Inf. Pauta Fortaleza, CE v. 3 Número especial nov. 2018 ISSN 2525-3468
DOI: https://doi.org/10.32810/2525-3468.ip.v3iEspecial.2018.39723.161-181

ARTIGO

APLICAÇÃO DE MINERAÇÃO DE DADOS EM INFORMAÇÔES ORIUNDAS DE PRONTUÁRIOS DE PACIENTE[i]

APPLICATION OF DATA MINING IN INFORMATION CONCERNING PATIENT RECORDS

Ricardo César de Carvalho[1]

[1] Mestre em Ciência da Informação pela Universidade Estadual Paulista (UNESP).

E-mail: ricardo.cc@ifsp.edu.br

Conflito de interesses: O autor declara que não há conflito de interesses.

Financiamento: Não há.

Declaração de Disponibilidade dos dados: Todos os dados relevantes estão disponíveis neste artigo.

Recebido em: 20/09/2018.
Revisado em: 01/10/2018.
Aceito em: 10/10/2018.

Como citar este artigo:
CARVALHO, Ricardo César. Aplicação de mineração de dados em informações oriundas de prontuários de paciente. **Informação em Pauta**, Fortaleza, v. 3, número especial, p. 161-181, nov. 2018. DOI: https://doi.org/10.32810/2525-3468.ip.v3iEspecial.2018.39723.161-181.

RESUMO

Este artigo procura investigar a aplicação da Mineração de Dados na descoberta de conhecimento oriundo de informações provenientes de prontuários do paciente. Diante disso, o objetivo foi examinar a bibliografia na busca da utilização, resultados e investimentos na área. A metodologia utilizada consistiu no levantamento bibliográfico, por meio de revisão de literatura e a aplicação de uma etapa da mineração de dados, a importação em dados provenientes da saúde. Conclui-se que a Mineração de Dados é eficiente, já existem muitas pesquisas e investimentos de grandes empresas e neste momento, possui um grande potencial de crescimento.

Palavras-chave: Mineração de Dados. Prontuário do Paciente. Descoberta de Conhecimento em Bases de Dados.

ABSTRACT

This article seeks to investigate the application of Data Mining in the discovery of knowledge derived from information from patient records. Therefore, the objective was to examine the bibliography in the search of the use, results and investments in the area. The methodology used consisted of a literature review, through literature review and the application of a step of data mining, the importation into health data. It is concluded that Data Mining is efficient, there are already a lot of researches and investments of large companies and now, it has a great growth potential.

Keywords: Data Mining. Medical Records. Knowledge Discovery in Databases.

1 INTRODUÇÃO

Hoje acontece em todos os seguimentos da sociedade um crescimento exponencial dos dados que são gerados, quantidades de dados sem precedentes, isso leva a uma necessidade de se extrair informações na busca de conhecimento (ISOTANI, BITTENCOURT, 2015). Para exemplificar esse volume dos dados, em 2014, a IDC *Digital Universe* estimou que o universo digital, dados criados e copiados, no ano de 2013 era de 4.4 trilhões de *gigabytes* e que em 2020 deve estar próximo de 44 trilhões de *gigabytes*, um crescimento anual de 40%, quase dobrando a cada 2 anos. (DIGITAL UNIVERSE, 2014).

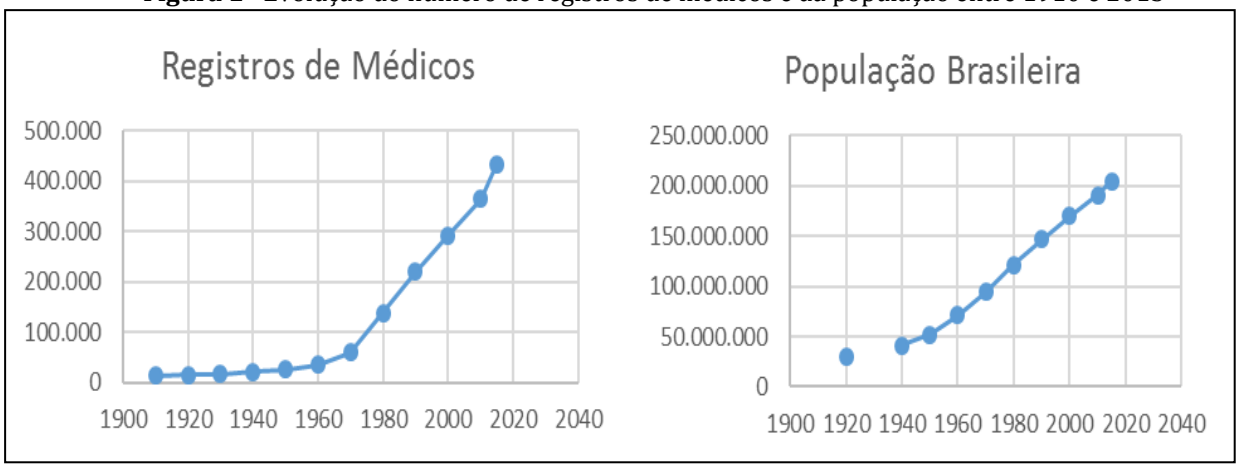
O Brasil, seguindo as tendências mundiais, também enfrenta problemas com o volume dos dados e a dificuldade de acesso e disponibilização dessas informações. Neste contexto, diversas tecnologias digitais de informação e de comunicação marcaram presença, alterando a forma de como as pessoas lidam com a informação. Na área da saúde, particularmente, elas fornecem recursos para a geração, controle, manutenção e arquivamento dos dados vitais dos pacientes, das pesquisas biomédicas e na captura e disponibilização de imagens diagnósticas, que, inclusive, "refletem" os nossos órgãos, átomos, células e moléculas mais internas. No caso do prontuário do paciente, conforme evidencia o Conselho Federal de Medicina (BRASIL, 2016) o uso dessas tecnologias é irreversível, o suporte de registro analógico (em papel) está migrando para o eletrônico, possibilitando inúmeras vantagens.

E os profissionais que trabalham diretamente com esses prontuários também cresce no país. A respeito do total de médicos em atividade no Brasil, em 12 de setembro de 2018, de acordo com Conselho Federal de Medicina (2018), existem 457.719 registros médicos para uma população de 208.822.860 habitantes, segundo o IBGE (BRASIL, 2018a). Para se ter uma ideia, no ano de 1970 haviam 58.994 médicos, em comparação com 2018, esse aumento foi de 675%, sendo que no mesmo período, a população brasileira cresceu 124% (segundo IBGE a população em 1970 era de 93.139.037 habitantes).

A pesquisa de Scheffer (2015), demonstra na figura 1, o crescimento na formação desses profissionais e uma previsão de continuidade de crescimento no futuro. Outro dado que a pesquisa apresenta, foi que médicos que atendem em consultórios tem um maior número de vínculos empregatícios, fazem jornadas mais longas, são na maioria especialistas e recebem os maiores salários. Em resumo, tais dados revelam que há

muitos médicos no Brasil que atendem pacientes em consultórios, tem pouco tempo e ganham mais. Essa constatação sugere que a existência de ferramentas que pudessem facilitar o trabalho destes profissionais pode ser bem-vinda, e até mesmo poderia ajudar no processo de diagnóstico, melhorando a qualidade de vida dos pacientes.

Figura 1– Evolução do número de registros de médicos e da população entre 1910 e 2015



Fonte: SCHEFFER (2015).

Neste contexto de muitos dados em saúde e muitos profissionais criando mais informações a cada dia, citando apenas o Brasil, é que podemos pensar em estratégias para recuperar informações relevantes aos usuários. Uma forma de tratar esse problema é a utilização de ferramentas de recuperação de informação, que pode ser descrita como "um processo ou método pelo qual um usuário da informação em potencial é capaz de converter a sua necessidade de informação em uma lista real de citações de documentos armazenados contendo informações úteis." (MOOERS, 1951, p. 25, tradução nossa).

E retomando o foco de dados em saúde, vários estudos concluem que a análise desses dados pode fornecer informações importantes para tomada de decisão e sua utilização no processo de diagnóstico médico, e até mesmo citando a Mineração de Dados como uma das ferramentas mais viáveis para isto (ANANIADOU; KELL; TSUJII, 2006; ZWEIGENBAUM *et al.*, 2007; FALCÃO *et al.*, 2009; SPASIĆ *et al.*, 2008; SONG, 2013).

Dessa forma, esta pesquisa procura investigar a recuperação de informações e descoberta de conhecimento oriundo de informações provenientes de sistemas de saúde por meio de ferramentas de Mineração de Dados, além de determinar a existência de pesquisas nesta área no Brasil e investimento por parte das empresas na disponibilização de meios para qualquer usuário consiga manipular seus dados e extrair

novas informações, tanto para o trabalho do diagnóstico médico, quanto para o planejamento de políticas públicas.

2 PRONTUÁRIOS DO PACIENTE E OS DADOS DE SAÚDE

De acordo com a Organização Mundial da Saúde (OMS), um sistema de informação para a saúde constitui-se em "[...] um conjunto de componentes que atuam de forma integrada por meio de mecanismos de coleta, processamento, análise e transmissão da informação necessária e oportuna para implementar processos de decisões no Sistema de Saúde." Logo, esse sistema tem o propósito de "[...] selecionar dados pertinentes e transformá-los em informações para aqueles que planejam, financiam, proveem e avaliam os serviços de saúde" (ORGANIZAÇÃO MUNDIAL DA SAÚDE, 1981, p. 42).

O processo de identificar uma informação que atenda às necessidades de um usuário dentre um montante de documentos é denominado recuperação de informação, Ferneda (2012) o descreve assim e também diferencia um sistema de recuperação de informação de um gerenciador de banco de dados devido ao fato de que ele é capaz de recuperar informações que atenderão à expressão de busca levando em consideração o conteúdo do texto recuperado e não somente demonstrar quais documentos contém as informações constantes na expressão de busca.

Estes sistemas de gerenciamento hospitalar permitem a pesquisa de termos a partir de seus dados, embora isso seja feito em informações registradas de forma estruturada no banco de dados, normalmente não oferecendo ferramentas automáticas ou buscas a dados não estruturados, devido a uma série de fatores e problemas encontrados nesse processo que poderiam não garantir a integridade das informações. Em outras situações, os dados dos pacientes se encontram em meio analógico, ou seja, em papel, devido ao fato de serem antigos ou anteriores à era da informatização, de qualquer modo, a busca das informações nesses documentos é mais complexa devido ao estado do papel, da grafia, da qualidade da escrita, entre outros fatores.

Estes documentos que retratam a saúde das pessoas, podem ser descritos, de acordo com o Conselho Federal de Medicina (CFM), no Artigo 1º da Resolução de nº 1.638/2002, definindo como Prontuário do Paciente

> [...] o documento único constituído de um conjunto de informações, sinais e imagens registradas, geradas a partir de fatos, acontecimentos e situações sobre a saúde do paciente e a assistência a ele prestada, de caráter legal, sigiloso e científico, que possibilita a comunicação entre membros da equipe multiprofissional e a continuidade da assistência prestada ao indivíduo. (BRASIL, 2002, p. 1).

É um documento multidisciplinar, temporal e abrangente devido aos diversos tipos de profissionais que registram informações nesse documento e acessam seus conteúdos informacionais. Nesse sentido, Bentes Pinto (2006, p. 36) entende que o prontuário do paciente é

> o documento que contém registradas todas as informações concernentes à condição de saúde de uma pessoa, desde o seu nascimento até a sua morte. Trata-se, portanto, de um documento que contém dados e informações clínicas e não clinicas, de natureza sensível e, portanto, protegidas pelo ordenamento jurídico nacional e internacional. Trata-se da memória escrita da história das condições de saúde de uma pessoa, sendo, portanto, indispensável para a comunicação intra e entre a equipe de saúde e entre ela e o paciente, para a continuidade, a segurança, a eficácia e a qualidade de seu tratamento e acompanhamento, bem como da gestão das organizações de saúde.

Na mesma linha de raciocínio, Galvão e Ricarte (2012) consideram que esse documento é complexo devido a sua forma de produção, conteúdo, organização, acesso e disponibilização, conforme descrito a seguir:

> Evidenciou-se que o prontuário do paciente é um documento informacionalmente complexo quanto ao modo de produção, quanto ao conteúdo que compreende, quanto ao modo de organização e quanto ao modo de acesso e disponibilização. Por esse motivo, seja em suporte papel, seja em suporte eletrônico, para ter melhor qualidade o prontuário demanda planejamento institucional, trabalho cooperativo e permanente da equipe de saúde, dos gestores, dos profissionais da informação e de informática que contam com o conhecimento necessário para sistematizar os aspectos informacionais relacionados ao prontuário. (RICARTE, 2012, p. 45)

Então é possível classificar o Prontuário do Paciente, resumidamente, como um documento: Universal, Temporal, Complexo, Informacional, Legal, Sigiloso, Interdisciplinar, Multidisciplinar e Transdisciplinar, e que para atender ao quesito de universalidade e multidisciplinaridade ainda conta com vocabulários das Equipes de Saúde, Paciente, Família do Paciente, Gestores da Saúde, Advogados, Juízes, Auditores, Docentes, Pesquisadores, Estudantes, entre outros. E se torna o elo de comunicação entre todos estes envolvidos com o intuito de garantir o atendimento à saúde do usuário paciente portador do mesmo.

A tecnologia pode fornecer meios para organizar esses documentos e facilitar o acesso ao conteúdo relevante aos usuários, por meio de ferramentas que a informática aplicada na saúde pode fornecer, descrita como o processo de se utilizar sistemas computacionais visando apoiar e agilizar a administração dos serviços de saúde, cuidados clínicos, investigação médica e treinamento. Isto requer a aplicação das tecnologias de computação, como a Mineração de Dados e de comunicações para otimizar o processamento de informações em saúde, em todas as etapas, como a coleta, armazenamento, recuperação efetiva (no seu devido tempo e lugar) e análise e apoio à decisão para os administradores, médicos, pesquisadores e educadores na medicina. (HOBBS, 2001).

3 MINERAÇÂO DE DADOS

A Mineração de Dados pode ser uma das tecnologias que poderia num tempo muito curto fornecer esses meios de acesso a essas informações, visto que ela é uma tecnologia que pode vasculhar grandes quantidades de informações e trazer resultados muito rapidamente.

Para demonstrar que neste momento já temos a possibilidade de aplicar essas novas tecnologias na busca desse conhecimento, em seu trabalho, Galvão e Ricarte (2015), explicam a transição tecnológica que vem acontecendo com o prontuário do paciente, e o fazem demonstrando através de quatro ondas. Na primeira onda, ocorreu a melhoria da qualidade enquanto suporte em papel, com o intuito de compreender os fluxos e processos envolvidos, com a transição para o suporte por meio de tecnologias e infraestrutura adequadas na segunda onda, em seguida, na terceira onda, começam novas pesquisas para melhorar os conteúdos registrados nos prontuários por meio de terminologias padronizadas e controle de qualidade, e finalmente na quarta onda, a busca da melhoria no planejamento e avaliação da assistência em saúde usando toda a infraestrutura tecnologia e informacional criada para esse fim.

A Mineração de Dados pode se enquadrar nesta quarta onda, porque fornece acesso a conhecimento escondido em grandes quantidades de informações permitindo melhores subsídios, tanto para o profissional da saúde tratar um paciente, quanto para gestores planejarem os investimentos, ou mesmo, a prevenção de endemias que possam acontecer.

Se analisarmos a quantidade de dados gerados diariamente em qualquer ambiente informacional, fenômeno chamado de *Big Data*, que classifica os dados gerados em quantidades monstruosas, produzidos em vários formatos e armazenados em uma grande quantidade de dispositivos e equipamentos. O *Big Data* vem trazendo grandes mudanças na forma de olhar os dados que são gerados, permitindo manipular de novas formas esses dados e trazendo muitas mudanças em todos os setores, por este motivo ele é considerado uma nova revolução industrial (AMARAL, 2016).

Neste cenário de massificação de produção e armazenamento de dados, bem como os processos e tecnologias para extraí-los e analisá-los, acabou por tornar inviável uma análise sistemática baseada em técnicas manuais de estatística de grandes bases de dados, nas quais pesquisadores da área da inteligência artificial, combinando técnicas de estatística e programação avançada conseguiram desenvolver sistemas que podem efetuar a extração e sumarização automática de informações úteis a partir de grandes bases de dados, o que foi chamado de Mineração de Dados (QUILICI-GONZALEZ; ZAMPIROLLI, 2015).

Este termo foi cunhado como alusão ao processo de mineração, porque explora bases de dados brutas (terreno) contendo muito material aparentemente sem utilidade, usando algoritmos especiais (ferramentas) adequados para se obter conhecimento (pedra preciosa) que permita uma tomada de decisão (CASTRO; FERRARI, 2016). Na Mineração de Dados, o objetivo é prover um método automático para descobrir padrões em dados sem a tendenciosidade de uma análise feita meramente por um humano e sua intuição (BRAGA, 2005).

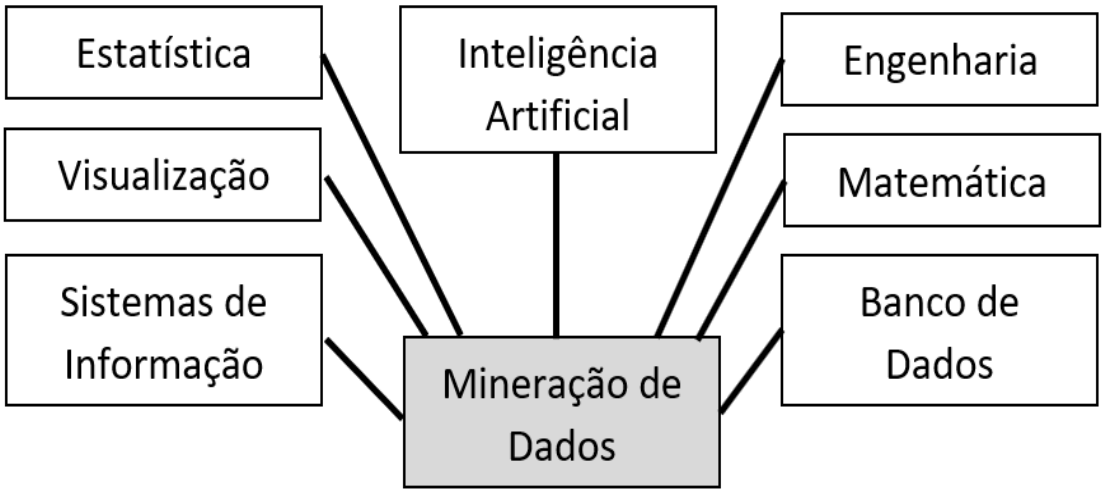
Pode se classificar o processo todo como Descoberta de Conhecimento em Dados Estruturados (*Knowledge Discovery in Databases* - KDD), e nesse contexto vários autores colocam a Mineração de Dados como apenas uma parte desse processo, sendo empregada na etapa de descoberta, que inclui a seleção e integração das bases de dados, a limpeza dessa base, seleção dos dados, transformação dos dados, mineração e avaliação dos dados (CASTRO; FERRARI, 2016).

Ainda assim é considerada uma área muito abrangente, vários autores citaram três grandes tarefas que deveriam ser consideradas as principais da Mineração de Dados: a predição, agrupamentos de dados e associação ou descoberta de regras de associação, e que trabalhando juntas podem compor muitas variações e até permitir a criação de

subtarefas na resolução dos problemas. Também é possível dividir a Mineração de Dados em dois grandes níveis de tarefas, as preditivas e as descritivas. As tarefas preditivas utilizam de valores de atributos descritivos para tentar predizer valores futuros ou desconhecidos, já as tarefas descritivas, têm o objetivo de encontrar padrões que podem descrever os dados de maneira que o ser humano possa interpretar (SILVA; PERES; BOSCARIOLI, 2016; CASTRO; FERRARI, 2016; BRAGA, 2005).

A Mineração de Dados por ser uma disciplina interdisciplinar e multidisciplinar envolve conhecimento de áreas como banco de dados, estatística, aprendizagem de máquina, computação de alto desempenho, reconhecimento de padrões, computação natural, visualização de dados, recuperação da informação, processamento de imagens e de sinais, análise espacial de dados, inteligência artificial, entre outras (CASTRO; FERRARI, 2016). A figura 2 mostra a ligação entre as diversas áreas que contribuem para com a Mineração de Dados, fornecendo ferramentas, técnicas e conhecimento.

Figura 2– A multidisciplinaridade da Mineração de Dados



Fonte: Adaptado de CASTRO; FERRARI (2016).

Os requerimentos para se fazer um processo de Mineração de Dados, na qual é necessário um grande conhecimento prévio a respeito dos dados e do processo, é muito complexo e envolve muitos atores e tecnologias, os autores o descreveram da seguinte forma:

> Contudo, minerar dados para descobrir conhecimento não é uma tarefa trivial. É preciso conhecer os dados, o processo de análise e descoberta, as tarefas e técnicas de mineração e as ferramentas matemáticas e computacionais que se aplicam nesse contexto. Portanto, a descoberta é um processo. Ainda, é preciso

> conhecer o ambiente em que os dados são produzidos e que tipo de conhecimento esse ambiente necessita e espera receber. Enfim, minerar dados exige conhecimento técnico, tempo e dedicação. (SILVA; PERES; BOSCARIOLI, 2016).

Existe uma grande quantidade de ferramentas de Mineração de Dados no mercado, tanto comerciais, quanto livres. As direcionadas para o público comercial são fornecidas por alguns dos maiores fornecedores de tecnologia, como Microsoft, SAS, IBM, Oracle. Já as ferramentas de código aberto mais populares são Weka, R e Orange. É importante destacar que estes conceitos e técnicas aplicadas numa dessas ferramentas podem ser aplicadas em qualquer ferramenta, pois o processo do aprendizado de máquina é uma disciplina universal (AMARAL, 2016).

Neste artigo, para a validação do processo foi escolhida a ferramenta R por alguns aspectos, como por exemplo, ser um ambiente de software livre para computação estatística e gráficos. Ela fornece uma ampla variedade de técnicas estatísticas e ferramentas para manipular os dados (R CORE TEAM, 2015).

Conforme descrito por Peng (2015), que utilizou dados importados para mostrar a utilização da ferramenta R, e por meio dos comandos da ferramenta pôde criar funções, expressões, simulações, estruturas de controle, vetores, além de imagens e gráficos. E ainda para facilitar mais seu uso, existe a plataforma RStudio, que é um ambiente de desenvolvimento integrado aberto e gratuito para o R, que permite a utilização da linguagem através de janelas gráficas e de maneira visual, sendo essa a maior vantagem sobre a ferramenta R pura, que deve ser manipulada somente em modo texto por meio de um terminal (RSTUDIO, 2018).

Como uma das vantagens da utilização da Mineração de Dados, Quilici-Gonzalez e Zampirolli (2015) citam vários exemplos da boa aplicação dessa técnica, inclusive na área da saúde, como usar o sistema para adquirir conhecimento prévio podendo demonstrar que algumas doenças e problemas de saúde podem estar mais associadas a uma certa etnia do que a outras, e dessa forma pudesse auxiliar o médico na hora de solicitar exames médicos, mesmo se o paciente não estiver apresentando sintoma prévio da enfermidade. O autor também demonstra preocupação ao citar que problemas podem ocorrer também, principalmente a respeito do sigilo e privacidade, como por exemplo, os dados de pagamentos feitos por cartão de crédito de uma pessoa que podem expor suas preferências religiosas, os hábitos de compra de livros revelando as preferências políticas das pessoas, ou até mesmo gastos elevados, normais para essa

pessoa, podem colocá-lo no grupo de clientes de alto risco para conseguir empréstimos bancários, sobre sua residência, o seu cep pode apontar que vive numa região problemática ou mesmo num bairro nobre, seus dados médicos podem revelar detalhes que podem custar uma vaga de emprego numa grande empresa que esteja pleiteando uma contratação.

4 DESENVOLVIMENTO

A metodologia utilizada consistiu no levantamento bibliográfico e revisão de literatura para fornecer subsídios teóricos para a compreensão do campo do estudo, por meio de um levantamento de informações e referências em livros, artigos, manuais, leis, decretos, para em seguida, a escolha de uma ferramenta e um conjunto de dados que pudesse representar o cenário escolhido, para a aplicação prática com a utilização dessa ferramenta de Mineração de Dados com o objetivo de testar a aplicabilidade das tecnologias sobre dados provenientes da saúde. Neste momento não foram utilizados dados de pacientes oriundos de prontuários de pacientes, mas a verificação da bibliografia, isto é, se havia estudos sendo desenvolvidos nesta área e a facilidade na utilização de ferramentas informacionais para a utilização por usuários que tenham acesso a tais dados de pacientes e a permissão necessária para fazer quaisquer tipos de consultas utilizando dados reais, mas que ainda desconhecem as ferramentas e técnicas citadas neste trabalho.

Partindo do pressuposto que os dados de um prontuário do pacientes são registrados por profissionais de saúde e outras áreas correlatas, independentemente do local de trabalho, para a criação desses dados existem regras definidas pelo governo ou por entidades de classe, para garantir a qualidade dos dados cadastrados, mas mesmo alguns dados serem estruturados, pois dependem de tabelas de procedimentos, tabelas de exames, siglários, bulário, entre outros, e estarem cadastrados nos banco de dados de uma forma legível para as máquina e para os humanos, existem também os dados não estruturados, cadastrados nos sistemas utilizando a linguagem natural e todos os seus vícios, siglas, símbolos, abreviaturas e outras formas da linguagem, impedindo a busca de forma eficiente por meio das máquinas. Podemos dizer que a qualidade dos dados recuperados está ligada à qualidade dos dados registrados. Outras tecnologias também

podem auxiliar o usuário no trabalho de levantamento de informações em bases de saúde, como a mineração de textos que podem encontrar informações em dados não estruturados, como a anamnese de um prontuário, que é redigido, normalmente, utilizando linguagem natural. E o aprendizado de máquina, que através do resultado da Mineração de Dados é possível criar modelos que aplicados a algoritmos especiais podendo "ensinar" um computador a procurar informações e até fazer inferências a respeito de situações encontradas nas grandes massas de dados.

Durante a pesquisa bibliográfica e documental, alguns artigos com aplicação da Mineração de Dados nessa área foram identificados. Além disso foram encontrados exemplos da utilização dessa tecnologia por parte de grandes empresas com alguns grandes projetos, inclusive já em produção em hospitais no Brasil, que descreveremos em seguida.

Um trabalho que avalia o processo de descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases* - KDD) como ferramenta de avaliação da usabilidade dos profissionais da enfermagem durante a utilização do prontuário eletrônico do paciente. Que por meio da Mineração de Dados poderia complementar o que já se sabe sobre as dificuldades da usabilidade. A opção pela Mineração de Dados em saúde, contemplada no processo KDD, pela eficácia evidenciada na descoberta de padrões que complementem o que os especialistas na área já conhecem (LOPES; CARVALHO; LAHM, 2017).

A Mineração de Dados como ferramenta de classificação de pacientes de fisioterapia, utilizando da metodologia na qual foi selecionado um subconjunto de dados, referentes a prontuários disponíveis em uma clínica de fisioterapia, sendo extraídos três grandes grupos-alvo de tarefas da Mineração de Dados: associação, classificação e agrupamento, explicitados no texto. E como resultado do experimento foram extraídos padrões a partir dos dados, de tal forma que se permitisse ao leitor entender passo a passo o processo, ampliando sua compreensão dos resultados obtidos. Foram descobertos padrões em diversos formatos, os quais evidenciaram as possíveis relações entre as variáveis disponíveis. Em seguida, não apenas os padrões foram discutidos, mas, também, a importância da qualidade dos dados coletados. Concluindo que as etapas de classificação, descoberta de regras de associação e agrupamento dos dados oportunizou melhor entendimento das especialidades de pacientes atendidos pela clínica em questão,

ampliando, assim, o conhecimento do profissional na identificação das condutas a serem adotadas (CARVALHO *et al.*, 2012).

A Mineração de Dados utilizada como ferramenta de descoberta de conhecimento a partir de dados de promoção à saúde foi o tema do artigo cujo objetivo foi buscar descobrir conhecimento com base nos dados de paciente armazenados em um sistema de informação de uma operadora de planos de saúde. A partir da informação minerada, profissionais da área de saúde (médicos, psicólogos, enfermeiros e afins) puderam verificar como se encontra a saúde das pessoas dos Vales do Taquari e Rio Pardo ao longo do tempo, assim como, acompanhar a tendência com base nos dados do passado, para que possam ser tomadas algumas ações em relação a essa situação, sejam elas ações proativas ou reativas. Foram encontradas informações importantes a respeito da qualidade dos dados, que após as avaliações, descobriu-se que estavam muito falhos, sendo que grande parte deles preenchidos de forma incorreta e muitos nem informados, devido a não obrigatoriedade dos mesmos. Então, foi realizado um plano de ação para tratar melhor as informações do sistema atual, ou até mesmo a troca do sistema buscando um software mais confiável, com consistência de dados, obrigatoriedade nos campos em informações cruciais para a futura análise, entre outros (GREGORY; PRETTO, 2016).

Outro trabalho consistia em obter, por meio de um processo de descoberta de conhecimento pela Mineração de Dados, um modelo de classificação que pudesse ser aplicado no auxílio à triagem de risco de vida, na medicina. Entende-se que atividades de triagem têm o objetivo de identificar o risco de vida em pacientes mediante a análise de seus sinais vitais. No conjunto inicial de dados havia 21.821 instâncias. Contudo, 10.499 armazenavam dados nulos e foram descartadas. Trezentos e vinte e cinco registros continham dados discrepantes, ou seja, apresentavam valores inaceitáveis, extremamente maiores ou menores que valores considerados para humanos, e foram eliminados. Assim, mantiveram-se 10.997 instâncias no conjunto de dados utilizado como entrada no estudo. Conclui-se que a aplicação de um processo de KDD utilizando classificação com árvores de decisão a dados de triagem pode ser muito importante para o entendimento de que tipo de característica é mais determinante para cada uma das classes de risco, assim como os intervalos de valores de cada uma dessas características. Esse conhecimento seria muito difícil de ser obtido a partir somente de análises visuais e consultas simples a esses dados de triagem (MACIEL *et al.*, 2015).

Com a aplicação cada vez maior de tecnologia na saúde, alguns usos se tornaram mais evidentes, um estudo mostra o uso de big data em saúde no Brasil através de algumas perspectivas para um futuro próximo:

> O uso de *big data* tem crescido em todas as áreas da ciência nos últimos anos. Existem três áreas auspiciosas para o uso de *big data* em saúde: medicina de precisão (*precision medicine*); prontuários eletrônicos do paciente; e internet das coisas (*internet of things*). Entre as linguagens de programação mais utilizadas em *big data*, duas têm se destacado nos últimos anos: R e Python. Em relação às novas técnicas estatísticas, espera-se que técnicas de *machine learning* (principalmente as árvores de classificação e regressão), metodologias para controlar por associações espúrias (como a correção de Bonferroni e a taxa de falsas descobertas) e metodologias para a redução da dimensão dos dados (como a análise de componentes principais e o *propensity score matching*) sejam cada vez mais utilizadas. A questão da privacidade será também cada vez mais importante na análise de dados. O uso de *big data* na área da saúde trará importantes ganhos em termos de dinheiro, tempo e vidas e precisa ser ativamente defendido por cientistas de dados e epidemiologistas. (CHIAVEGATTO FILHO, 2015).

Muitas outras pesquisas ainda estão em desenvolvimento, porém uma iniciativa da empresa de tecnologia IBM já se encontra em produção, um ambiente de recuperação de informação muito maior que apenas a mineração de dados, de acordo com a empresa. Um dos braços de trabalho da empresa chama-se Watson *Health*, que quando aplicado à análise de imagens médicas, utiliza uma plataforma cognitiva que busca extrair valor de dados de imagens médicas em constante crescimento, analisando dados de pacientes, populações e pesquisas médicas estruturadas e não estruturadas que residem em silos desconectados.

A plataforma pretende organizar as informações disponíveis e apresentá-las de maneira contextualmente relevante e orientada para a probabilidade, para auxiliar os especialistas em diagnóstico, bem como tratar os médicos no consultório. Apenas em 2015 nos Estados Unidos, foram realizados cerca de 800 milhões de exames de ressonância e tomografia, esses estudos geraram aproximadamente 60 bilhões de imagens médicas. Nesse volume, cada um dos cerca de 31.000 radiologistas dos EUA teria que visualizar uma imagem a cada dois segundos de cada dia útil durante um ano inteiro, a fim de extrair informações potencialmente salvadoras de um punhado de imagens escondidas em um mar de dados. O sistema procura analisar dados de imagem, ler relatórios de diagnóstico e comparar informações clínicas com a lista de problemas dos exames e o registro de faturamento para encontrar possíveis discrepâncias dignas

de uma segunda olhada. Além disso, visa ajudar os administradores do hospital a monitorar com eficiência a qualidade dos cuidados sem treinamento especializado, e projetar programas de qualidade orientados por dados para padronizar a consistência dos cuidados em todos os locais e níveis dentro de uma empresa (INTERNATIONAL BUSINESS MACHINES, 2016).

O Watson *for Oncology* é um dos primeiros sistemas de apoio à decisão clínica de oncologia orientados pela inteligência artificial, que está sendo usado em todo o mundo para ajudar os médicos a avançar no tratamento do câncer. Este sistema se integra aos sistemas hospitalares e obtém centenas de atributos do registro de saúde eletrônico de um paciente, incluindo notas de médicos e relatórios de laboratório, e analisando-os com a tecnologia de processamento de linguagem natural (PNL). Em seguida, fornece aos médicos opções de tratamento de confiança e evidências de apoio para ajudá-los a tomar decisões de tratamento para seus pacientes. A ingestão de mais de 600 mil peças de evidências médicas pelo Watson, mais de dois milhões de páginas de revistas médicas e a capacidade adicional de pesquisar até 1,5 milhão de registros de pacientes para obter mais informações dão a ele uma amplitude de conhecimento que nenhum médico humano pode igualar. A taxa de diagnóstico bem-sucedido para o câncer de pulmão é de 90%, comparado a 50% para os médicos humanos (INTERNATIONAL BUSINESS MACHINES, 2017).

É humanamente impossível acompanhar a proliferação diária de dados de saúde. É necessário criar um ecossistema conectado em todo o setor de saúde para aproveitar o conhecimento dessas informações e determinar o seu valor compartilhado. É estimada uma explosão de 2.310 *exabytes* de dados de saúde projetados até 2020, que em 1 semana de internação pode ser igual a centenas de páginas em registros eletrônicos de saúde, $47 trilhões de dólares são estimados pelo impacto econômico global das doenças crônicas até 2030, em pesquisa e desenvolvimento, apenas 10% das drogas atualmente em desenvolvimento chegam ao mercado, dessa forma, é demostrada que existe uma demanda muito grande em formas de se analisar esse volume de dados que cresce e já existem pesquisas e produtos em funcionamento nessa área (INTERNATIONAL BUSINESS MACHINES, 2018).

Em continuidade com a pesquisa, para aquisição de dados para testes de importação para a ferramenta de Mineração de Dados R, foi acessado o sitio do Portal Brasileiro de Dados Abertos, mantido por Brasil (2018b), na Internet, que é um local

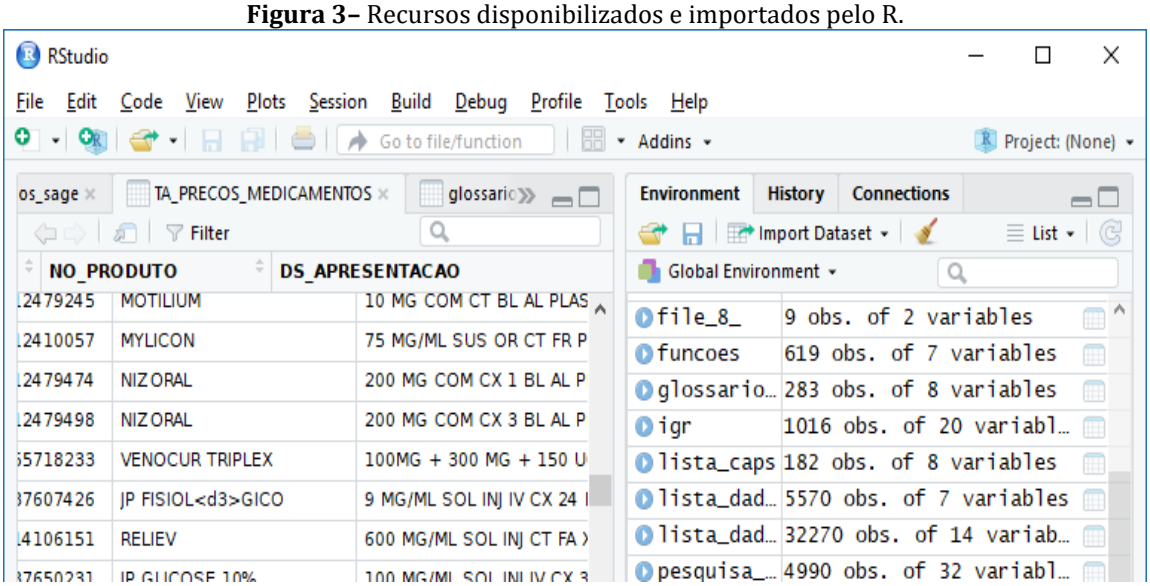
mantido pelo Governo Federal para que todos os cidadãos possam encontrar e utilizar dados e informações públicas, além disso, este portal segue princípios internacionais e deve estar de acordo com a Lei de Acesso a Informação Pública (Lei 12.527/2011), sancionada em 18 de novembro de 2011, e em seguida iniciar uma busca por dados de saúde. Em sua tela inicial são mostradas algumas informações, como por exemplo, um campo de pesquisa no qual são adicionados os termos a serem buscados, algumas publicações mais recentes e notícias ligadas ao tema.

Na data desta pesquisa, setembro de 2018, na tela inicial havia uma informação na qual o sistema mostra que é possível pesquisar em 5.972 conjuntos de dados com 30.576 recursos, porém estes números crescem diariamente, são os dados catalogados do portal e formam as fontes de dados que são retornadas a partir das buscas. Prosseguindo com o levantamento das informações por meio do sistema de busca integrado, ao utilizar o termo "saúde" como palavra de busca, obtém-se um número de 262 conjuntos de dados encontrados. Cada conjunto de dados resgatado a partir da busca é mostrado nas telas seguintes, cada um deles possui uma pequena descrição a respeito do conteúdo e elementos visuais identificando os formatos de dados em que são disponibilizados.

A metodologia usada neste trabalho para servir de métrica da capacidade de aproveitamento dos dados disponibilizados a partir de informações oriundas de estabelecimentos de saúde e prontuários de pacientes, e permitir uma grande interação com a possibilidade de resultados promissores, a partir da importação dos dados disponibilizados no sítio, por meio da ferramenta R, utilizando o ambiente de desenvolvimento chamado RStudio.

Para efetuar o procedimento de importação, foram escolhidos diversos recursos resultantes da busca do termo "saúde", neste momento foi possível ler uma pequena descrição desta fonte de dados, e ao entrar nelas é possível obter mais informações a seu respeito e escolher qual o formato a ser utilizado. Para estes exemplos, foi escolhido o formato CSV para ser efetuado o *download* para o computador do usuário, e posteriormente a importação pela ferramenta R, como mostrado na figura 3.

Figura 3– Recursos disponibilizados e importados pelo R.



Fonte: Elaborado pelo autor (2018).

Utilizando a ferramenta RStudio para manipular os dados na linguagem R, foi efetuada a importação com sucesso, e os dados ficaram disponíveis dentro da ferramenta, pode-se analisar apenas uma fonte de dados ou, depois de importar outras fontes, fazer um cruzamento entre diversas fontes, ou formatos, e ter a possibilidade de criar regras e extrair informações.

Milhões de registros contendo dados a respeito da área da saúde foram importados com sucesso, vide quadro 1. Os dados disponibilizados neste portal são provindos de estabelecimento de saúde que registram, em boa quantidade, os dados nos prontuários do paciente, como dados da doença, origem do paciente, procedimentos efetuados, causa da mortalidade, dados do nascimento, resultados de exames, entre muitos outros. A escolha para a utilização destes dados para este artigo foi devido ao fato que o acesso aos dados dos pacientes de qualquer estabelecimento de saúde depende de muitas regras de autorização e acesso por autorização por meio de comissão de ética, entre outros, e por isso, se fosse possível a utilização dos dados no final de sua cadeia, na disponibilização pelo governo como dado aberto, a sua utilização na origem também seria justificada e até melhor aproveitada, pois contêm mais dados e mais rapidez de acesso a todos os documentos. Neste caso, como proposta de trabalho futuro é conseguir o acesso a dados médicos, aplicando as mesmas ferramentas e verificando os resultados dos dados resgatados e do conhecimento descoberto. (Quadro 1).

Quadro 1– Quantidade de registros importados para o R

Recurso	Quantidade registros
Beneficiários Identificados SUS/ABI	5.341.053
Procedimentos Hospitalares por UF (São Paulo)	362.403
Exames Ambulatoriais e de Internação	174.870
Atendimentos e Consultas (07/2017)	172.324
Preços de Medicamentos	51.321
Aqui Tem Farmácia Popular	32.270
Taxa de câmbio - Livre - Dólar americano (compra)	8.466
Agentes Comunitários de Saúde	5.570
Postos de trabalho médicos no setor privado por mil habitantes	5.566
Projetos e Pesquisas	4.990
Cirurgias	4.862
Cobertura do SAMU	3.760

Fonte: Elaborado pelo autor (2018).

Não faz parte desse artigo uma análise do conteúdo importado nem mesmo como fazer a mineração pela linguagem R, apenas se por meio da importação, de dados provindos do portal de dados abertos a respeito de saúde poderiam ser utilizados para a Mineração de Dados. Foram efetuadas diversas importações com sucesso para a ferramenta, e todos os seus dados ficaram disponíveis em diversas abas na parte superior do programa, como mostrado na figura 3, permitindo neste momento serem manipulados e interagirem entre si. No quadro 2 apresenta-se a importação dos dados.

Quadro 2 – Dados aproveitados de forma nativa pela ferramenta R.

Formato	Todo o Portal	Termo "saúde"
TXT	64	6
CSV	5.076	175
XLS	222	8
JSON	4.157	74
XML	398	26
HTML	3.976	140
Total aproveitado	13.893	429
Total de dados	19.747	822
Aproveitamento %	**70,35%**	**52,18%**

Fonte: Elaborado pelo autor (2018).

Uma informação importante sobre a possibilidade de importação dos dados, no quadro 2, é que 52,18% dos conjuntos de dados fornecidos pelo portal dos dados

abertos foram retornados a partir da busca pelo termo “saúde” e 70,35% de todos os conjuntos de dados do portal podem ser importados pela ferramenta R de maneira nativa, sem a necessidade de outros aplicativos, de uma forma bem simples, permitindo a qualquer usuário com um pouco de conhecimento da metodologia da Mineração de Dados pode começar suas pesquisas. Fato este positivo, devido ao fato de existirem muitos formatos de distribuição pelo portal, podendo gerar dúvidas sobre a taxa de aproveitamento dos dados.

5 CONSIDERAÇÕES FINAIS

A conclusão é que a Mineração de Dados funciona bem sobre dados estruturados, neste caso, sobre dados provenientes do portal de dados abertos cuja origem, em muitos casos, foi de prontuários de paciente no momento do atendimento.

A existência de ferramentas livres e gratuitas já estão à disposição para utilização , mas também existem muitos investimentos nessa área, inclusive pelas gigantes de tecnologia, como exemplo o IBM Watson, na qual a maior parte dos esforços de desenvolvimento se concentram no estudo da Oncologia e Imagens Médicas, e nos próximos anos muitas outras ferramentas na área de inteligência artificial e aprendizado de máquina estarão disponíveis para aplicações na área da saúde e nos prontuários dos pacientes.

A utilização da ferramenta R na importação dos dados foi bastante positiva, pois segundo a pesquisa, quando os dados são importados para a ferramenta, um “mundo” de possibilidades torna-se possível. De acordo com informações levantadas e citadas anteriormente, essa ferramenta como função de Mineração de Dados, está ganhando mais espaço entre todas as áreas, como as da computação e da estatística, pois permite a manipulação de grandes volumes de dados e a criação de relatórios, dados visuais, gráficos e dados estatísticos de forma automática, se os dados forem corretamente, disponibilizados e atualizados.

Conclui-se que a Mineração de Dados como ferramenta de descoberta de conhecimento oriundo de informações provenientes de sistemas de saúde é possível, e até mesmo já encontra-se sendo pesquisada e aplicada, tanto por pesquisadores, quanto por grandes empresas de tecnologia que, dentre muitas áreas, já têm investimentos na

saúde. A aplicação da ferramenta R como plataforma de Mineração de Dados é uma possibilidade devido a facilidade de encontrar o programa, ser gratuito e muito poderoso, mas como descrito anteriormente, o processo da Mineração de Dados não depende apenas do seu programa ou aplicativo, mas sim de profundo conhecimento do usuário a respeito dos dados e seu fluxo, tempo e conhecimento a respeito das necessidades informacionais, do mais, se aplicado corretamente e com as devidas permissões de acesso às informações a prontuários do paciente em estabelecimentos de saúde, pode trazer grandes avanços e permitir acesso aos médicos e gestores à um conhecimento que sempre esteve presente nos documentos, porém faltava a tecnologia que pudesse cruzar todos esses dados em tempo hábil e com a devida qualidade e segurança.

REFERÊNCIAS

AMARAL, F. **Aprenda Mineração de Dados**: teoria e pratica. Rio de Janeiro: Alta Books, 2016a.

AMARAL, F. **Introdução à Ciência de Dados**: mineração de dados e big data. Rio de Janeiro: Alta Books, 2016b.

ANANIADOU, S.; KELL, D. B.; TSUJII, J. Text mining and its potential applications in systems biology. **Biotechnology**, [*S.l.*], v. 24, n. 12, p. 571-579, 2006. Disponível em: <https://bit.ly/2RaQ1Y1>. Acesso em: 07 set. 2018.

BENTES PINTO, V. Prontuário eletrônico do paciente: documento técnico de informação e comunicação do domínio da saúde. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 11, n. 21, 2006**.** Disponível em:<https://bit.ly/2QgqD5V>. Acesso em: 03 ago. 2018.

BRAGA, L. P. V. **Introdução à Mineração de Dados**: 2. ed. rev. ampl. Rio de Janeiro: E-papers Serviços Editoriais, 2005.

BRASIL. **Resolução CFM nº 1638/02**, de 10 de julho de 2002. Brasília: Diário Oficial da União; 09 de agosto de 2002.

BRASIL. Conselho Federal de Medicina. Sociedade Brasileira de Informática em Saúde. **Manual de certificação para Sistemas de Registro Eletrônico em Saúde** (S-RES) versão 4.2. Brasil, 2016.

BRASIL. Instituto Brasileiro de Geogradia e Estatística. **Projeção da população do Brasil e das Unidades da Federação,** 2018a. Disponível em: <https://bit.ly/2EIT9bH>. Acesso em: 12 set. 2018.

BRASIL. **Portal brasileiro de dados abertos**. Brasília, 2018b. Disponível em: <http://dados.gov.br>. Acesso em: 31 ago. 2018.

CARVALHO, D. R. et al. Mineração de Dados aplicada à fisioterapia**. Fisioterapia Mov.**, Rio de Janeiro, v. 25, n. 3, p. 595-605, jul., 2012. Disponível em: <https://bit.ly/2Qjf2mL>. Acesso em: 26 ago. 2018.

CASTRO, L. N.; FERRARI, D. G. **Introdução à Mineração de Dados**. São Paulo: Saraiva, 2016.

CHIAVEGATTO FILHO, A. D. P. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. **Epidemiologia e Serviços de Saúde**, Brasília, v. 24, n. 2, p. 325-332, jun., 2015.

BRASIL. Conselho Federal de Medicina. **PORTAL MÉDICO.** 2018. Disponível em: <http://portal.cfm.org.br>. Acesso em: 12 set. 2018.

FALCÃO, A. E. J. *et al*. InDeCS: método automatizado de classificação de páginas Web de Saúde usando mineração de texto e Descritores em Ciências da Saúde (DeCS). **Journal of Health Informatics**, São Paulo, v. 1, n. 1, p. 1-6, jul. 2009. Disponível em: <https://bit.ly/2PS6Vyd>. Acesso em: 17 jun. 2018.

FERNEDA, E. **Introdução aos Modelos Computacionais de Recuperação de Informação**. Rio de Janeiro: Ciência Moderna, 2012.

GALVÃO, M. C. B.; RICARTE, I. L. M. **Prontuário do Paciente**. Rio de Janeiro: Guanabara Koogan, 2012.

GALVÃO, M. C. B.; RICARTE, I. L. M.. **Prontuário eletrônico do paciente.** 2015. 203 slides. Disponível em: <https://bit.ly/2PR8A6V>. Acesso em: 10 jul. 2018.

GREGORY, G.; PRETTO, F. Mineração de Dados para Descoberta de Conhecimento em Dados de Promoção à Saúde**. Revista Destaques Acadêmicos,** [*S.l.*], v. 8, n. 4, p. 51-65, 2016.

HOBBS, G. R. Data mining and healthcare informatics. **American journal of health behavior**, [*S.l.*], v. 25, n. 3, p. 285-289, 2001.

INTERNATIONAL BUSINESS MACHINES. **IBM Watson Health**. 2016. Disponível em: <https://ibm.co/2R9sBCj>. Acesso em: 14 maio. 2018.

INTERNATIONAL BUSINESS MACHINES. **Watson Health**. 2018. Disponível em: <https://ibm.co/2bHwY2s>. Acesso em: 10 set. 2018.

INTERNATIONAL BUSINESS MACHINES. **Watson and Cancer**: Get the Facts. 2017. Disponível em: <https://ibm.co/2wbxuzK>. Acesso em: 19 set. 2017.

INTERNATIONAL DATA CORPORATION. **The Digital Universe of Opportunities**: Rich Data and the Increasing Value of the Internet of Things. 2014. Disponível em: <http://idcdocserv.com/1678>. Acesso em: 15 ago. 2018.

ISOTANI, S.; BITTENCOURT, I. **Dados Abertos Conectados**. São Paulo: Novatec Editora, 2015. Disponível em: <https://bit.ly/2L5Wctr>. Acesso em: 10 jun. 2018.

LOPES, V. J.; CARVALHO, D. R.; LAHM, J. V. KDD na Avaliação da Usabilidade do Prontuário Eletrônico do Paciente por Profissionais da Enfermagem**. Revista Brasileira de Inovação Tecnológica em Saúde**, Rio Grande do Norte, v. 6, n. 3, p. 20-31, 2017.

MACIEL, T. V. *et al*. Mineração de Dados em triagem de risco de saúde. **Revista Brasileira de Computação Aplicada**, [*S.l.*], v. 7, n. 2, p. 26-40, 2015.

MOOERS, C. N. Zatocoding applied to mechanical organization of knowledge. **American documentation**, [*S.l*], v. 2, n. 1, p. 20-32, 1951.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Evaluación de los programas de salud**: normas fundamentales para su aplicación en el procesode gestion para el desarollo nacional de la salud. Genebra, 1981. p. 49.

PENG, R. D. **R Programming for Data Science**. Online: Leanpub, 2015. Disponível em: <https://bit.ly/1JIouGj>. Acesso em: 03 ago. 2018.

QUILICI-GONZALEZ, J. A.; ZAMPIROLLI, F. A**. Sistemas Inteligentes e Mineração de**

Dados. Santo André: Triunfa Gráfica e Editora, 2015.

R CORE TEAM. **Comprehensive R Archive Network**. R Data Import/Export. 2015. Disponível em: <https://bit.ly/2KtBc0j>. Acesso em: 10 ago. 2018.

RSTUDIO. **Take control of your R code**. 2018. Disponível em: <https://bit.ly/1VLmlgA>. Acesso em: 13 ago. 2018.

SCHEFFER, M. *et al.*, **Demografia Médica no Brasil 2015**. Departamento de Medicina Preventiva, Faculdade de Medicina da USP. Conselho Regional de Medicina do Estado de São Paulo. Conselho Federal de Medicina. São Paulo, 2015, 284 p.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de Dados:** com Aplicações em R. São Paulo: Elsevier, 2016.

SONG, M. **Opinion**: Text Mining in the Clinic. The Scientist. Midland, abr. 2013. Opinion. Disponível em: <https://bit.ly/2FEPYTd>. Acesso em: 14 jun. 2018.

SPASIĆ, I. *et al*. Text mining and ontologies in biomedicine: Making sense of raw text. **Briefings in Bioinformatics**, Oxford: University Press, v. 6, n. 3, p. 239-251, 2005.

ZWEIGENBAUM, P. et al. Frontiers of biomedical text mining: current progress**. Briefings in Bioinformatics**, Oxford University, v. 8, n. 5, p. 358-375, 2007.

NOTAS

[i] A revisão ortográfica, gramatical e em Língua Portuguesa é de responsabilidade do autor.