tx_incidencia_tuberculose.csv”

	0
0	8.2
1	19
2	29
3	40
4	50
5	60
6	71
7	81
8	92
9	1e+02

Fonte: Autor

3º Conhecendo as classes de cada *itemset*, foram feitas funções SQL (veja *Figura 6*) que retornam respectivamente à identificação abreviada do *itemset*, juntamente com o intervalo de classe, da maneira que segue, “Abreviatura_Itemset + |Intervalo_Classe|”;

Figura 6: Função SQL que Retorna o Nome do *Itemset* da “tx_incidencia_bacilifera” com seu Intervalo de Classe

```

1 CREATE DEFINER='root'@'localhost' FUNCTION 'GetTxIncBacilifera'(tx_inc_bac varchar(45))
2 RETURNS varchar(45) CHARSET utf8
3 begin
4     declare faixa varchar(45);
5     if (tx_inc_bac >= 0 and tx_inc_bac < 12.9) then
6         set faixa = 'tx_inc_bac 0 - 12.9';
7     elseif (tx_inc_bac >= 12.9 and tx_inc_bac < 25.8) then
8         set faixa = 'tx_inc_bac 12.9 - 25.8';
9     elseif (tx_inc_bac >= 25.8 and tx_inc_bac < 38.7) then
10        set faixa = 'tx_inc_bac 25.8 - 38.7';
11    elseif (tx_inc_bac >= 38.7 and tx_inc_bac < 51.6) then
12        set faixa = 'tx_inc_bac 38.7 - 51.6';
13    elseif (tx_inc_bac >= 51.6 and tx_inc_bac < 64.5) then
14        set faixa = 'tx_inc_bac 51.6 - 64.5';
15    elseif (tx_inc_bac >= 64.5 and tx_inc_bac < 77.4) then
16        set faixa = 'tx_inc_bac 64.5 - 77.4';
17    elseif (tx_inc_bac >= 77.4 and tx_inc_bac < 90.2) then
18        set faixa = 'tx_inc_bac 77.4 - 90.2';
19    elseif (tx_inc_bac >= 90.2 and tx_inc_bac < 103) then
20        set faixa = 'tx_inc_bac 90.2 - 103';
21    elseif (tx_inc_bac >= 103 and tx_inc_bac <= 116) then
22        set faixa = 'tx_inc_bac 103 - 116';
23    end if;
24    return faixa;
25 end

```

Fonte: Autor

4º Após codificada todas as funções SQL, através do MySQL Workbench (veja *Figura 7*), para todos os *itemset*, foi criado o *dataset* final em CSV, que recebeu o nome “tuberculose_paraiba_2001-2011_final.csv”. Nem todos os *itemset* do *dataset*

inicial, chamado de “dataset_tuberculose_paraiba_2001-2015_inicial.csv” entraram na seleção, pois foi verificado que os *itemset* com os óbitos por etnia, casos ignorados e índice GINI retornaram muitos dados com poucas classes, o que iria, retornar regras repetitivas ou vazias, dessa forma, o *dataset* final possui 18 *itemset* dos 25 iniciais. Os *Itemset* discretos obtiveram a aparência da Figura 9;

Figura 7: Seleção de Valores de Todos o *Itemset* para criar o *dataset* final com Valores Discretizados

```

1 select
2     GetIncidenciaTuberculose(tx_incidencia_tuberculose) as Tx_Inc_TB,
3     GetTxIncBacilifera(tx_incidencia_bacilifera) as Tx_Inc_Bacilifera,
4     GetBaciliferoCurado(baciliferos_curados) as Bacilifera_Curado,
5     GetTxMortalidade(tx_mortalidade) as Tx_Mortalidade_TB,
6     GetRetratamentoCultura(retratamento_realizaram_cultura) as Retratamento_Cultura,
7     GetTesteHIVRealizado(teste_hiv_realizado) as Teste_HIV,
8     GetCasosNegros(casos_negros) as Casos_Negros,
9     GetCasosAmarelos(casos_amarelos) as Casos_Amarelos,
10    GetCasosPardos(casos_parda) as Casos_Pardos,
11    GetCasosIndigenas(casos_indigena) as Casos_Indigena,
12    GetCasosIgnorados(casos_ignorado) as Casos_Ignorados,
13    GetPercapitaMeioSalario(renda_percapita_meio_salario) as Percapita_Meio_Salario,
14    GetEsgotamentoSanitario(esgotamento_sanitario) as Esgotamento_Sanitario,
15    GetMediaMoradoresDomicilio(media_moradores_domicilios) as Moradores_Domicilios,
16    GetCasosBranco(casos_branco) as Casos_Branco,
17    GetTxEscolarizacao6a4Anos(tx_escolarizacao_6_14_anos) as Tx_Escolarizacao_6a4,
18    GetAcessoSaudeBasica(acesso_saude_basica) as Acesso_Saude_Basica,
19    GetIDM(idm_m) as IDM_M,
20
21 from
22     tb_paraiba;
23

```

Fonte: Autor

Figura 8: Resultado para Alguns *Itemset* do *Dataset* Final Discretizados

Tx_Inc_TB	Tx_Inc_Bacilifera	Bacilifera_Curado	Tx_Mortalidade_TB	Retratamento_Cultura	Teste_HIV	Casos_Negros	Casos_Amarelos	Casos_Pardos
tx inc tb 8 - 19	tx inc bac 12.9 - 25.8	bacilifero curado 100 - 120	tx mortalidade 7.4 - 13	retratamento cultura 0 - 11	teste hiv realizado 78 - 89	casos negros 0 - 11	casos amarelos 0 - 11	casos pardos 89 - 100
tx inc tb 29 - 40	tx inc bac 12.9 - 25.8	bacilifero curado 78 - 91	tx mortalidade 18 - 23	retratamento cultura 0 - 11	teste hiv realizado 89 - 100	casos negros 0 - 11	casos amarelos 0 - 11	casos pardos 78 - 89
tx inc tb 19 - 29	tx inc bac 12.9 - 25.8	bacilifero curado 78 - 91	tx mortalidade 2 - 7.3	retratamento cultura 0 - 11	teste hiv realizado 44 - 56	casos negros 11 - 22	casos amarelos 11 - 22	casos pardos 56 - 67
tx inc tb 8 - 19	tx inc bac 0 - 12.9	bacilifero curado 78 - 91	tx mortalidade 7.4 - 13	retratamento cultura 89 - 100	teste hiv realizado 44 - 56	casos negros 33 - 44	casos amarelos 0 - 11	casos pardos 56 - 67
tx inc tb 19 - 29	tx inc bac 12.9 - 25.8	bacilifero curado 78 - 91	tx mortalidade 7.4 - 13	retratamento cultura 0 - 11	teste hiv realizado 67 - 78	casos negros 33 - 44	casos amarelos 0 - 11	casos pardos 67 - 78
tx inc tb 19 - 29	tx inc bac 12.9 - 25.8	bacilifero curado 100 - 120	tx mortalidade 18 - 23	retratamento cultura 89 - 100	teste hiv realizado 78 - 89	casos negros 0 - 11	casos amarelos 0 - 11	casos pardos 78 - 89
tx inc tb 60 - 71	tx inc bac 51.6 - 64.5	bacilifero curado 100 - 120	tx mortalidade 39 - 45	retratamento cultura 0 - 11	teste hiv realizado 67 - 78	casos negros 44 - 56	casos amarelos 0 - 11	casos pardos 89 - 100
tx inc tb 29 - 40	tx inc bac 12.9 - 25.8	bacilifero curado 65 - 78	tx mortalidade 2 - 7.3	retratamento cultura 89 - 100	teste hiv realizado 56 - 67	casos negros 22 - 33	casos amarelos 11 - 22	casos pardos 67 - 78
tx inc tb 19 - 29	tx inc bac 12.9 - 25.8	bacilifero curado 78 - 91	tx mortalidade 2 - 7.3	retratamento cultura 0 - 11	teste hiv realizado 56 - 67	casos negros 44 - 56	casos amarelos 44 - 56	casos pardos 78 - 89
tx inc tb 19 - 29	tx inc bac 12.9 - 25.8	bacilifero curado 78 - 91	tx mortalidade 7.4 - 13	retratamento cultura 44 - 56	teste hiv realizado 56 - 67	casos negros 44 - 56	casos amarelos 22 - 33	casos pardos 67 - 78
tx inc tb 8 - 19	tx inc bac 12.9 - 25.8	bacilifero curado 100 - 120	tx mortalidade 2 - 7.3	retratamento cultura 0 - 11	teste hiv realizado 56 - 67	casos negros 89 - 100	casos amarelos 0 - 11	casos pardos 56 - 67
tx inc tb 8 - 19	tx inc bac 0 - 12.9	bacilifero curado 65 - 78	tx mortalidade 2 - 7.3	retratamento cultura 0 - 11	teste hiv realizado 44 - 56	casos negros 0 - 11	casos amarelos 0 - 11	casos pardos 67 - 78
tx inc tb 8 - 19	tx inc bac 0 - 12.9	bacilifero curado 78 - 91	tx mortalidade 2 - 7.3	retratamento cultura 0 - 11	teste hiv realizado 44 - 56	casos negros 0 - 11	casos amarelos 0 - 11	casos pardos 89 - 100
Casos_Indigena	Casos_Ignorados	Percapita_Meio_Salario	Esgotamento_Sanitario	Moradores_Casos_Branco	Tx_Escolarizacao_6a4	Acesso_Saude_Basica	IDM_M	
casos indigenas 0 - 11	casos ignorados 0 - 11	percapita meio salario 47 - 54	esgotamento sanitario 0 - 6.4	media ... casos brancos 67 - 78	tx escolarizacao 6a4 anos 89 - 100	acesso saude basica 89 - 100	idm m 0.5 - 0.6	
casos indigenas 0 - 11	casos ignorados 0 - 11	percapita meio salario 47 - 54	esgotamento sanitario 13 - 19	media ... casos brancos 56 - 67	tx escolarizacao 6a4 anos 89 - 100	acesso saude basica 89 - 100	idm m 0.6 - 0.7	
casos indigenas 0 - 11	casos ignorados 22 - 33	percapita meio salario 0 - 6.7	esgotamento sanitario 25 - 51	media ... casos brancos 22 - 33	tx escolarizacao 6a4 anos 89 - 100	acesso saude basica 89 - 100	idm m 0.5 - 0.6	
casos indigenas 0 - 11	casos ignorados 89 - 100	percapita meio salario 47 - 54	esgotamento sanitario 25 - 51	media ... casos brancos 67 - 78	tx escolarizacao 6a4 anos 89 - 100	acesso saude basica 89 - 100	idm m 0.5 - 0.6	
casos indigenas 0 - 11	casos ignorados 56 - 67	percapita meio salario 47 - 54	esgotamento sanitario 6.4 - 13	media ... casos brancos 33 - 44	tx escolarizacao 6a4 anos 89 - 100	acesso saude basica 89 - 100	idm m 0.6 - 0.7	
casos indigenas 0 - 11	casos ignorados 0 - 11	percapita meio salario 47 - 54	esgotamento sanitario 0 - 6.4	media ... casos brancos 56 - 67	tx escolarizacao 6a4 anos 89 - 100	acesso saude basica 89 - 100	idm m 0.5 - 0.6	
casos indigenas 0 - 11	casos ignorados 0 - 11	percapita meio salario 47 - 54	esgotamento sanitario 25 - 51	media ... casos brancos 89 - 100	tx escolarizacao 6a4 anos 89 - 100	acesso saude basica 89 - 100	idm m 0.5 - 0.6	
casos indigenas 0 - 11	casos ignorados 11 - 22	percapita meio salario 47 - 54	esgotamento sanitario 6.4 - 13	media ... casos brancos 22 - 33	tx escolarizacao 6a4 anos 89 - 100	acesso saude basica 89 - 100	idm m 0.5 - 0.6	
casos indigenas 0 - 11	casos ignorados 0 - 11	percapita meio salario 47 - 54	esgotamento sanitario 0 - 6.4	media ... casos brancos 56 - 67	tx escolarizacao 6a4 anos 89 - 100	acesso saude basica 89 - 100	idm m 0.5 - 0.6	
casos indigenas 0 - 11	casos ignorados 0 - 11	percapita meio salario 54 - 60	esgotamento sanitario 6.4 - 13	media ... casos brancos 22 - 33	tx escolarizacao 6a4 anos 89 - 100	acesso saude basica 89 - 100	idm m 0.5 - 0.6	
casos indigenas 0 - 11	casos ignorados 0 - 11	percapita meio salario 47 - 54	esgotamento sanitario 25 - 51	media ... casos brancos 67 - 78	tx escolarizacao 6a4 anos 89 - 100	acesso saude basica 89 - 100	idm m 0.5 - 0.6	
casos indigenas 22 - 33	casos ignorados 0 - 11	percapita meio salario 47 - 54	esgotamento sanitario 25 - 51	media ... casos brancos 56 - 67	tx escolarizacao 6a4 anos 89 - 100	acesso saude basica 89 - 100	idm m 0.5 - 0.6	

Fonte: Autor

Com os dados pré-processados e transformados é possível minerar os dados, através da obtenção de regras de associação, que deu origem aos perfis epidemiológicos sobre a TB. As funções em MySQL podem ser baixado em <<https://github.com/tuberculoseparaiba/funcoesMySQL.git>>.

3.4.1.3. Mineração de Dados

Nesta fase foi utilizada a IDE Spyder 3 para codificar o script em *Python*, chamado “bigData_tuberculose_paraiba_2001-2015.py”, mostrado na Figura 9, responsável por ler o *dataset* final, “tuberculose_paraiba_2001-2011_final.csv”.

O script fez uso da biblioteca pandas, que foi capaz de ler os dados em CSV e armazená-lo em uma estrutura de dados do tipo *Dataframe* (tabela), desta maneira, usando uma implementação do algoritmo apriori para *Python*, chamada, Apyori, foram geradas regras associativas.

O script em Python podem ser baixados em <<https://github.com/tuberculoseparaiba/scriptPython.git>>.

Figura 9: Script em *Python* com a Biblioteca Pandas e Apyori para Minerar o *Dataset*

```

1 import pandas as pd
2
3 # Importa arquivo csv sem cabeçalho (header)
4 dados = pd.read_csv('tuberculose_paraiba_2001-2011_final.csv', header = None)
5
6 #Busca Dimensões do Dataframe linhasxcoluas dimensoes[]
7 dimensoes = dados.shape
8
9 #Ler o Dataframe e armazena-lo em uma lista, chamada transacoes
10
11 transacoes = []
12
13 for i in range(0, dimensoes[0]):
14     transacoes.append([str(dados.values[i, j]) for j in range(0, dimensoes[1])])
15
16 #Gerar Regras de Associação
17 from apyori import apriori
18
19 #Criar Regras Sem Definir Suporte e Confiança
20 regras = apriori(transacoes, min_support = 0.06, min_confidence = 0.8, min_lift = 3 , min_length = 2)
21
22 #Armazenando as Regras Geradas
23
24 resultados = list(regras)
25 resultados2 = [list(x) for x in resultados]
26 #Exibir Resultado Formatado
27 #Refinando a Exibição das Regras de Associação
28 print()
29 print("Resultado Refinado")
30 print()
31
32 resultadoFormatado = []
33
34 for j in range (0,30):
35     #Em resultados [j][2] busca-se os valores da confiança e do lift
36     resultadoFormatado.append([list(x) for x in resultados2[j][2]])
37 print (resultadoFormatado)

```

Fonte: O Autor

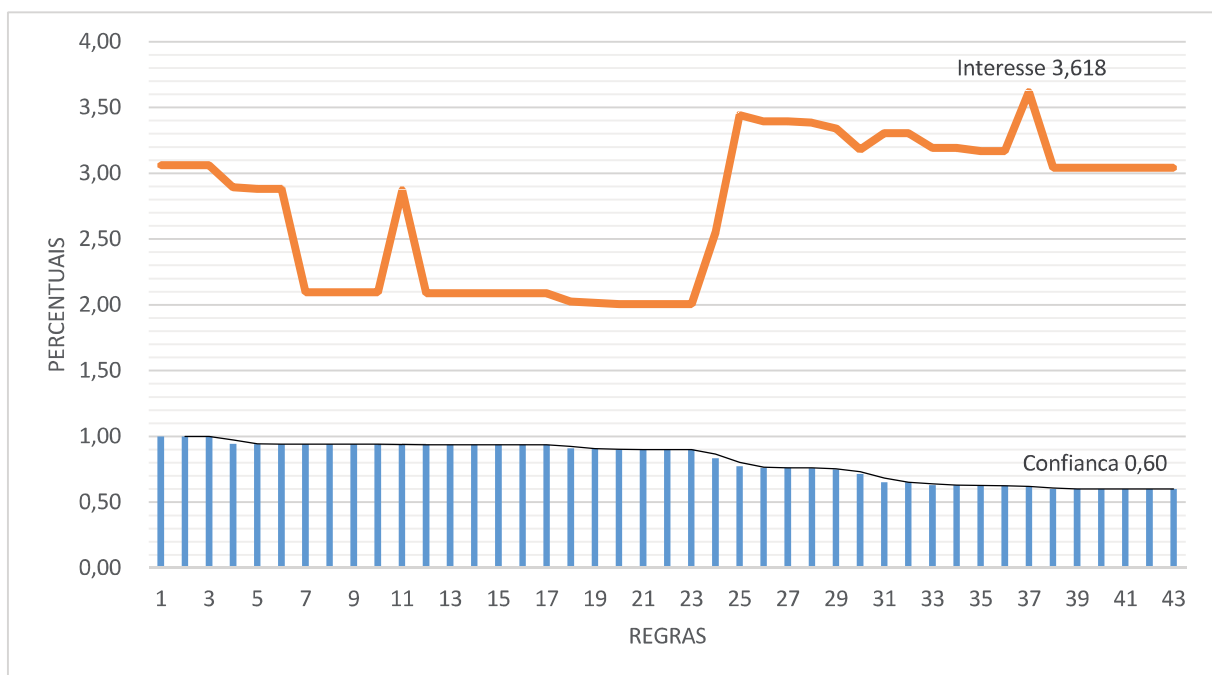
4. RESULTADOS

Para os experimentos, vários valores de parâmetros foram utilizados no algoritmo apriori, afim, de encontrar as regras mais condizentes segundo a bibliografia consultada a respeito da TB, eliminando regras triviais.

A abrangência de uma regra de associação é o número de instâncias para as quais a regra prevê corretamente, comumente conhecida como o suporte da regra. A precisão, convencionada como confiança, é o número de instâncias que a regra prevê corretamente, demonstrada, como o percentual de todas as instâncias a que a regra se aplica (WITTEN *et al.*, 2011) e o interesse é uma medida que demonstra a importância de uma regra, mede o quão frequente um Y (consequente) e X (predecessor) ocorrem juntos se fossem comparados estatisticamente independente. (IBM, 2018).

Os valores utilizados para os parâmetros de suporte mínimo foram mantidos em 0.1, fazendo com que as regras fossem geradas livremente. Os parâmetros confiança mínima e interesse mínimo foram definidos como 0.6 e 2, respectivamente. No total, 43 regras de associação foram geradas, conforme mostra o Gráfico 1.

Gráfico 1: Demonstrativo de Crescimento - Confiança x Interesse em Relação às Regras Geradas



4.1. Resumo dos Resultados

Analisando o Gráfico 1, é possível perceber que das 43 regras geradas, as 23 primeiras possuem predominantemente a confiança entre 1.0 (100%) e 0.9 (90%), mantendo um interesse aproximadamente entre 3 e 2, dentro do suporte escolhido. Da regra 24 a 43, à medida que a confiança diminui, gradativamente o interesse da regra aumenta, respectivamente de 0.9 (90%) a 0.6 (60%) e de 2 a 3 aproximadamente

Desta forma, o Quadro 1 exibe as 23 regras com os perfis gerados que possuem as melhores confianças e interesse. Podem ser interpretadas da esquerda para a direita. O valor anterior ao símbolo “==>” indica o suporte da regra, isto é, os itens abrangidos pela(s) sua(s) premissa(s), separados por ponto e vírgula (;). O valor que aparece após o atributo consequente consiste no número de itens para os quais o consequente da regra é válido. Entre parênteses, há o valor da confiança e do interesse da regra. Desta forma, conforme descrito na regra 1, podemos lê-la da seguinte forma:

“**SE** acesso_saude_basica entre 89 – 100 (%) e percatpta_meio_salario entre 40 – 47 (%) **ENTÃO** idh_m entre 0,6 - 0,7”.

Quadro 1: Resultado das 23 Melhores Regras Geradas

1	{acesso_saude_basica 89 - 100'; 'percatpta_meio_salario 40 - 47'}	==>	{'idh_m 0,6 - 0,7'}	Conf.	Interesse
				(1,00)	(3,063)
2	{acesso_saude_basica 89 - 100'; 'casos_indigenas 0 - 11'; 'percatpta_meio_salario 40 - 47'}	==>	{'idh_m 0,6 - 0,7'}	Conf.	Interesse
				(1,00)	(3,063)
3	{acesso_saude_basica 89 - 100'; 'esgotamento_sanitario 0 - 6,4'; 'percatpta_meio_salario 40 - 47'}	==>	{'idh_m 0,6 - 0,7'}	Conf.	Interesse
				(1,00)	(3,063)
4	{'percatpta_meio_salario 40 - 47'}	==>	{'idh_m 0,6 - 0,7'}	Conf.	Interesse
				(0,94)	(2,892)
5	{'casos_indigenas 0 - 11'; 'percatpta_meio_salario 40 - 47'}	==>	{'idh_m 0,6 - 0,7'}	Conf.	Interesse
				(0,94)	(2,882)
6	{'esgotamento_sanitario 0 - 6,4'; 'percatpta_meio_salario 40 - 47'}	==>	{'idh_m 0,6 - 0,7'}	Conf.	Interesse
				(0,94)	(2,882)
7	{'tx_mortalidade 13 - 18'; 'percatpta_meio_salario 47 - 54'; 'casos_amarelos 0 - 11'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,94)	(2,096)
8	{'tx_mortalidade 13 - 18'; 'percatpta_meio_salario 47 - 54'; 'tx_inc_bac 12,9 - 25,8'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,94)	(2,096)
9	{acesso_saude_basica 89 - 100'; 'percatpta_meio_salario 47 - 54'; 'casos_amarelos 0 - 11'; 'tx_mortalidade 13 - 18'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,94)	(2,096)
10	{acesso_saude_basica 89 - 100'; 'percatpta_meio_salario 47 - 54'; 'tx_inc_bac 12,9 - 25,8'; 'tx_mortalidade 13 - 18'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,94)	(2,096)
11	{'casos_indigenas 0 - 11'; 'esgotamento_sanitario 0 - 6,4'; 'percatpta_meio_salario 40 - 47'}	==>	{'idh_m 0,6 - 0,7'}	Conf.	Interesse
				(0,94)	(2,871)
12	{'tx_mortalidade 13 - 18'; 'casos_indigenas 0 - 11'; 'casos_amarelos 0 - 11'; 'percatpta_meio_salario 47 - 54'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,94)	(2,088)
13	{'tx_mortalidade 13 - 18'; 'percatpta_meio_salario 47 - 54'; 'casos_amarelos 0 - 11'; 'esgotamento_sanitario 0 - 6,4'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,94)	(2,088)
14	{'tx_mortalidade 13 - 18'; 'percatpta_meio_salario 47 - 54'; 'tx_inc_bac 12,9 - 25,8'; 'esgotamento_sanitario 0 - 6,4'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,94)	(2,088)
15	{'tx_mortalidade 13 - 18'; 'acesso_saude_basica 89 - 100'; 'percatpta_meio_salario 47 - 54'; 'casos_indigenas 0 - 11'; 'casos_amarelos 0 - 11'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,94)	(2,088)
16	{'tx_mortalidade 13 - 18'; 'acesso_saude_basica 89 - 100'; 'percatpta_meio_salario 47 - 54'; 'esgotamento_sanitario 0 - 6,4'; 'casos_amarelos 0 - 11'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,94)	(2,088)
17	{'tx_mortalidade 13 - 18'; 'acesso_saude_basica 89 - 100'; 'percatpta_meio_salario 47 - 54'; 'tx_inc_bac 12,9 - 25,8'; 'esgotamento_sanitario 0 - 6,4'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,94)	(2,088)
18	{acesso_saude_basica 89 - 100'; 'tx_mortalidade 13 - 18'; 'casos_amarelos 0 - 11'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,91)	(2,025)
19	{acesso_saude_basica 89 - 100'; 'tx_mortalidade 13 - 18'; 'casos_amarelos 0 - 11'; 'esgotamento_sanitario 0 - 6,4'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,90)	(2,015)
20	{'tx_mortalidade 13 - 18'; 'percatpta_meio_salario 47 - 54'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,90)	(2,005)
21	{acesso_saude_basica 89 - 100'; 'percatpta_meio_salario 47 - 54'; 'tx_mortalidade 13 - 18'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,90)	(2,005)
22	{'tx_mortalidade 13 - 18'; 'casos_indigenas 0 - 11'; 'casos_amarelos 0 - 11'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,90)	(2,005)
23	{acesso_saude_basica 89 - 100'; 'casos_indigenas 0 - 11'; 'casos_amarelos 0 - 11'; 'tx_mortalidade 13 - 18'}	==>	{'tx_inc_tb 19 - 29'}	Conf.	Interesse
				(0,90)	(2,005)

As regras descobertas, apresentadas no Quadro 1, estabelecem uma relação com o objetivo inicial da pesquisa que era responder a questão da existência de relação entre as regras obtidas com uso do *Big Data Analytic*, com os indicadores epidemiológicos da tuberculose nos municípios da Paraíba. As regras foram validadas pela especialista em epidemiologia como importantes para o entendimento de como a tuberculose manifesta-se na Paraíba.

4.2. Análise Comparativa entre Medidas de Interesse

Todas as regras abordam pontos que são relevantes como indicadores epidemiológico de ocorrência da TB na Paraíba, citados na bibliografia consultada, como o acesso à saúde, renda, acesso a saneamento básico, taxa de mortalidade e relacionam-se sempre com IDH-M e taxa de incidência de tuberculose no município.

Para o melhor entendimento da força das regras criadas foram analisadas as três primeiras regras do Quadro 1, validas pela especialista em epidemiologia, que as considerou como inferências coerentes. Podendo ser compreendidas conforme é demonstrado abaixo:

- **Regra 1 - {'acesso_saude_basica 89 - 100'; 'percatpta_meio_salario 40 - 47'}==>{'idh_m 0,6 - 0,7'} / Conf. (1,00) / Interesse(3,063)** – Se os habitantes dos municípios da Paraíba que possuem acesso à saúde básica entre 89% a 100% e que possuem renda per capita de meio salário entre 40% e 47%, então possui IDHM entre 0,6 e 0,7. Fica evidente a ocorrência de casos de tuberculose em populações que tem acesso à saúde básica, porém, uma baixa renda, mesmo possuindo um IDHM aceitável. Com uma confiança de 1,00 (100%) e interesse de 3,063;
- **Regra 2 - {'acesso_saude_basica 89 - 100'; 'casos_indigenas 0 - 11'; 'percatpta_meio_salario 40 - 47'} / Conf. (1,00) / Interesse(3,063)** – Se os habitantes dos municípios da Paraíba que possuem acesso à saúde básica entre 89% a 100%, e os casos de incidência de TB são de índios entre 0(%) e 11(%), e possui renda per capita de meio salário entre 40% e 47%, então possui

IDHM entre 0,6 e 0,7. Fica evidente a ocorrência de casos de tuberculose em populações que tem acesso à saúde básica, quando ocorre em comunidades indígenas é baixa, possuindo baixa renda, mesmo, com um IDHM aceitável. Com uma confiança de 1,00 (100%) e interesse de 3,063;

- **Regra 3 - {'acesso_saude_basica 89 - 100'; 'esgotamento_sanitario 0 - 6,4'; 'percatpta_meio_salario 40 - 47'}/ Conf. (1,00) / Interesse(3,063)** – Se os habitantes dos municípios da Paraíba que possuem acesso à saúde básica entre 89% a 100%, e esgotamento sanitário entre 0% a 6,4%, e renda per capita de meio salário entre 40% e 47%, então possui IDHM entre 0,6 e 0,7. Fica evidente a ocorrência de casos de tuberculose em populações que tem acesso à saúde básica, o acesso a esgotamento sanitário é precário e possuem baixa renda, mesmo, possuindo um IDHM aceitável. Com uma confiança de 1,00 (100%) e interesse de 3,063.

5. CONCLUSÃO

As regras de associação encontradas neste trabalho com o *Big Data Analytics* utilizando o algoritmo apriori estão em concordância com a bibliografia consultada para elaboração deste trabalho, sua semântica foi validada pela especialista. A mineração de dados utilizada como análise exploratória dos dados, e, neste caso, usando regras de associação, proporcionou a descoberta de conhecimento esperada para os dados sobre a tuberculose na Paraíba.

As regras geradas foram relevantes e estão alinhadas com as características da tuberculose de forma geral. Analisando as regras 1, 2 e 3, do Quadro 1, pode-se concluir, que, os municípios que apresentaram incidência de tuberculose em sua grande parte tem acesso à saúde básica, num percentual de 89% a 100%, os casos de tuberculose estão estritamente ligados à falta de saneamento básico e à baixa renda dos municípios. De forma paradoxal, algumas regras apontam que o IDH-M não é considerado baixo (IPEA, 2013) e está em consonância com o IDH do Brasil, que é 0,699. (IBGE, 2010), dessa forma, evidencia-se a falta de distribuição de renda nos municípios envolvidos na pesquisa.

Todas as outras regras tinham como antecedente pelo menos um item relacionado a valores socioeconômicos baixos, que em sua maioria implicaram na ocorrência de taxas de incidência de tuberculose entre 19 e 29 casos por 100.000 hab/ano.

Dessa forma, levando em consideração os expostos até agora, o uso *Big Data Analytics* foi capaz de apontar a necessidade de melhorar a qualidade de vida nos municípios da Paraíba, na tentativa de eliminar os casos de tuberculose. A necessidade de aumentar o acesso ao saneamento básico e a distribuição de renda ficam evidenciados pelos perfis encontrados para alcançar o êxito na eliminação da TB.

REFERÊNCIAS

AGRAWAL, R.; IMIELINSKI, T.; Swami, A. "**Database Mining: A Performance Perspective**", IEEE Transactions on Knowledge and Data Engineering, Special issue on Learning and Discovery in KnowledgeBased Databases. Vol. 5, pp.914, 925.

AGRAWAL, R; SRIKANT, R. "**Fast Algorithms for Mining Association Rules in Large Databases**". In **Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)**, pp. 487 – 499.

APYORI. Sobre o Apyori 1.0.0. Disponível em:
<<https://pypi.python.org/pypi/apyori/1.0.0>>. Acesso em: 21 jan. 2018.

ARAÚJO, K. M. F. Azevedo. **Distribuição espacial da tuberculose e a correlação com as desigualdades sociais**. Dissertação apresentada à Universidade Estadual da Paraíba – UEPB. 95f. Universidade Estadual da Paraíba. Campina Grande: 2012.

ÁVILA, Thiago. **O que faremos com os 40 trilhões de gigabytes de dados disponíveis em 2020**. Disponível em: <<http://thiagoavila.com.br/sitev2/dados-abertos/o-que-faremos-com-os-40-trilhoes-de-gigabytes-de-dados-disponiveis-em-2020/>>. Acesso em: 11 Jan. 2018.

BARANAUSKAS, José Augusto. Slide - Regras de Associação. Departamento de Física e Matemática – FFCLRP-USP. Disponível em: <<http://dfm.ffclrp.usp.br/~augusto>>. Acesso em: 09 jan. 2018.

BIGDATABUSINESS. **Big Data Analytics: você sabe o que é?** Disponível em: <<http://www..com.br/voce-sabe-o-que-e-big-data-analytics/>>. Acesso em: 09 jan. 2018.

CABENA, P; HADJINIAN, P; STADLER, R; JAAPVERHEES; ZANASI, A. **Discovering Data Mining: From Concept to Implementation**. Prentice Hall, 1998.

CAMILO, Cássio Oliveira. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. Technical Report - RT-INF_001-09 - Relatório Técnico**. Universidade Federal de Goiás. Goiás: 2009.

CASANOVA, A. A; LABIDI, S. **Algoritmo da Confiança Inversa para Mineração de Dados Baseado em Técnicas de Regras de Associação e Lógica Nebulosa**. XXV. Congresso da Sociedade Brasileira de Computação, 2005.

CHIAVEGATTO FILHO, Alexandre Dias Porto. **Uso de big data em saúde no Brasil: perspectivas para um futuro próximo.** *Epidemiol. Serv. Saúde* [online]. 2015, vol.24, n.2, pp.325-332. ISSN 1679-4974. Disponível em: <<http://dx.doi.org/10.5123/S1679-49742015000200015>>. Acesso em: 03 nov. 2017.

CIO. **Qual linguagem de programação voltada a Big Data devo usar?** Disponível em: <<http://cio.com.br/tecnologia/2016/04/06/qual-linguagem-de-programacao-voltada-a-big-data-devo-usar/>>. Acesso em: 10 ago. 2017.

CIOS, K. J; PEDRYCZ, W; SWINIARSKI, R. W; KURGAN, L. A. **Data Mining - A Knowledge Discovery Approach.** Springer: 2007.

CIR – Comissões Intergestores Regionais. COSEMSPB. Disponível em: <<http://cosemspb.org/cir/>>. Acesso em: 08 jan. 2018.

COELI, Cláudia Medina; PINHEIRO, Rejane Sobrino; CAMARGO JR., Kenneth Rochel de. **Conquistas e desafios para o emprego das técnicas de record linkage na pesquisa e avaliação em saúde no Brasil.** *Epidemiol. Serv. Saúde*, Brasília, v. 24, n. 4, p. 795-802 . Disponível em: <http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742015000400023&lng=pt&nrm=iso>. Acesso em: 22 dez. 2017.

COMISSÃO INTERGESTORES BIPARTITE (Paraíba). **Resolução nº 488, de 28 agosto de 2008. Aprova o plano diretor de regionalização.** João Pessoa, 2008.

COMISSÃO NACIONAL DE ÉTICA EM PESQUISA (Brasil). **Manual operacional para comitês de ética em pesquisa.** Brasília, DF, 2002. (Série CNS – Cadernos Técnicos; Série A. Normas e Manuais Técnicos, n. 133).

COUTINHO, L. A. S. de Araújo et al. **Perfil Epidemiológico da Tuberculose no Município de João Pessoa – PB, entre 2007 – 2010.** João Pessoa, v. 16, n. 1 (2012), p. 35-42, abr. 2012. ISSN 1415-2177. Disponível em: <<http://periodicos.ufpb.br/ojs/index.php/rbcs/article/view/10172>>. Acesso em: 20 set. 2017. doi:10.4034/RBCS.2012.16.01.06.
DATASUS, Gerenciador de Ambiente Laboratorial – GAL. Disponível em: <<http://gal.datasus.gov.br/GALL/index.php>>. Acesso em: 19 fev. 2018.

DARONCO, Alexandre et al. **Aspectos relevantes sobre tuberculose para profissionais de saúde.** *Revista de Epidemiologia e Controle de Infecção*, Santa Cruz do Sul, v. 2, n. 2, p. 61-65, abr. 2012. ISSN 2238-3360. Disponível em:

<<https://online.unisc.br/seer/index.php/epidemiologia/article/view/2599/2063>>. Acesso em: 27 dez. 2017. doi:<http://dx.doi.org/10.17058/reci.v2i2.2599>.

DAVENPORT, T.H. **Big data no trabalho: derrubando mitos e descobrindo oportunidades**. Tradução Cristina Yamagami. Rio de Janeiro: Elsevier, 2014.

FAYYAD, U; PIATETSKY-SHAPIO, G; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. American Association for Artificial Intelligence, 1996.

FELDMAN, R.; SANGER, J. **The text mining handbook: advanced approaches in analyzing unstructured data**. Reino Unido: Cambridge University Press, 2007.

FERRAZ, Inhaúma Neves. Conhecimento do mundo como instrumento enriquecedor dos resultados obtidos na mineração de dados. Tese apresentada ao Curso de Pós-Graduação em Computação da Universidade Federal. Fluminense, 155f. Niterói: 2008.

FIGUEIREDO, T.M.R.M. *et al.* **Avaliação dos serviços de atenção secundária e primária à saúde no controle da tuberculose, município de Campina Grande-PB-Brasil**. 2º Congresso Brasileiro De Política, Planejamento e Gestão Em Saúde:

FOGO, José Carlos. **Análise Descritiva de Dados**. Departamento de Estatística UFSCar. Disponível em: <http://www.ufscar.br/jcfogo/IEP/arquivos/PE_Variaveis_Quantitativas.pdf>. Acesso em: 21 jan. 2018.

Free Software Foundation(FSF). Disponível em: <<http://www.fsf.org/>>. Acesso em: 19 dez. 2017.

Gantz, John and Reinsel. **The Digital Universe In 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East**. EMC Corporation. Disponível em: <http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf>>. Acesso em: Jan. 2018.

GREGO, M. **Conteúdo digital dobra a cada dois anos no mundo: Se todo conteúdo digital do mundo fosse armazenado em iPads, eles formariam uma pilha com altura igual a dois terços da distância entre a Terra e a Lua**. Disponível em:

<<https://exame.abril.com.br/tecnologia/conteudo-digital-dobra-a-cada-dois-anos-no-mundo/>>. Acesso em: 19 dez. 2017.

HAN, J; KAMBER, M. **Data Mining: Concepts and Techniques**. 2nd Edition. Morgan Kaufmann Publishes. Elsevier. 745f. USA: 2006.

HERNANDEZ-LEAL, Emilcy J.; DUQUE-MENDEZ, Néstor D.; MORENO-CADAVID, Julián. **Big Data: una exploración de investigaciones, tecnologías y casos de aplicación. Tecno. Lógicas**. v. 20, n. 39, p. 17-24. Medellín: 2017. Disponível em: <http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0123-77992017000200002&lng=en&nrm=iso>. Acesso em: 27 Dez. 2017.

HOPPEN, J. Características importantes para diferenciar BI, Data Mining e Big. Aquare. Abr. 2015. Disponível em: <<https://aquare.la/7-caracteristicas-importantes-para-diferenciar-bi-data-mining-e-big-data/>>. Acesso em: 10 nov. 2017.

IBGE. **Informações sobre os municípios brasileiros**. Disponível em: <<http://www.ibge.gov.br/cidadesat/topwindow.htm?1>>. Acesso em 20 dez. 2017.

IBM. **O Processo de Data Mining**. Disponível em: <https://www.ibm.com/support/knowledgecenter/pt-br/SSEPGG_9.5.0/com.ibm.i.m.easy.doc/c_dm_process.html>. Acesso em: 09 Jan. 2018.

IPEA. **O Índice de Desenvolvimento Humano Municipal Brasileiro: Série Atlas do Desenvolvimento Humano no Brasileiro**. Edição PNUD, Brasil: 2013. Disponível em: <http://www.ipea.gov.br/portal/images/stories/PDFs/130729_AtlasPNUD_2013.pdf>. Acesso em: 28 Fev. 2018

LANGE, L. C. **Mineração de dados em sistema eficiente de iluminação pública incluindo parâmetros sócio-comportamentais**. Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre-RS, 2007.

LIBRELOTTO, Solange Rubert; MOZZAQUATRO, Patricia Mariotto. **Análise dos Algoritmos de Mineração J48 e apriori Aplicados na Detecção de Indicadores da Qualidade de Vida e Saúde**. RevInt- Revista Interdisciplinar de Ensino, Pesquisa e Extensão, vol.1 n°1, 2013. Disponível em: <<http://www.revistaeletronica.unicruz.edu.br/index.php/eletronica/article/viewFile/26-37/pdf>>. Acesso em: 08 jan. 2018.

Manual de Recomendações para o Controle da Tuberculose no Brasil, Ministérios da Saúde, Brasília - DF, 2011.

MELLO, Raquel Gama Soares de. **Utilização de *Big Data Analytics* nos Sistemas de Medição de Desempenho: Estudos De Caso**. 111f. Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de São Carlos, São Carlos: 2015.

Ministério da Saúde. **Portfólio de Projetos BI. 2017**. Disponível em: <http://www2.datasus.gov.br/DATASUS/APRESENTACAO/BI/PROJETOS%20DE%20BI%20NA%20CGDIS_html/html/index.html#2>. Acesso em: 09 jan. 2018.

MYSQL. Sobre o MySql Community. Disponível em: <<https://dev.mysql.com/downloads/mysql/>>. Acesso em: 21 jan. 2018.

ORACLE. **Big data fortalecerá a tomada de decisão estratégica em 2016**. São Paulo, 2016. Disponível em: <<https://www.oracle.com/br/corporate/pressrelease/big-data-will-strengthen-strategic-decision-making-in-2016-20160105.html>>. Acesso em: 09 jan. 2018.

PACHIAROTTI, Juan Francisco Beis. **Aplicação de Técnicas de Mineração de Dados no aprimoramento do Atendimento Médico em um Cenário de Plano de Saúde**. Monografia apresentada para obtenção do Grau de Bacharel em Ciência da Computação pela Universidade Vila Velha. 126f. Vila Velha: 2012. Disponível em: <http://www.uvv.br/edital_doc/Aplica%C3%A7%C3%A3o%20de%20T%C3%A9cnicas%20de%20Minera%C3%A7%C3%A3o%20de%20Dados%20no%20aprimoramento%20do%20Atendimento%20M%C3%A9dico%20em%20um%20Cen%C3%A1rio%20de%20Plano%20de%20Sa%C3%BAde.pdf>. Acesso em: 08 jan. 2018.

PALANCAR, J; LEÓN, R; PAGOLA, J. M; HECHAVARRÍA, A. **A compressed vertical binary algorithm for mining frequent patterns**. STUDIES IN COMPUTATIONAL INTELLIGENCE. pP. 197-211. Springer-Verlag: 2008.

PEREIRA, V. A. S. **Big Data: Um Estudo Em Gestão Empresarial**. 80f. Monografia (TCC) apresentado ao Curso de Biblioteconomia e Gestão de Unidades de Informação da Universidade Federal do Rio de Janeiro, 2016. Disponível em: <<http://pantheon.ufrj.br/bitstream/11422/169/1/Big%20data%20-%20um%20estudo%20em%20gest%C3%A3o%20empresarial.pdf>>. Acesso em: 19 dez. 2017.

RIBEIRO, Carla Danielle Silva. **19º Congresso Brasileiro dos Conselhos de Enfermagem (CBCENF)**. 2016. Disponível em:

<<http://apps.cofen.gov.br/cbcenf/sistemainscricoes/arquivosTrabalhos/I40909.E10.T8050.D6AP.pdf>>. Acesso em: 27 out. 2017.

RAGHUPATHI, Wullianallur; RAGHUPATHI, Viju. **Big Data Analytics in healthcare: promise and Potential. Health Information Science and Systems 2014**. BioMed Central. Disponível em: <<http://www.hissjournal.com/content/2/1/3>>. Acesso em: 09 jan. 2018.

REZENDE, S.O. **Mineração de dados**. In: XXV Congresso da Sociedade Brasileira de Computação. São Leopoldo: SBC, 2005.

SANTOS, Ivânia Ramos; FERREIRA, Juliana Pagnoncelli. Big Data: Armazenamento, análise e gerenciamento. Disponível em: <<https://www.devmedia.com.br/big-data-armazenamento-analise-e-gerenciamento/30918>>. Acesso em: 27 dez. 2017.

Secretaria de Estado da Saúde. **O que é tuberculose. 2017**. Disponível em: <<http://www.saude.sp.gov.br/ses/perfil/cidadao/temas-de-saude/tuberculose/o-que-e-tuberculose>>. Acesso em: 18 nov. 2017.

SILVA, Michel de A. *et al.* **Aplicação do algoritmo apriori para uma base de dados de ictioplâncton em um reservatório de água doce da Amazônia Legal**. Programa de Pós-Graduação Modelagem Computacional de Sistemas – Universidade Federal do Tocantins (UFT). Palmas: 2013. Disponível em: <<http://www.lbd.dcc.ufmg.br/colecoes/eniac/2013/0038.pdf> >. Acesso em: 08 Fev. 2018.

SILVA, Talina Carla da. **Tuberculose e sua Relação com a Vulnerabilidade Social: Uma Abordagem Espacial**. Dissertação apresentada ao Programa de Pós-graduação Mestrado em Saúde Pública da Universidade Estadual da Paraíba. 53f. Universidade Estadual da Paraíba. Campina Grande: 2014.

SOUSA, Selda Gomes de. **Avaliação da Vigilância Epidemiológica do Estado da Paraíba**. Doutorado em Saúde Pública. Fundação Oswaldo Cruz: Centro De Pesquisas Aggeu Magalhães. 146f. Recife: 2014.

SOUZA, Amanda Damasceno et al. **A Informação em Oncologia na Era do Big Data**. XVI Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação. Out. 2015. Universidade Federal da Paraíba. ISSN 2177-3688. Disponível em: <<http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/enancib2015/paper/viewFile/3046/1270>>. Acesso em: 02 nov. 2017.

SPYDER. Sobre o Spyder 3. Disponível em: <<https://pythonhosted.org/spyder/>>. Acesso em: 21 jan. 2018.

Stop TB Partnership. **Políticas de TB em 29 países: Uma pesquisa de prevenção, Teste, tratamento, políticas e práticas** [tradução nossa]. Jul. 2017. Disponível em: <<http://www.stoptb.org/>>. Acesso em: 27 Dez. 2017.

TAURION, Cezar. **As oportunidades que trazem não podem nem devem ser desperdiçadas**. Jun. 2016. Disponível em: <<http://cio.com.br/tecnologia/2016/06/17/os-5-vs-do-big-data/>>. Acesso em: 27 dez. 2017.

THE PARADIGM SHIFT: 2016-2020. Plano Global para Acabar com TB [tradução nossa]. Genebra, Suíça. Disponível em: <<http://www.stoptb.org/>>. Acesso em: 27 Dez. 2017.

Universalidade, Igualdade e Integralidade Da Saúde: Um Projeto Possível. BELO HORIZONTE: 2013.

VERDE, M. A. R. F. Baptista. **Market Basket Analysis e aplicação de Regras de Associação Hierárquica: um caso de estudo numa empresa de retalho portuguesa**. Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão. 74f. Portugal: 2016

VIANNA, William Barbosa; DUTRA, Moisés Lima; FRAZZON, Enzo Morosini. **Big Data e Gestão da Informação: Modelagem do Contexto Decisional Apoiado pela Sistemografia**. *Inf., Londrina*, v. 21, n. 1, p. 185 – 212, jan./abr. 2016. DOI: 10.5433/1981-8920.2016v21n1p185.

WHO. **GLOBAL TUBERCULOSIS REPORT 2017**. ISBN 978-92-4-156551-6. Disponível em: <http://www.who.int/tb/publications/global_report/en/>. Acesso em: 08 jan. 2018.

WU *et al.* **"Top 10 algorithms in data mining"**. Knowledge and Information Systems, 2007. Vol. 14, pp. 1 – 37.

ZAKI, M. J.; GOUDA, Karam. **Fast Vertical Mining Using Diffsets**. The Blavatnik School of Computer Science.