UNIVERSIDADE FEEVALE

DIEISON MEDINGER

APLICAÇÃO DE DATA MINING PARA EXTRAIR PADRÕES EM DATASETS DE PACIENTES COM CÂNCER.

Anteprojeto de Trabalho de Conclusão

Novo Hamburgo, agosto de 2016.

DIEISON MEDINGER

APLICAÇÃO DE DATA MINING PARA EXTRAIR PADRÕES EM DATASETS DE PACIENTES COM CÂNCER.

Anteprojeto de Trabalho de Conclusão de Curso, apresentado como requisito parcial à obtenção do grau de Bacharel em Sistemas de Informação pela Universidade Feevale

Orientador: Juliano Varella de Carvalho

Novo Hamburgo, agosto de 2016.

RESUMO

Câncer é uma das doenças que mais intriga cientistas e pesquisadores por não haver uma cura 100% eficaz e nem uma única causa comprovada para todos os tipos de câncer, podendo variar de caso para caso. Estimativas para 2016 e 2017 apontam 596.070 casos novos de câncer, sendo 49% em mulheres e 51% em homens, reforçando a magnitude do problema no país. Estudos e pesquisas ligadas a esta doença vêm ganhando mais espaço com o passar dos anos, não só na esfera nacional, mas também mundial. Em todo o mundo existem entidades e ONGs (Organizações não governamentais) que buscam fomentar discussões, estudos e promover fóruns sobre o tema. Assim como na Europa há o *Cancer Research UK*, nos Estados Unidos da America o *American Cancer Society* que incentivam e promovem pesquisas sobre a doença. No Brasil há o INCA (Instituto Nacional de Câncer) que além de investimentos para área de pesquisa disponibiliza uma base de dados sobre pacientes, com informações de idade, sexo, estado, município de nascimento e residência, bem como onde o paciente foi diagnosticado com câncer, tipo de tumor, histórico de alcoolismo e tabagismo, entre outros. Esta base mantém registros entre os anos de 1985 e 2015. Em um primeiro momento, para profissionais da área de tecnologia da informação, esta base de dados pode ser de grande valia, demonstrando um potencial enorme para exploração de padrões através de técnicas de *data mining* ou até mesmo um ponto de partida para comparação dos casos ali contidos com bases de institutos de outros países. Porém, quando esta base de dados, de forma crua, é apresentada a um profissional da área da saúde, ele não tem a mesma capacidade e conhecimento técnico para explorar todo o potencial destes dados como um profissional de TI. Pensando nesta dificuldade que um pesquisador não ligado a área da tecnologia venha a ter para interpretar e utilizar de forma eficiente estas informações disponibilizadas é que propõe-se o desenvolvimento deste trabalho. A fim de aplicar técnicas de mineração de dados sobre esta base, para então buscar novos padrões, agregando assim novas perspectivas aos dados já existentes. Em um primeiro passo busca-se entender os algoritmos existentes para mineração de dados, avaliando a melhor opção para então iniciar experimentos utilizando a base do INCA. No intuito de facilitar a compreensão destes dados, da base existente e dos novos dados gerados após a mineração, recursos e bibliotecas disponíveis na linguagem de programação R serão utilizados para criar visualizações gráficas dinâmicas e estáticas. Buscando facilitar a disponibilização destes dados, todo o conteúdo gerado nas fases anteriores será disponibilizado de forma online, em um servidor R rodando uma aplicação Shiny onde alguns profissionais da área serão convidados a testar essa nova ferramenta e deixar *feedback* quanto a facilidade de uso e relevância desta nova solução para a comunidade de pesquisadores.

Palavras-chave: *Data Mining*. Câncer. Visualização de dados. Mineração de Dados. Descoberta de Conhecimento.

SUMÁRIO

MOTIVAÇÃO

Câncer é um termo genérico para um vasto grupo de doenças que podem afetar qualquer parte do corpo, assim como câncer há outros termos para descrever esta doença como tumores malignos ou neoplasias. Uma característica marcante desta doença, é o crescimento descontrolado de células que além de atingirem qualquer parte do corpo, multiplicam-se além dos seus limites usuais, que podem atingir partes adjacentes do corpo e se espalharem para outros órgãos, processo conhecido como metástase e é o maior causador de mortes por câncer.

> O termo "câncer" (do latim câncer = caranguejo) é a tradução latina da palavra grega *Karkinos* (crustáceo, caranguejo). Foi usado pela primeira vez por GALENO (aproximadamente 138-201 d.C) para designar um tumor maligno da mama cuja veias superficiais apareciam inturgescidas e ramificadas, lembrando as patas de um caranguejo. O emprego da palavra então generalizou-se para indicar tumores malignos de qualquer natureza. (Teixeira *et al*., 1997, p.13)

Células com sua proliferação fora de controle, dão origem ao que é chamado de tumor ou neoplasia, uma massa compacta de células anormais continuamente em crescimento. No entanto, se essas células neoplásicas permanecem agregadas formando uma massa única, o tumor é dito como benigno. Nesse estágio, existe uma grande possibilidade de cura completa pela remoção cirúrgica da massa. Um tumor passa a ser chamado de câncer, ou tumor maligno, quando suas células adquirem a capacidade de invadir os tecidos adjacentes. Essa invasão pode implicar na capacidade de desagregação do tumor, penetração na corrente sanguínea ou nos vasos linfáticos e formação de tumores secundários em outros locais do corpo, processo chamado de metástase. (Alberts et al., 2009)

Além de ser uma doença extremamente agressiva, os números de casos e mortes é um dos fatores que mais assusta a comunidade de cientistas e pesquisadores desta área da saúde. Conforme mostra o portal da World Health Organization (2016), 8.2 milhões de pessoas morrem todos os anos pela doença, esse número representa 13% de todas as mortes no mundo. É esperado um aumento de 70% nos casos da doença nas próximas 2 décadas, sendo que atualmente já foram diagnosticados mais de 100 tipos diferentes de câncer, onde cada tipo requer único diagnóstico e tratamento.

Além dos dados mundiais alarmantes, no Brasil estas estatísticas não são diferentes, conforme estimativas divulgadas pelo INCA para 2016-2017 é esperado cerca de 420 mil casos novos de câncer dos mais variados tipos e em torno de mais 180 mil incidências de câncer de pele. Dentre os novos casos esperados, destaca-se o câncer da cavidade oral, se não

considerado os tumores de pele não melanoma, este é o sexto mais frequente em homens na região Sul e totalizando 15.490 novos casos entre homens e mulheres no Brasil.

Levando em conta os dados apresentados acima, é possível ver que este é um problema de âmbito global, muitas entidades apoiam e financiam pesquisas, estudos e debates sobre o tema. Entidades internacionais como *World Health Organization (WHO), American Cancer Society, UK Cancer Society* e outras nacionais como o Instituto Oncoguia, Fundação do Câncer e Instituto Nacional de Câncer (INCA) fornecem dados, estatísticas e até mesmo *dataset* com as mais variadas informações sobre pacientes com históricos da doença.

Como visto, informações relevantes aos pesquisadores que estudam esta doença são encontradas de várias maneiras nas mais variadas fontes e disponibilizadas em diferentes formatos. Tan, Steinbach e Kumar (2006), afirmam que muitas vezes não é possível usar técnicas tradicionais para analisar dados, mesmo em pequenos conjuntos de dados, necessitando assim desenvolver novos métodos para efetuar esta análise.

Porém, devido à natureza da formação de pesquisadores da área da saúde compilar, manipular e analisar de forma eficiente estes dados pode ser uma difícil tarefa. Pensando nesta dificuldade o atual trabalho propõe o uso da informática e suas técnicas para analisar informações coletadas sobre a doença, aplicar técnicas na busca de padrões e gerar visualizações que estimulem a criatividade e o senso de investigação dos mesmos.

Conforme descrito por Passos e Goldschmidt (2005, p. 1)

> A análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas. Portando, torna-se imprescindível o desenvolvimento de ferramentas que auxiliem o homem, de forma automática e inteligente, na tarefa de analisar, interpretar e relacionar esses dados para que se possa desenvolver e selecionar estratégias de ação em cada contexto de aplicação.

Na busca de concretizar o objetivo proposto, um primeiro passo é a coleta dos dados a serem estudados, nas mais variadas fontes, tais como as disponibilizadas nas entidades citadas acima. Para então em um momento seguinte aplicar técnicas de mineração de dados.

Coleman e Ahlemeyer-Stubbe (2014) escrevem que a habilidade de extrair conhecimento oculto em dados tem se tornado muito importante no mundo atual. Quando os dados são usados para predição e comportamentos futuros isso representa uma grande vantagem na área em que estes dados são estudados.

Buscando-se uma maneira de extrair essas informações dos dados, de maneira mais acurada e utilizando técnicas eficazes optou-se por usar a linguagem R. Além de ser uma

linguagem de uso livre e código aberto, esta apresenta bibliotecas que atendem a todas as necessidades deste trabalho, como por exemplo a biblioteca Ratlle, que prove solução para usuários carregar, transformar e explorar dados e ainda conta com uma interface gráfica. (Williams *et al*., 2016)

O trabalho pretende não apenas extrair conhecimento dos dados colhidos, mas também gerar visualizações, dinâmicas e estáticas, capazes de instigar a investigação dos pesquisadores, dando-lhes a liberdade de manipular os gráficos gerados. Para isso serão investigadas várias bibliotecas da linguagem R, tais como:

- googleVis – "Interface de R para o Google Charts API, permitindo aos usuários criar gráficos interativos baseados em *data frames*." (Gesmann, Castillo e Cheng, 2016, tradução nossa)
- ggplot2 – "um sistema de plotagem para R, com base na gramática da gráfica[...], bem como fornecendo um modelo poderoso de gráficos que faz com que seja fácil de produzir gráficos complexos multi-camadas." (Wickham, 2013, tradução nossa)
- plotly – "[...] uma versão baseada em web interativo e cria visualizações baseadas na web personalizadas diretamente do R." (Sievert et al., 2016, tradução nossa)

A pesquisa de métodos e bibliotecas para representar os dados graficamente se faz de extrema importância para o resultado final deste trabalho. Conforme Chen, Härdle e Unwin (2008) a visualização de dados é um termo novo, e expressa mais do que a representação dos dados em forma de gráfico, ela deve ajudar os leitores a ver a estrutura dos dados, assim como deduzir informações sobre estes dados em vez de apenas concentrar-se na apresentação de informações.

> Ferramentas de visualização ajudam as pessoas em situações em que vendo a estrutura do *dataset* em detalhe é melhor do que ver apenas um breve resumo do mesmo. Uma dessas situações ocorre quando explorar os dados para encontrar padrões, tanto para confirmar os padrões esperados quanto encontrar aqueles inesperados. Outra situação ocorre quando se avalia a validade de um modelo estatístico, para julgar se o modelo de fato se ajusta aos dados. (Munzner, 2014, p. 7, Tradução do Autor)

"A evolução do poder computacional têm sido de grande benefício para a geração de gráficos nos últimos anos. Tornou-se possível desenhar precisos e complexos gráficos com grande facilidade e imprimí-los com uma qualidade impressionante em alta resolução." (Chen, Härdle e Unwin, 2008, p.5, tradução nossa). Mesmo encontrando-se gráficos com as

mais variadas informações sobre o câncer na internet, poucos exploram de forma eficiente o poder da computação ou possibilitam a interação e manipulação das visualizações fornecidas ao público que acessa o conteúdo.

De forma mais pobre e rudimentar são encontradas visualizações sobre esta doença com dados e estatísticas no âmbito nacional no portal do INCA. Este portal fornece algumas visualizações onde se é possível ter alguma interação com os gráficos ali apresentados.

O gráfico abaixo foi gerado no portal mortalidade.inca.gov.br selecionando-se os parâmetros de tipo de câncer como esôfago, lábio, outras partes da língua, base da língua e gengiva para homens e mulheres da região sul nos períodos de 2004 a 2013.

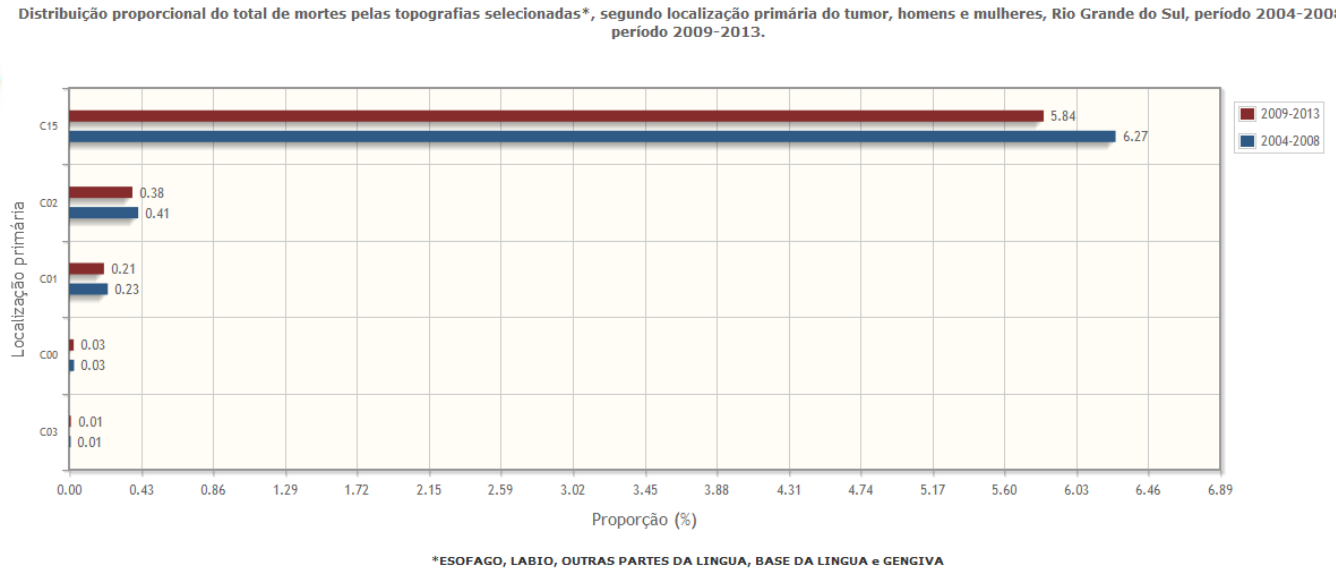

Figura 1: Distribuição proporcional do total de mortes pelas topografias selecionadas*, segundo localização primária do tumor, homens e mulheres, região Sul, período 2004-2008 e período 2009-2013.(INCA, 2016)

Preocupando-se não apenas com a exibição dos dados, faz-se necessário disponibilizar os mesmos de uma forma de fácil acesso para utilização dos profissionais da área da saúde. Para este problema a linguagem R também apresenta uma solução de fácil implementação e com recursos robustos permitindo a disponibilização das visualizações criadas de forma *online*, a biblioteca Shiny.

Shiny é um framework que "torna incrivelmente fácil de construir aplicações web interativas com R.[...] possível construir belas, responsivas, e poderosas aplicações com o mínimo esforço". (Chang *et al.*, 2016, tradução nossa).

Oferecendo uma solução de servidor que roda em linux que permite os usuários hospedar e gerenciar aplicações Shiny através da internet. Esta solução permite gerenciar

processos R sendo executados em URLs e portas diferentes com outras vantagens como hospedar várias aplicações simultaneamente.

Com os avanços ocorridos na área da computação nos últimos anos é possível analisar e exibir informações de maneira nunca antes vista. Utilizando-se destes recursos será possível, após a coleta de dados sobre câncer nas mais diferentes fontes, aplicar mineração de dados e gerar modelos gráficos das informações obtidas. A aplicação dessas técnicas com ajuda da linguagem de programação R, permitirá disponibilizar todo o conteúdo online. Assim, ao final do desenvolvimento deste trabalho, busca-se obter uma ferramenta online que, além de consolidar informação sobre incidências de câncer em uma única plataforma, ainda utilizará de recursos gráficos para exibir informações sobre a doença de maneira interativa e com possibilidade de manipular estes gráficos em tempo de execução.

OBJETIVOS

Objetivo geral

Criar um ambiente de visualizações, gráficos interativos e de fácil compreensão, a partir de *datasets* sobre câncer e padrões encontrados a partir da aplicação de *data mining*, com o intuito de auxiliar pesquisadores dessa área de estudo.

Objetivos específicos

- Investigar bases de dados abertas a respeito de casos de câncer;
- Montar diferentes visualizações dos dados obtidos;
- Preparar dados para mineração;
- Extrair conhecimento da base de dados;
- Interagir com pesquisadores da área da Saúde para que eles avaliem as visualizações e resultados gerados.

METODOLOGIA

O propósito deste trabalho é criar gráficos e visualizações de dados dinâmicas, visando auxiliar pesquisadores da área da saúde, mais especificamente pesquisadores ligados ao estudo do câncer. Para gerar tais visualizações, será feita uma pesquisa de natureza aplicada, pois será necessário investigar a linguagem R, mineração de dados e visualização de dados para então aplicá-los na prática.

Do objetivo do estudo, podemos afirmar se tratar de explicativo, pois a ideia do projeto é, através da aplicação de técnicas já conhecidas de mineração de dados, buscar padrões no *dataset* selecionado, para então criar visualizações onde será possível melhor compreender os dados ali contidos, estimulando a investigação por parte dos usuários da ferramenta, através da possibilidade de interagir de forma dinâmica com os gráficos.

Com a avaliação dos materiais existentes, como artigos, livros, trabalhos de conclusão de curso, teses de mestrado e doutorado, se dará o embasamento teórico para construção do trabalho, realizando-se assim uma pesquisa bibliográfica prévia a fim de melhor conhecer técnicas e investigar trabalhos já realizados nesse campo de pesquisa. Para finalizar este procedimento técnico, serão realizadas pesquisas experimentais com o objetivo de selecionar as bibliotecas a serem utilizadas, bem como algoritmos de mineração de dados a serem aplicados.

Após a avaliação do material disponível e finalização da pesquisa bibliográfica, se inicia o passo de mineração de dados, com o intuito de agregar mais valor à base de dados. Para que então no passo seguinte sejam geradas visualizações destes dados, com o intuito de expressar os padrões encontrados no passo anterior e também transformar os dados contidos na base em informação visual e dinâmica para melhor compreensão e interpretação do conteúdo.

Finalizado os experimentos de mineração e montagem das visualizações dos dados, todo o material será submetido a avaliação de profissionais da área de odontologia para captação de ideias e *feedback* para confecção final do ambiente, produzindo assim uma ferramenta orientada as necessidades dos profissionais da área, levando em consideração suas necessidades, ideias e expectativas.

Quanto a abordagem do trabalho, esta será qualitativa pois como expresso acima, a intenção deste trabalho não é quantificar a qualidade da base ou dos gráficos que serão

construídos, mas sim criar novas visualizações em cima de dados já disponíveis para elaborar uma nova ferramenta para utilização no estudo do câncer no Brasil.

Assim sendo, com base na metodologia acima, o trabalho propõe-se a completar o objetivo geral e todos os objetivos específicos apresentados. Além disto, também está encarregado de responder a seguinte questão de pesquisa: A partir dos *datasets* disponibilizados por órgãos governamentais, é possível consolidar dados e estatísticas sobre câncer em uma ferramenta com visualizações interativas, a fim de auxiliar pesquisadores da área da saúde?

CRONOGRAMA

Trabalho de Conclusão I

Etapa	Meses			
	Ago	Set	Out	Nov
Anteprojeto				
Pesquisa bibliográfica sobre Câncer e Mineração de dados;				
Estudar Linguagem R e suas bibliotecas;				
Estudar os métodos de *data mining;*				
Estudar técnicas de visualizações de dados;				
Verificar existência de outras bases com dados referentes a câncer bocal;				
Elaborar TC I;				

Trabalho de Conclusão II

Etapa	Meses			
	Mar	Abr	Mai	Jun
Preparar os dados para análise;				
Aplicar *Data Mining;*				
Extrair conhecimento das análises realizadas				
Montar gráficos e visualizações;				
Disponibilizar visualizações no ambiente;				
Elaborar TC II;				

BIBLIOGRAFIA

AHLEMEYER-STUBBE, Andrea; COLEMAN, Shirley. **A Practical Guide to Data Mining for Business and Industry.** United Kingdom: John Wiley & Sons Ltd., 2014. 296p.

ALBERTS, Bruce; JOHNSON, Alexander; LEWIS, Julian; RAFF, Martin; ROBERTS, Keith; WALTER, Peter. Biologia Molecular da Célula. Porto Alegre: Artmed, 2009. 1463p.

CHANG, Winston *et al*. **shiny: Web Application Framework for R**. CRAN, 2016. Disponível em < https://cran.r-project.org/web/packages/shiny/index.html >. Acesso em: 20 de Agosto de 2016.

CHEN, Chun-houh; HÄRDLE, Wolfgang; UNWIN, Antony. **Handbook of Data Visualization.** 3º ed. German: Springer-Verlag BerlinHeidelberg, 2008. 899p.

GESMANN, Markus; CASTILLO, Diego de; CHENG, Joe. **googleVis: R Interface to Google Charts.** CRAN, 2016. Disponível em: < https://cran.r-project.org/web/packages/googleVis/index.html>. Acesso em: 20 de Agosto de 2016.

INCA, **Atlas On-line de Mortalidade.** Disponível em: < https://mortalidade.inca.gov.br/MortalidadeWeb/pages/Modelo02/consultar.xhtml#panelResultado>. Acesso em: 21 de agosto de 2016.

Ministério da Saúde. **Estimativa/2016 Incidência de Câncer no Brasil.** Rio de Janeiro, 2015. 121p.

MUNZNER, Tamara. **Visualization Analysis & Design.** New York: CRC Press, 2014. 375p.
GOLDSCHMIDT, Ronaldo; EMMANUEL, Passos. **Data Minining Um Guia Pratico.** Rio de Janeiro: Elsevier Editora Ltda., 2005. 253p.

SIEVERT, Carson *et al.* **plotly: Create Interactive Web Graphics via 'plotly.js'** CRAN, 2016. Disponível em < https://cran.r-project.org/web/packages/plotly/index.html >. Acesso em: 20 de Agosto de 2016.

TAN, Pang – Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao DATAMINING Mineração de Dados.** Rio de Janeiro: Editora Ciência Moderna Ltda., 2009. 900p.

TEIXEIRA, Antonio C. B. *et al* **Câncer de Boca Noções Básicas para Prevenção e diagnóstico.** São Paulo: Editora Fundação Peirópolis, 1997. 86p.

WICKHAM, Hadley. **ggplot2.** ggplot2.org. 2013. Disponível em: < http://ggplot2.org/>. Acesso em: 20 de Agosto de 2016.

WILLIAMS, Graham *et al.* rattle: **Graphical User Interface for Data Mining in R.** CRAN, 2016. Disponível em: < https://cran.r-project.org/web/packages/rattle/index.html>. Acesso em: 20 de Agosto de 2016.

World Health Organization, **Cancer.** Disponível em: <http://www.who.int/cancer/en/>. Acesso em: 21 de agosto de 2016.