
# UM AMBIENTE PARA RECOMENDAÇÃO DE TÉCNICAS DE VISUALIZAÇÃO DE INFORMAÇÃO

Fernanda Cristina Ribeiro

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Jano Moreira de Souza

Rio de Janeiro

Setembro de 2016

UM AMBIENTE PARA RECOMENDAÇÃO DE TÉCNICAS DE VISUALIZAÇÃO DE INFORMAÇÃO

Fernanda Cristina Ribeiro

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____

Profº. Jano Moreira de Souza, Ph.D.

_____

Prof.ª Melise Maria Veiga de Paula, D.Sc.

_____

Profº. Cláudio Esperança, Ph.D.

_____

Profº. Marcos Roberto da Silva Borges, Ph.D.

RIO DE JANEIRO, RJ - BRASIL

SETEMBRO DE 2016

Ribeiro, Fernanda Cristina

Um Ambiente para Recomendação de Técnicas de Visualização de Informação/ Fernanda Cristina Ribeiro. – Rio de Janeiro: UFRJ/COPPE, 2016.

XIII, 105 p.: il.; 29,7 cm.

Orientador: Jano Moreira de Souza

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2016.

Referências Bibliográficas: p. 92-97.

1. Recomendação de Visualização. 2. Visualização de Informação. 3. Sistema de Recomendação. I. Souza, Jano Moreira de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

# Agradecimentos

Sou grata a Deus por ter me sustentado e me abençoado nesses anos, especialmente, durante o período do mestrado. Em 2013, deixei minha família e amigos para iniciar uma nova etapa. Foi desafiador e cresci muito como pessoa e profissionalmente também. Tenho certeza que não terminaria essa etapa senão fosse a vontade de Deus.

Agradeço a Deus pela família que me deu. O apoio, as conversas e as orações do meu pai (Evaristo), da minha mãe (Dinha), do meu irmão (Filipe) e da minha irmã (Juliana) foram fundamentais nessa etapa. Estive ausente em muitos momentos que gostamos de passar em família. Mas isso nos fortaleceu e nos fez valorizar ainda mais os momentos que estamos juntos. Família linda e querida, obrigada pela compreensão e apoio!!! Também agradeço a minha cunhada (Luciana) pelo apoio e orações. Aos demais familiares e amigos, obrigada.

Agradeço ao professor Jano por me orientar durante o mestrado. Pelas ideias, críticas e sugestões que surgiram durantes nossas conversas. Também agradeço o apoio financeiro com a revisão de artigos em inglês. Agradeço a professora Melise por aceitar a me coorientar neste trabalho. Durante nossas conversas também surgiram ideias, críticas e sugestões que me ajudaram na elaboração desta dissertação. Tenho aprendido muito com meu orientador e coorientadora, são pessoas especiais e espero continuar convivendo e aprendendo com eles.

Agradeço aos professores Cláudio Esperança e Marcos Borges por aceitarem participar da banca examinadora.

Desde 2011 trabalho em projetos de pesquisa e desenvolvimento no Centro de Apoio a Políticas de Governo (CAPGov/COPPE). Agradeço ao Sérgio, ao professor Jano e a professora Melise pela oportunidade de participar desses projetos. A toda equipe do CAPGov, obrigada. Com vocês tenho aprendido a trabalhar melhor em equipe.

Agradeço a professora Isabela Drummond (da Unifei) por esclarecer minhas dúvidas em relação à área de aprendizado de máquina. O Gustavo Lima e o Luan Garrido também contribuíram nesse assunto, obrigada! Agradeço ao Luiz Felipe por me ajudar na elaboração da interface da ferramenta. Agradeço ao Bernardo por avaliar a ferramenta do primeiro estudo de caso e sugerir melhorias de usabilidade. Bárbara, obrigada pelas caronas neste último ano de mestrado.

Agradeço a Patrícia, Ana Paula, Eliah, Solange, Guty, Mercedes, Cláudia e Ricardo por esclarecerem minhas dúvidas e por me ajudarem com as documentações do mestrado e do CAPGov. A Lourdes pelo café e a Rosa e a Déda por cuidarem da limpeza dos laboratórios, obrigada.

Agradeço ao CNPq pela ajuda financeira com a bolsa do mestrado.

A todos, fica aqui registrado: “Muito obrigada!!!”.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UM AMBIENTE PARA RECOMENDAÇÃO DE TÉCNICAS DE VISUALIZAÇÃO DE INFORMAÇÃO

Fernanda Cristina Ribeiro

Setembro/2016

Orientador: Jano Moreira de Souza

Programa: Engenharia de Sistemas e Computação

O crescimento e a disseminação das tecnologias de informação e comunicação geram uma quantidade de informações cada vez maior e mais complexa para o usuário. Neste cenário, as técnicas de Visualização de Informação podem ser importantes ferramentas para auxiliar o usuário na análise e utilização dos dados. No entanto, atualmente, é notável a diversidade de possibilidades para gerar representações visuais. Essa grande variedade de técnicas de visualização dificulta a escolha da técnica mais apropriada para a representação dos dados. A proposta desta dissertação é construir uma ferramenta que auxilie o usuário na seleção e criação de visualizações através da recomendação. Para construir a ferramenta, primeiro foi realizado um estudo para elaborar um conjunto de regras para recomendação de técnicas de visualização de informação considerando as características dos dados e a tarefa que o usuário pretende realizar a partir da visualização. Em seguida, este conjunto foi refinado através da aplicação de técnicas de aprendizado de máquina. O algoritmo C4.5 (QUINLAN, 1993) (um algoritmo de árvore de decisão) obteve o melhor desempenho na classificação das regras e foi escolhido como o algoritmo da ferramenta de recomendação. Por fim, a ferramenta foi avaliada através de um estudo de caso com usuários. Os resultados mostraram que a abordagem de recomendação (através da classificação das técnicas de visualização) pode auxiliar os usuários na criação do gráfico mais adequado para seus dados e para a tarefa que deseja realizar.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

AN ENVIRONMENT FOR RECOMMENDATION OF INFORMATION VISUALIZATION TECHNIQUES

Fernanda Cristina Ribeiro

September/2016

Advisor: Jano Moreira de Souza

Department: Systems and Computer Engineering.

The growth and spread of information and communication technologies generate an increasingly large and more complex amount of information to the user. In this scenario, information visualization techniques can be important tools in assisting users in analysis and use of data. Currently, there is a great variety of information visualization techniques, which complicates the process of choosing the most appropriate technique to represent data. The purpose of this dissertation is to build a tool that assists the user in selecting and creating visualizations through recommendation. To build the tool, first a study was conducted to develop a set of rules for the recommendation of information visualization techniques, by considering the characteristics of the data sets and the tasks the user intends to accomplish from visualization. Then this set was refined by applying machine learning techniques. The C4.5 algorithm (QUINLAN, 1993) (a decision tree algorithm) achieved the best performance in the classification of rules and it was chosen as the algorithm recommendation tool. Finally, the tool was evaluated through a case study with users. The results show that the recommendation approach (by classification of visualization techniques) can assist users in creating the most appropriate chart for your data and for the task you want to accomplish.

# Sumário

# Figuras

# Tabelas

# Lista de abreviaturas, siglas e símbolos

BDVR – *Behavior-Driven Visualization Recommendation*

CSV – *Comma-Separated Values*

IHC – Interface Humano-Computador

InfoVis – *Information Visualization*

kNN – *k Nearest Neighbor*

MLP – *Multilayer Perceptron*

PDF – *Portable Document Format* (Formato Portátil de Documento)

PNG – *Portable Network Graphics* (formato de imagem)

RVis – Recomendação de Visualização

TXT – Extensão de arquivo texto

XLS(X) – Extensão de arquivos do Excel

# 1. Introdução

## 1.1 Contextualização

Atualmente, a possibilidade de criação e acesso à informação pode resultar em oportunidades interessantes para empresas, governos e outras classes de usuários divulgarem seus dados e melhorar seus processos. Segundo DIX (2013), as técnicas de Visualização de Informação podem ser utilizadas como solução nestes contextos. O avanço da tecnologia nesta área tem gerado diversas possibilidades para utilização destas técnicas de maneira que a publicação de uma determinada informação (um dos requisitos essenciais de diferentes classes de sistemas) seja adequadamente atendida.

Desta forma, fica estabelecido um cenário onde a escolha da estratégia mais apropriada para a análise e divulgação de um determinado dado pode representar um problema. Estudos têm sido realizados com o objetivo de descobrir quais aspectos da visualização influenciam na escolha da técnica. Alguns desses aspectos são: cores, atributos visuais, contexto e o conjunto de técnicas de visualização (TUFTE, 2001), (FEW, 2012), (VOIGT *et al.*, 2012), (BORKIN *et al.*, 2013).

Além disso, atualmente, é notável a diversidade de possibilidades para gerar representações visuais. Por exemplo, RIBECCA (2016) catalogou 54 [1] técnicas de visualização e, em AIGNER *et al.* (2011), foram apresentadas 101 técnicas de visualização de dados temporais. Essa grande variedade de técnicas de visualização dificulta a escolha da

[1] Até o dia 18 de agosto de 2016 tinham sido catalogados 54 de um total de 60 técnicas de visualização.

estratégia mais adequada para representar um determinado conjunto de dados e auxiliar o usuário na execução de uma tarefa.

Na literatura, diversos autores discutem diferentes abordagens para facilitar a escolha de uma determinada técnica de visualização. A classificação das técnicas é uma estratégia utilizada por RIBECCA (2016), KEIM (2002), LENGLER & EPPLER (2007a), FEW (2012) e SHNEIDERMAN (1996).

Os sistemas de recomendação também são apresentados como alternativa para facilitar a escolha de técnicas de visualização. Exemplos de aplicações que disponibilizam a recomendação como recurso para facilitar a escolha de uma determinada técnica de visualização são: *Microsoft Excel*, *Tableau Public* (TABLEAU, 2016), ViSC (DE SOUSA & BARBOSA, 2013), *Exploration Views* (ELIAS & BEZERIANOS, 2011), *Watson Analytics* (IBM, 2015a), VizAssist (BOUALI *et al.*, 2015) e Voyager (WONGSUPHASAWAT *et al.*, 2016).

A proposta deste trabalho foi elaborar uma ferramenta que auxilie o usuário na seleção e criação de visualizações através da recomendação. O mecanismo definido para a elaboração da recomendação foi baseado na classificação das técnicas de visualização considerando a tarefa que o usuário deseja realizar e as características do conjunto de dados.

O diferencial deste trabalho quando comparado às demais iniciativas encontradas está relacionado à estratégia utilizada para a recomendação, mais especificamente, a utilização de um modelo de classificação que combina elementos importantes deste contexto que são as tarefas e as características dos dados.

## 1.2 Objetivos

Portanto, os objetivos deste trabalho são:

1. Definir um conjunto de regras para a classificação e recomendação de técnicas de visualização de informação.
2. Aplicar técnicas de aprendizado de máquina, mais especificamente de modelos de classificação, para otimizar a representação da regra de maneira que a recomendação automática seja a mais próxima possível da recomendação humana.
3. Elaborar uma ferramenta para apoiar a seleção e criação de técnicas de visualização de informação considerando os aspectos que influenciam essa escolha, como a tarefa e as características do conjunto de dados. A recomendação foi realizada através do modelo de classificação que obteve o melhor desempenho na otimização do conjunto de regras.

## 1.3 Organização do Trabalho

Este trabalho está organizado em seis capítulos. Este é o primeiro capítulo que contém a Introdução. No segundo capítulo é apresentada a fundamentação teórica com uma breve revisão da literatura acerca dos temas necessários para o entendimento deste trabalho: Visualização de Informação, Recomendação de Técnicas de Visualização de Informação e Aprendizado de Máquina. No terceiro capítulo é apresentada a proposta desta dissertação, detalhando os passos para elaboração e avaliação do conjunto de regras. Ainda no terceiro capítulo é descrita a evolução e o refinamento do conjunto de regras com a aplicação de técnicas de aprendizado de máquina. A ferramenta desenvolvida RVis (Recomendação de Visualização) é descrita no quarto capítulo. O quinto capítulo apresenta a metodologia utilizada na avaliação da ferramenta, bem como os resultados obtidos. No sexto capítulo

estão as conclusões, considerando as contribuições, as limitações do trabalho e os direcionamentos para trabalhos futuros.

# 2. Fundamentação Teórica

Este capítulo visa descrever os principais conceitos necessários para o entendimento da proposta deste trabalho. São descritos conceitos referentes às áreas de visualização de informação, recomendação de técnicas de visualização e aprendizado de máquina.

## 2.1 Visualização

De acordo com o dicionário online Dicio (DICIO, 2016), visualizar significa "transformar conceitos abstratos em imagens mentalmente visíveis" ou formar uma imagem mental. Nesta definição, estão incluídas quaisquer representações feitas através de imagens podendo ser geradas até mesmo por alucinações. Já para WARE (2004), visualização pode se referir tanto ao uso de diagramas para transmissão de um sentido quanto ao processamento e transformação de dados em formas gráficas.

Ainda de acordo com WARE (2004), os significados de visualização remetem a um meio para auxiliar o entendimento ou a cognição seja colocando em evidência características de um fenômeno ou exibindo resultados de operações com informações. Essa similaridade aponta o principal objetivo da visualização que é aprimorar o sistema cognitivo explorando o sentido da visão e possibilitando uma maior transmissão de informações (SHNEIDERMAN, 1996), (WARE, 2004).

A visualização pode ser classificada com base na natureza dos dados manipulados (DO NASCIMENTO & FERREIRA, 2005):

- **Visualização Científica:** representa dados científicos, ou seja, dados que geralmente correspondem a objetos físicos, fenômenos da natureza ou posições em domínio espacial e possuem uma representação correspondente no mundo

físico (DO NASCIMENTO & FERREIRA, 2005). Por exemplo, visualização de órgãos do corpo humano.

- **Visualização de Informação:** representa dados abstratos para os quais, geralmente, não existe uma representação visual correspondente no mundo físico (VALIATI, 2008). Segundo DO NASCIMENTO & FERREIRA (2005), neste caso, a representação visual pode ser gerada com base nos relacionamentos e informações inferidos acerca dos dados abstratos.

Neste trabalho, será utilizado o conceito de Visualização de Informação como esse apresentado por VALIATI (2008), não sendo consideradas as visualizações científicas.

### 2.1.1 Visualização de Informação

Especificando mais o conceito, a Visualização de Informação pode ser considerada uma técnica de visualização que auxilia no processo de cognição, mas está ligada intrinsecamente com a computação (WARE, 2004). CARD & MACKINLAY (1997) e SHNEIDERMAN (1996) afirmam que o interesse por esta área de estudo foi ampliado devido ao aumento da disponibilidade de computadores, da capacidade de processamento gráfico e da quantidade de dados.

Na literatura, é possível encontrar outras definições para o termo Visualização de Informação. Por exemplo, para CARD *et al.* (1999) Visualização de Informação é o uso de representações visuais e interativas dos dados suportadas por computador para ampliar a cognição. Em outro trabalho CARD (2003) afirma que a promessa da Visualização de Informação é acelerar nosso entendimento e nossas ações em um mundo de volumes crescentes de informações. Segundo KEIM *et al.* (2006), Visualização de Informação é a comunicação de dados abstratos através do uso de interfaces visuais interativas. De acordo

com PURCHASE *et al.* (2008), a Visualização de Informação utiliza computação gráfica e recursos interativos para auxiliar os seres humanos na resolução de problemas.

Nestas definições, é possível identificar aspectos importantes do termo, como: o uso do computador e o uso de recursos visuais e interativos. O objetivo principal, no entanto, é ampliar a capacidade cognitiva do visualizador para a tomada de decisão. Desta forma, a visualização não pode ser considerada estritamente como uma representação gráfica ou o fim de um processo, a visualização deve ser entendida como um meio de se chegar a um fim ou uma ferramenta catalizadora do processo de criação do conhecimento e da tomada de decisão (RIBEIRO *et al.*, 2016).

A escolha da técnica de visualização é uma etapa importante no processo de elaboração de representações visuais aplicadas a uma determinada situação (FREITAS *et al.*, 2001). Para FEW (2012), o processo de seleção e construção de um gráfico pode ser abordado seguindo seis etapas fundamentais: (1) determinar qual mensagem será transmitida e identificar os dados necessários para comunicar essa mensagem, (2) determinar se será necessário uma tabela, um gráfico ou uma combinação de ambos para comunicar a mensagem, (3) determinar o melhor meio para codificar os valores, (4), determinar onde exibir cada variável no gráfico, (5) determinar o melhor design para os dados (6) determinar se um dado em particular deve ser destacado dos demais e se for, determinar como será o destaque.

Esse processo de seleção e construção dos gráficos está relacionado à etapa de mapeamento visual do modelo de referência de visualização proposto por CARD *et al.* (1999). Nesse modelo, conforme mostra a Figura 1, os dados abstratos passam por um processo de transformação tendo como resultado a representação visual (CARD *et al.*, 1999). Ao todo são três transformações: (1) transformação dos dados, (2) mapeamento visual e (3) transformação visual. Na primeira transformação, os dados brutos são transformados e

armazenados em tabelas. Em seguida ocorre o mapeamento visual, ou seja, essa etapa está relacionada à escolha de representações ou estruturas visuais dos dados. Por fim, ocorrem as transformações visuais que consistem nas possíveis interações que o usuário realiza sobre a visualização (CARD *et al.*, 1999).

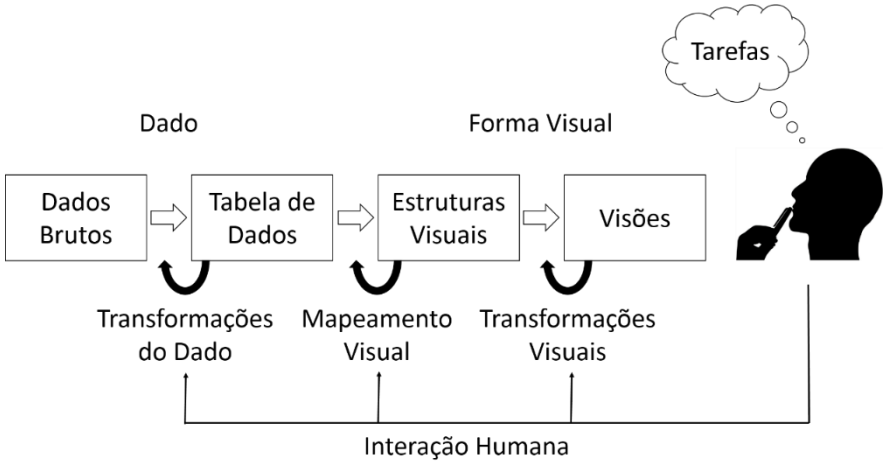


Figura 1. Modelo de Visualização proposto por CARD *et al.* (1999). Adaptado de CARD *et al.* (1999).

De acordo com LUZZARDI (2003) e FREITAS *et al.* (2001), alguns autores propõem classificações (ou taxonomias) para auxiliar os usuários na escolha das visualizações. Por exemplo, SHNEIDERMAN (1996) criou o Mantra da Visualização (*Visual Information Seeking Mantra: overview first, zoom and filter, then details-on-demand*) que define os princípios considerados fundamentais no processo de construção de visualizações. Esse processo inicia com a visão geral dos dados e continua com a possibilidade do usuário focar e filtrar as informações para, só então, buscar detalhes conforme necessário (YAMAGUCHI, 2010). Com base nesses critérios, SHNEIDERMAN (1996) criou uma taxonomia que classifica as técnicas de visualização de acordo com as tarefas realizadas pelo usuário e os tipos de dados manipulados na representação visual. Em relação às tarefas, as classes definidas foram (SHNEIDERMAN, 1996), (YAMAGUCHI, 2010):

- **Visão geral**: fornece uma visão geral da coleção de dados.

- ***Zooming***: foco nos itens de interesse do usuário.
- **Filtragem**: descarta os itens que não são interessantes ao usuário.
- **Detalhes sob demanda**: seleção de um item ou de um grupo e obtenção de detalhes quando necessário.
- **Relações**: visualiza o relacionamento entre os itens.
- **Histórico**: mantém um histórico de ações para dar apoio às ações de desfazer e refazer.
- **Extração**: permite a extração de subconjuntos da coleção de dados e parâmetros de consulta.

Em relação aos tipos de dados manipulados, SHNEIDERMAN (1996) classificou as técnicas em: unidimensional, bidimensional, tridimensional, temporal, multidimensional, hierárquico e representado por grafos.

Na taxonomia de KEIM (2002), as técnicas de visualização são classificadas com base em três critérios:

- **Tipo de dado a ser visualizado:** unidimensional, bidimensional, multidimensional, texto/hipertexto, hierarquias/grafos e algoritmos/softwares.
- **Técnica de visualização:** gráficos convencionais (para representação de um a três atributos), técnicas geométricas, icnográficas, orientadas a pixel e técnicas baseadas em dimensões.
- **Técnica de distorção e interação utilizada:** projeção, filtragem, zoom, distorção e ligação e seleção (*link & brush*).

LENGLER & EPPLER (2007a) organizaram as técnicas de visualização de acordo com as áreas de aplicação: visualização de dados, visualização de informação, visualização conceitual, visualização estratégica, visualização metafórica e visualização composta.

Conforme mostra a Figura 2, essa classificação foi apresentada em uma tabela periódica, onde cada elemento da tabela é uma técnica de visualização. As linhas representam a complexidade da visualização que é referente ao número de regras aplicadas para utilização e / ou ao número de interdependência dos elementos a ser visualizado. Quanto mais à direita uma visualização, maior é a sua complexidade. As colunas representam às áreas de aplicação. A cor das letras de cada sigla ilustra o tipo de informação representada: azul para informação de processo e preto para informação estruturada (LENGLER & EPPLER, 2007a).

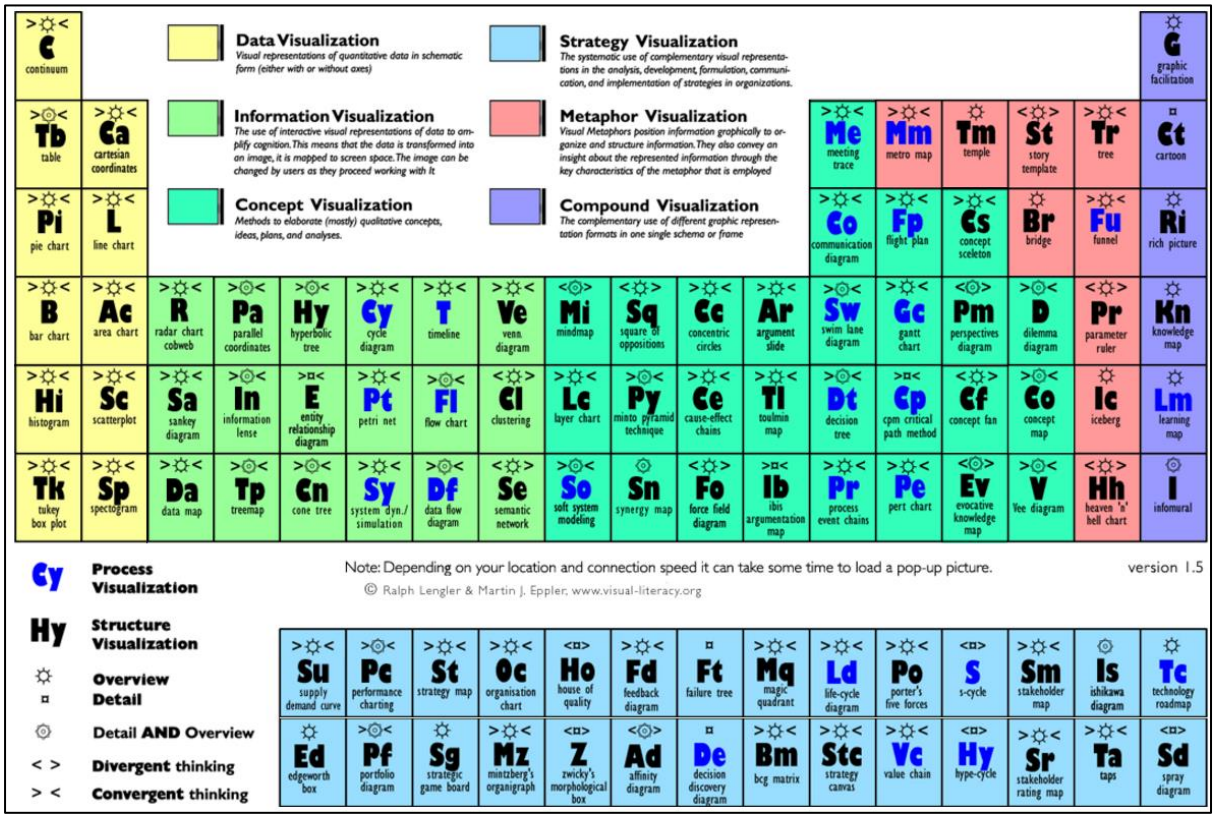


Figura 2. Tabela periódica das técnicas de visualização proposta por LENGLER & EPPLER (2007a). Retirado de (LENGLER & EPPLER, 2007b).

Conforme ilustra a Figura 2, acima da sigla de cada visualização há símbolos que representam dois aspectos da técnica: o tipo de visão dos dados e o processo cognitivo. O tipo de visão pode ser subdividido em (LENGLER & EPPLER, 2007a):

- Detalhe (🔲): a técnica de visualização destaca os itens individuais dos dados e suas características.
- Visão geral (☼): a técnica de visualização enfatiza a visão global dos dados e permite uma visão geral no primeiro contato com a visualização.
- Detalhe + visão geral (⚙): a técnica de visualização permite a visão geral dos dados e também destaca os itens individuais.

Já o processo cognitivo pode ser classificado em (LENGLER & EPPLER, 2007a):

- Pensamento convergente (> <): reduz a complexidade através de análise e síntese.
- Pensamento divergente (< >): adiciona complexidade com o objetivo de obter *insight*.

Quanto ao tipo do dado, FEW (2012) afirma que os dados podem ser classificados em dois tipos: dados quantitativos e dados categóricos. Dados quantitativos são os números e dados categóricos são os rótulos que dão significado aos números. Por exemplo, no gráfico da Figura 3, os dados categóricos estão representados pela cor verde e os dados quantitativos pela cor azul.

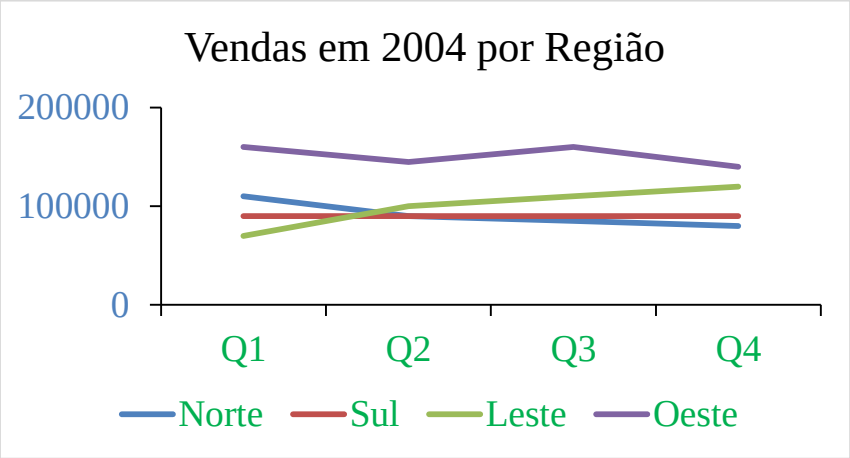


Figura 3. Representação e distinção entre dados categóricos e quantitativos no mesmo gráfico. Adaptado de FEW (2012).

O gráfico da Figura 3 contém uma escala quantitativa no eixo vertical (Y) e uma escala categórica no eixo horizontal (X). Segundo FEW (2012), quando usada em gráficos, a escala categórica pode ser de três tipos: nominais, ordinais e intervalares. Escalas nominais consistem de itens de dados discretos que não possuem relação entre si. Esses itens se diferem apenas no nome (por isso, dados nominais). Além disso, os itens da escala categórica não possuem uma ordem particular e não representam valores quantitativos (FEW, 2012). Exemplos de dados nominais: regiões (norte, sul, leste, oeste) e departamentos (vendas, marketing, finanças).

Ao contrário da escala nominal, os itens em uma escala ordinal possuem uma ordem intrínseca, porém os itens em si não representam valores quantitativos (FEW, 2012). Por exemplo, a escala Likert com 5 opções (como: concordo fortemente, concordo, neutro, discordo, discordo fortemente) e o tamanho de peça de roupa: pequeno (P), médio (M) e grande (G).

Na escala intervalar, os itens de dados têm uma ordem intrínseca e representam valores quantitativos. De acordo com FEW (2012), uma escala intervalar começa como uma escala quantitativa, mas é convertida em uma escala categórica por meio da subdivisão do intervalo completo de valores em uma série sequencial de intervalores menores de igual tamanho, cada um com seu próprio rótulo. Por exemplo, considere o intervalo de valores que aparece no gráfico da Figura 3 acima. Este intervalo, de 0 a 200.000, pode ser convertido em uma escala categórica contendo os seguintes intervalos menores: (1) > 0 e <= 50.000, (2) > 50.000 e <= 100.000, (3) > 100.000 e <= 150.000 e (4) > 150.000 e <= 200.000.

FEW (2012) ressalta que unidades de tempo tais como anos, semestres, bimestres, meses, semanas, dias e horas, mesmo que os itens correspondam a valores quantitativos (exemplo: 2014, 2015, 2016), uma escala formada por unidades de tempo é uma escala do

tipo (categórica) intervalar. Além disso, os meses (que não possuem a mesma quantidade de dias) também formam uma escala intervalar, pois para propósito de relatórios e análises considera-se que os meses possuem a mesma quantidade de dias (FEW, 2012).

## 2.2 Recomendação de Visualização

A escolha de visualizações adequadas para representação de um determinado conjunto de dados é um problema recorrente em diversos cenários. Na literatura, há diversos trabalhos que apresentam sugestões que podem orientar esta escolha, como TUFTE (2001), FEW (2012), ILIINSKY & STEELE (2011), RIBECCA (2016) e HARDIN *et al.* (2012). Outra vertente de estudo nesta área é denominada recomendação de visualização e pode ser definida como a área que estuda e desenvolve abordagens e sistemas para recomendar e/ou criar visualizações de forma automática (YANG *et al.*, 2014a) (VOIGT *et al.*, 2012). Nesta seção são apresentados trabalhos dessas duas vertentes de pesquisa.

Os trabalhos da primeira vertente de pesquisa (orientações para escolha de visualizações) foram analisados para elaboração do conjunto inicial de regras de recomendação. Um dos trabalhos pesquisados foi o de Edward Tufte que apresenta princípios ou maneiras efetivas de representar os dados visualmente (TUFTE, 2001). Segundo TUFTE (2001) todo gráfico deve:

- Exibir os dados.
- Induzir o visualizador a pensar sobre o dado e não sobre a metodologia, o design gráfico ou a tecnologia usada para gerar o gráfico.
- Evitar distorções que o dado não contém.
- Tornar coerentes conjuntos de dados volumosos.
- Encorajar o olho do visualizador a comparar pedaços dos dados.

- Revelar o dado em vários níveis de detalhes.
- Servir a um propósito claro.

FEW (2012) apresentou princípios que são aplicados, especificamente, ao processo de criação de gráficos e tabelas para representar as informações quantitativas no cenário de negócios. Segundo o autor, dois desafios fundamentais envolvem a apresentação (de forma efetiva) dos dados quantitativos. O primeiro é determinar a melhor maneira de contar a estória dos dados, ou seja, de apresentá-los ao visualizador (ou público-alvo) e o segundo desafio é projetar/desenvolver os componentes que contarão essa estória de maneira clara e o mais eficiente possível (FEW, 2012).

Para solucionar esses desafios, FEW (2012) propõe práticas de design de gráficos definidas em função de sete tipos de relacionamentos que são específicos para dados quantitativos no cenário de negócios. A Tabela 1 apresenta a definição de FEW (2012) para esses relacionamentos e as técnicas de visualização sugeridos pelo autor.

Tabela 1. Técnicas de visualização sugeridos por Few. Adaptado de FEW (2012).

| **Relacionamento** | **Definição** | **Técnicas de Visualização** |
|---|---|---|
| Comparação Nominal | O relacionamento comparação nominal refere-se à comparação entre valores de um conjunto de dados não ordenados e que não possuem um relacionamento particular entre esses valores. Como exemplo, considere o gráfico da Figura 4, a escala categórica ao longo do eixo X é nominal e as quatro regiões geográficas não se relacionam entre si em alguma ordem em particular. | Gráfico de barras<br>Gráfico de colunas<br>Gráfico de pontos |

| | | |
|---|---|---|
| Correlação | Quando pares de valores[2] quantitativos são exibidos para revelar se há um relacionamento significativo entre esses valores, esse relacionamento é denominado correlação. FEW (2012) afirma que o entendimento da correlação entre variáveis quantitativas pode auxiliar a prever ou evitar determinados comportamentos. | Gráfico de dispersão |
| Desvio | Quando os dados quantitativos são exibidos para caracterizar como um ou mais conjunto de valores diferem de um conjunto de referência de valores, esse relacionamento é conhecido por relacionamento de desvio. | Gráfico de linhas<br>Gráfico de pontos (somente quando os dados não incluem zero) |
| Distribuição | No relacionamento distribuição, os valores quantitativos estão espalhados em um intervalo. Segundo FEW (2012), o formato da distribuição de um conjunto de valores revela se há *gaps* ou concentrações nos dados. | Histograma<br>Gráfico de linhas |
| Parte-todo | Revela a porção que cada valor quantitativo representa em algum todo. FEW (2012) afirma que esse tipo de relacionamento é útil para mostrar alguma informação que está dividida em partes e para mostrar a porcentagem de cada parte no todo. | Gráfico de barras<br>Gráfico de colunas<br>Gráfico de barras empilhadas<br>Gráfico de colunas empilhadas |
| Série Temporal | Representa valores quantitativos com intervalos de tempo iguais (como anos, meses, semanas, dias e horas). FEW (2012) afirma que esse relacionamento revela tendências e padrões nos dados. | Gráfico de linhas<br>Gráfico de colunas<br>Gráfico de pontos |

[2] Com cada valor representando uma medida diferente sobre uma entidade – como pessoa, departamento ou produto.

| | | |
|---|---|---|
| Ranqueamento | Quando valores quantitativos são apresentados em sequência de acordo com o tamanho, de grande a pequeno e vice-versa, este relacionamento é conhecido como ranqueamento. Segundo FEW (2012), esse relacionamento revela a sequência dos dados e facilita a comparação entre os valores com base no posicionamento dos elementos que representam os valores quantitativos. | Gráfico de barras<br>Gráfico de colunas |

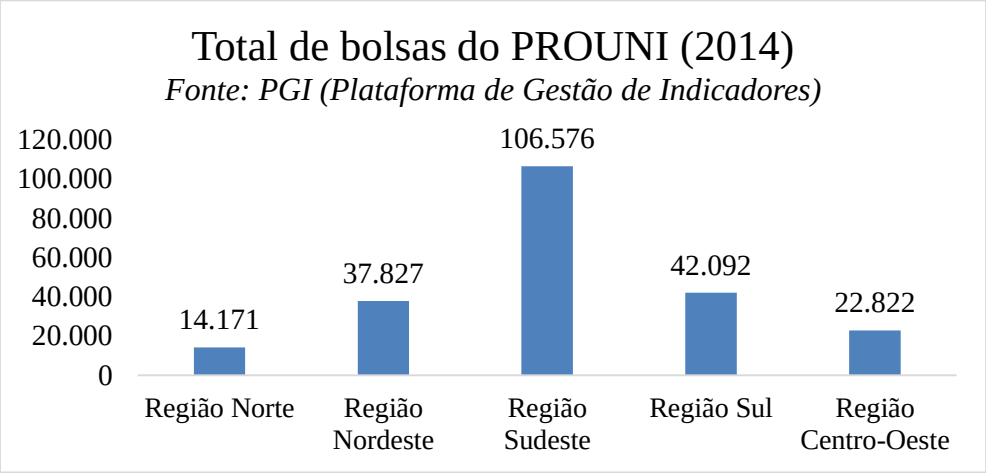


Figura 4. Exemplo de gráfico para o relacionamento comparação nominal.

Ainda em FEW (2012), o autor sugere que o gráfico de pizza não é adequado para exibir dados com relacionamento parte-todo, apesar dessa técnica de visualização ser frequentemente utilizado para esse propósito. Segundo o autor, os gráficos mais indicados são colunas e barras, pois comparar o comprimento das barras/colunas é mais fácil do que comparar o tamanho das fatias (área) (FEW, 2012).

Outro aspecto discutido por FEW (2012) se refere ao espaço disponível para apresentação do gráfico que pode representar um problema quando é necessário representar uma considerável quantidade de variáveis. Por exemplo, se um conjunto de dados possui três

variáveis a serem exibidas em um gráfico, duas variáveis podem ser codificadas nos eixos X e Y e a terceira pode ser representada por múltiplas linhas ou por conjuntos de barras ou de pontos que são codificadas em diferentes maneiras (por exemplo, através do uso de diferentes cores). Contudo, caso seja necessário representar uma quarta variável, qual seria a melhor solução?

Uma das possíveis soluções seria aumentar a dimensão da visualização. Em um gráfico 2D, por exemplo, poderia ser adicionando o eixo Z. Contudo, devem ser feitas algumas considerações. Um gráfico 3D pode impor uma sobrecarga à análise que prejudicaria a exploração do dado. Em FEW (2012), por exemplo, o autor não recomenda o uso dessa prática, pois a terceira dimensão pode dificultar a leitura. Segundo o mesmo autor, uma possível solução seria usar o conceito de "múltiplos pequenos" gráficos (do inglês "*small multiples*") proposto por TUFTE (2001). Essa solução envolve uma série de gráficos menores, todos dispostos de forma que possam ser vistos simultaneamente (TUFTE, 2001), conforme mostra a Figura 5.

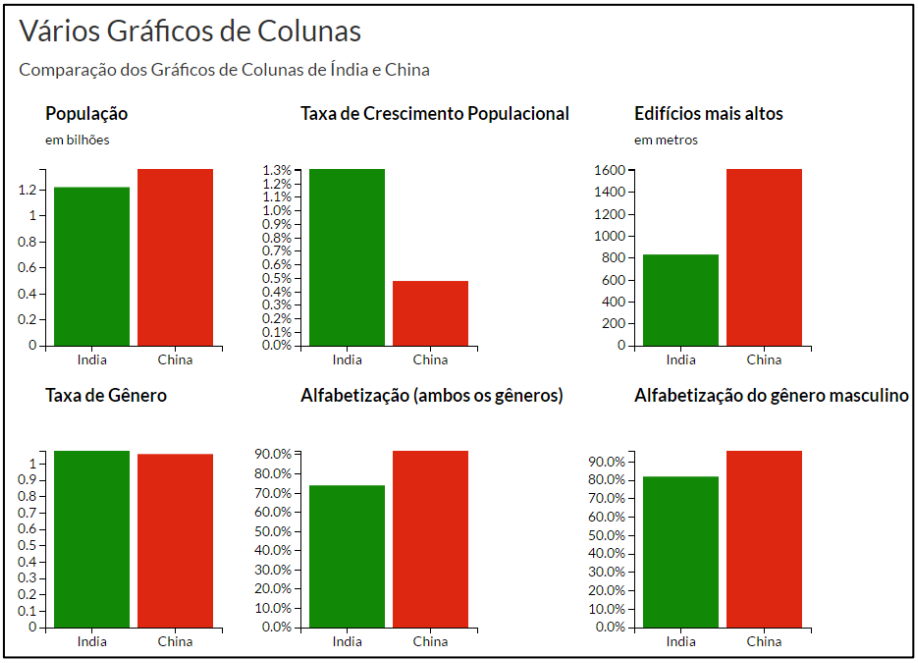


Figura 5. Exemplo de "*small multiples*". Retirado de DO (2015).

Segundo ILIINSKY & STEELE (2011), para criar visualizações são necessários três "ingredientes": o projetista, o leitor e os dados. O projetista é quem cria a visualização com algum propósito, o leitor é o público-alvo da visualização e os dados são codificados na visualização para revelar aspectos interessantes. ILIINSKY & STEELE (2011) afirmam que dados diferentes requerem abordagens e técnicas diferentes para comunicar ao leitor o objetivo do projetista. A Tabela 2 apresenta algumas visualizações que, segundo ILIINSKY & STEELE (2011), podem ser usadas na codificação dos dados.

Tabela 2. Técnicas de visualização sugeridos por Iliinsky e Steele. Adaptado de ILIINSKY & STEELE (2011).

| **Técnica de Visualização** | **Recomendação** |
|---|---|
| Gráfico de barras | Comparação entre e inter categorias, ideal para dados discretos. |
| Gráfico de linhas | Mostra a tendência dos dados, ideal para dados contínuos. |
| Histograma | Mostra a distribuição de valores em um intervalo possível. |
| Gráfico de áreas | Ideal para valores que se acumulam. |
| Gráfico de pizza | Comparação de frações de um todo. Recomendado quando há poucas frações relevantes. |
| Gráfico de barras empilhadas | Alternativa ao gráfico de pizza quando há muitas frações. |
| Gráfico de dispersão | Mostra a correlação entre duas variáveis quantitativas. Também é recomendado para exibir os dados que variam ao longo de duas dimensões. |

RIBECCA (2016) disponibilizou na web um catálogo com 60 técnicas de visualização[3]. Nesse catálogo, o autor apresenta as seguintes informações: descrição da técnica de visualização; funções (ou tarefas) relacionadas à técnica; exemplo de variações dessa técnica

[3] Até o dia 18 de agosto de 2016 tinham sido catalogados 54 gráficos.

(por exemplo, o gráfico de área é apresentado como uma variação do gráfico de linhas[4]); links para as principais ferramentas e softwares usados para gerar visualizações da técnica descrito e links de outros sites que criaram exemplos de visualização da técnica descrita. Em relação às funções (ou tarefas), RIBECCA (2016) apresenta alguns tipos de tarefas que o usuário pode realizar com a visualização, conforme mostra a Tabela 3.

Tabela 3. Técnicas de visualização sugeridos por Ribecca. Adaptado de RIBECCA (2016).

| **Tarefa** | **Técnicas de Visualização** |
|---|---|
| Comparar as diferenças ou similaridades entre os valores | Gráfico de barras, gráfico de bolhas, histograma, gráfico de linhas, pirâmide populacional, gráfico de radar, gráfico de pizza, treemap e diagrama de Venn. |
| Mostrar as diferenças e similaridades entre os valores ou em relação ao todo usando o tamanho ou a área | Gráfico de bolhas, gráfico de colunas empilhadas, gráfico de área proporcional, nuvem de palavras, gráfico de pizza, gráfico de rosca, gráfico de colunas empilhadas e treemap. |
| Mostrar o relacionamento ou conexões entre os dados. | Gráfico de barras, gráfico de bolhas, gráfico de linhas, gráfico de dispersão, gráfico de áreas empilhadas, gráfico de colunas empilhadas e diagrama de Venn. |
| Visualizar a hierarquia do conjunto de dados | Treemap e diagrama de árvore. |
| Mostrar parte (ou partes) de uma variável em relação ao seu total. | Gráfico de pizza, gráfico de rosca, gráfico de colunas empilhadas e treemap. |
| Visualizar como os dados estão distribuídos ao longo de um intervalo ou como estão agrupados. | Gráfico de bolhas, histograma, gráfico de colunas agrupadas, linha do tempo, mapa (distribuição exibida geograficamente), pirâmide populacional e nuvem de palavras (para visualizar a distribuição das palavras no corpo do texto). |

[4] Veja o exemplo em http://www.datavizcatalogue.com/methods/line_graph.html

| | |
|---|---|
| Exibir as variações entre os limites superior e inferior em uma escala. | Histograma, gráfico de caixa (*box plot*). |
| Analisar os dados ao longo de um período de tempo. | Gráfico de linhas, histograma, linha do tempo, gráfico de bolhas, gráfico de áreas, série temporal em espiral (*spiral plot*) |

HARDIN *et al.* (2012) também apresentam recomendações para criar visualizações. Essas recomendações são apresentadas em função do tipo de dado que está sendo analisado e das questões que o usuário quer responder com a visualização. Por exemplo, HARDIN *et al.* (2012) afirmam que o gráfico de barras é eficiente para dados numéricos divididos em categorias diferentes. Segundo os autores, esse gráfico auxilia o usuário a comparar os dados entre as categorias. Em relação ao gráfico de pizza, HARDIN *et al.* (2012) recomendam que esse gráfico tenha no máximo seis fatias, pois é mais difícil interpretar esse gráfico com muitas fatias. Se há mais de seis proporções a serem comunicadas, os autores recomendam o gráfico de barras. Além das recomendações, os autores apresentam exemplos de conjuntos de dados que podem ser representados pelas técnicas de visualização consideradas. A Tabela 4 sumariza os gráficos recomendados por HARDIN *et al.* (2012).

Tabela 4. Gráficos recomendados por HARDIN *et al.* (2012).

| **Técnica de Visualização** | **Quando usar?** |
|---|---|
| Gráfico de Barras | Para comparar dados entre categorias. |
| Múltiplos Gráficos de Barras | Para comparar informações relacionadas. |
| Gráfico de Barras Empilhadas | Para exibir dados relacionados (um em cima do outro). |
| Gráfico de Barras Agrupadas | Para exibir dados relacionados (um ao lado do outro). |
| Gráfico de Linhas | Para visualizar tendências nos dados ao longo do tempo. |
| Gráfico de Áreas | Para mostrar valores que se acumulam verticalmente. |
| Gráfico de Pizza | Para mostrar proporções. |

| | |
|---|---|
| Gráfico de Dispersão | Para investigar a relação entre variáveis diferentes. |
| Gráfico de Bolhas | Para mostrar a concentração dos dados. |
| Histograma | Para entender a distribuição dos dados. |
| Treemap | Para mostrar dados hierárquicos como uma proporção de um todo. |
| Gráfico de caixa (*box plot*) | Para mostrar a distribuição de um conjunto de dados. |

Já em 2009, GOTZ & WEN (2009) classificaram em três categorias os sistemas que construíam ou recomendavam automaticamente visualizações. Essas categorias eram (GOTZ & WEN, 2009), (FREYNE & SMYTH, 2010): (1) baseados em tarefa, (2) baseados nas propriedades dos dados e (3) sistemas híbridos.

Na primeira categoria, os autores citam a ferramenta BOZ (CASNER, 1991) proposta em 1991. Essa ferramenta usava a descrição da tarefa do usuário como entrada para o algoritmo que constrói e recomenda a visualização mais apropriada (GOTZ & WEN, 2009). Sistemas baseados nas propriedades dos dados utilizam as características do conjunto de dados como entrada para o algoritmo de recomendação. Exemplo de sistemas dessa segunda categoria: APT (MACKINLAY, 1986), SAGE (ROTH *et al.*, 1994), VizDeck (KEY *et al.*, 2012), Tableau Public (TABLEAU, 2016) e *Microsoft Excel*. Por fim, os sistemas híbridos combinam as propriedades dos dados e a tarefa do usuário para recomendar visualizações. Exemplo de sistemas dessa categoria são: *Exploration Views* (ELIAS & BEZERIANOS, 2011), ViSC (DE SOUSA & BARBOSA, 2013), *Watson Analytics* (IBM, 2015a), HARVEST (GOTZ & WEN, 2009) e IMPROVISE+ (ZHOU & CHEN, 2003).

Também existem estudos relacionados à área de recomendação que pesquisam as preferências dos usuários na criação de visualizações. Por exemplo GOTZ & WEN (2009) apresentaram a abordagem BDVR (*Behavior-Driven Visualization Recomendation*) que monitora o comportamento do usuário para encontrar padrões de interação do usuário com a

visualização. Com base nesses padrões, o algoritmo BDVR infere a tarefa que o usuário deseja realizar e então sugere automaticamente visualizações que dão suporte a essa tarefa.

Nessa linha de pesquisa, a análise do comportamento do usuário, YANG *et al.* (2014) realizaram estudos empíricos para entender os vários aspectos de design da visualização e o impacto desses aspectos nas preferências dos usuários. Com os resultados desses estudos, YANG *et al.* (2014) pretendem construir um sistema de visualização automática que auxilie os usuários a combinar dois ou mais gráficos para formar uma visualização integrada.

### 2.2.1 Trabalhos relacionados

Nesta subseção são apresentados os trabalhos e sistemas da área de recomendação de visualização relacionados à proposta desta dissertação. Exemplos de sistemas que recomendam gráficos: Microsoft Excel, Tableau (TABLEAU, 2016), Exploration Views (ELIAS & BEZERIANOS, 2011), ViSC (DE SOUSA & BARBOSA, 2013), Many Eyes (VIEGAS *et al.*, 2007), Watson Analytics (IBM, 2015a), VizAssist (BOUALI *et al.*, 2015) e Voyager (WONGSUPHASAWAT *et al.*, 2016).

O Excel é um editor de planilhas da Microsoft[5], na versão 2013, foi adicionada a funcionalidade *Recommended Charts* cujo objetivo é recomendar gráficos (JELEN, 2013). De acordo com JELEN (2013), o Excel aplica regras e heurísticas para sugerir os gráficos. Essas regras e heurísticas são aplicadas sobre os dados selecionados na planilha, ou seja, a recomendação é realizada apenas com base no tipo de dado selecionado (JELEN, 2013).

Já o *software* Tableau Public[6] (versão 10) utiliza a funcionalidade *Show Me* para recomendar os gráficos (TABLEAU, 2016), (MACKINLAY *et al.*, 2007). Essa

[5] https://products.office.com/pt-br/excel
[6] https://public.tableau.com/s/download

funcionalidade utiliza uma linguagem de especificação algébrica chamada VizQL (*a language for query, analysis and visualization*) para automaticamente apresentar os dados como uma tabela de visualizações (comumente chamada de pequenas múltiplas visões) (MACKINLAY *et al.*, 2007). Essa tabela de visualizações é a paleta *Show Me*. A linguagem de especificação VizQL foi originalmente usada no desenvolvimento do sistema Polaris (STOLTE *et al.*, 2002). VizQL é uma linguagem formal e declarativa usada para realizar consultas em banco de dados e apresentar o resultado através de tabelas, gráficos e mapas, ou seja, o resultado é expresso de uma maneira visual (HANRAHAN, 2006).

Os usuários do Tableau especificam expressões VizQL através do recurso interativo de arrastar as instâncias dos campos apresentados no painel de Dados. Esse painel de Dados é organizado em campos relacionados à dimensão (que são normalmente campos categóricos) e campos relacionados às medidas (que são normalmente campos numéricos) (MACKINLAY *et al.*, 2007). Segundo MACKINLAY *et al.* (2007), a expressão VizQL é formada pelos campos categóricos e numéricos arrastados pelo usuário para as áreas específicas na ferramenta. A funcionalidade *Show Me* recomenda os gráficos com base nos campos arrastados (TABLEAU SOFTWARE, 2016).

Tanto o Excel quanto o Tableau fazem a recomendação dos gráficos com base nas propriedades dos dados. As ferramentas apresentadas a seguir, utilizam outros aspectos (além das propriedades dos dados) para recomendar os gráficos. Essas ferramentas são: Exploration Views, ViSC e IBM Watson Analytics, Voyager.

Exploration Views é um sistema para criação e customização de vários gráficos em um ambiente de *dashboard* (ELIAS & BEZERIANOS, 2011). Esse sistema foi projetado para entender como usuários novatos na área de visualização constroem gráficos e *dashboards*. Durante o processo de criação dos gráficos, após o usuário selecionar um ou mais atributos

dos dados e selecionar as categorias, o Exploration Views apresenta os tipos de análise mais apropriados para os dados do usuário (ELIAS & BEZERIANOS, 2011). Por exemplo, para um conjunto de dados sobre vendas por estado no período de 2 anos, os tipos de análise possíveis incluem: comparação (vendas por estado ou por ano), contribuição (a porcentagem de vendas por estado sobre todas as vendas) e tendências (a evolução das vendas ao longo do tempo). Logo, os tipos de análise correspondem às tarefas.

Com base no tipo de análise que o usuário escolhe e nas características do conjunto de dados, o sistema Exploration Views faz a recomendação dos gráficos aos usuários. Segundo ELIAS & BEZERIANOS (2011), os gráficos sugeridos são apresentados na ordem do mais comumente usados aos mais complexos, permitindo que os usuários inexperientes explorem outras alternativas de visualização.

A ferramenta ViSC (*Visualization with Smart Charts*) também foi projetada para usuários inexperientes. Essa ferramenta apoia a construção de visualizações de dados estatísticos através de recomendações de visualização baseadas em perguntas mais comuns que os usuários costumam formular durante a construção e leitura de gráficos. Desta forma, os gráficos são construídos por meio de um diálogo da ferramenta com o usuário que deve selecionar a pergunta a ser respondida (DE SOUSA & BARBOSA, 2013).

Uma limitação da ferramenta ViSC é que o usuário não consegue fazer upload dos seus dados. Ao utilizar a ferramenta, primeiro o usuário seleciona um dos temas (conjuntos de dados) disponíveis, define as variáveis dos eixos e escolhe os dados que deseja visualizar. Em

seguida a ferramenta faz a recomendação[7]. A recomendação da ferramenta ViSC é definida a partir de uma ontologia com cinco classes de nível superior (DE SOUSA, 2013):

- **Dado**: possui dimensões dadas pelo número de variáveis. Cada variável pode ser classificada quanto à estrutura (escalar, vetorial e sensorial), quanto à natureza (discreta ou contínua), quanto ao tipo (nominal, ordinal, intervalar e razão) e quanto ao tipo de componente (temporal, geográfico ou nulo).
- **Atributos de exibição**: essa classe baseia-se nas propriedades das marcas (posicional, de retina e temporal) e se relaciona com a classe Dado através de quatro tipos de percepção (associativa, seletiva, ordenada e quantitativa).
- **Visualização**: essa classe é composta pelas representações atômicas (como séries, colunas, dispersão, barras) e pelas visualizações que agregam novos atributos às representações atômicas (como séries múltiplas, barras agregadas, colunas empilhadas). Ao todo são oito técnicas de visualização suportadas pela ferramenta ViSC: colunas agregadas, colunas múltiplas, colunas empilhadas, séries, séries múltiplas, séries empilhadas, dispersão e tabela.
- **Tarefa**: é o objetivo do usuário com a visualização e o que deve ser feito com a visualização. Segundo DE SOUSA (2013), a tarefa tem relação com a pergunta que o usuário quer responder e com a eficiência da representação visual. As tarefas e as perguntas da ferramenta ViSC foram definidas com base na taxonomia de AMAR *et al.* (2005). As perguntas são geradas automaticamente na ferramenta ViSC com base em *templates* armazenados em um banco de

[7] Este vídeo (http://taissasousa.com/visctutorial.mp4) mostra que o usuário primeiro seleciona um dos temas disponíveis e as variáveis e os dados que deseja visualizar para em seguida a ferramenta fazer a recomendação.

dados. Os parâmetros desses *templates* para criar as perguntas são: o dado selecionado e as características desse dado.

- **Transformação**: essa classe define quais variáveis podem ir para cada atributo de exibição, de modo que a visualização seja eficiente para a tarefa ou pergunta.

Outro sistema que também utiliza o recurso de diálogo no processo de construção de visualizações é a aplicação web IBM Watson Analytics[8]. Essa aplicação foi criada como uma alternativa ao Many Eyes, que foi descontinuado conforme comunicado no site da aplicação (IBM, 2015b).

Em 2010, quando o Many Eyes ainda estava em uso, FREYNE & SMYTH (2010) aplicaram técnicas de raciocínio baseado em casos para produzir recomendações de visualizações e auxiliar os usuários da ferramenta na seleção da visualização mais apropriada. Essa recomendação baseada em casos funcionava da seguinte maneira: quando um novo conjunto de dados era selecionado, o sistema convertia o conjunto de dados em um conjunto de características e usava essas características para encontrar um conjunto de casos similares. As visualizações associadas com os casos similares eram ranqueadas e retornadas aos usuários como um conjunto de recomendações (FREYNE & SMYTH, 2010).

A ferramenta Watson Analytics oferece três tipos de serviços para análise dos dados: *Explore*, *Predict* e *Assemble* (IBM, 2015a). O serviço *Explore* permite que o usuário faça uma análise exploratória sobre o conjunto de dados e descubra padrões e relacionamentos nesses dados. O serviço *Predict* utiliza algoritmos para realizar predições e descobrir ideias, padrões e correlações no conjunto de dados do usuário. As predições auxiliam na identificação de potenciais associações entre os campos do conjunto de dados que podem ser

[8] http://www.ibm.com/analytics/watson-analytics/

úteis na análise visual dos dados. Com o serviço *Assemble*, o usuário pode criar e compartilhar painéis com os *insights* (ideias) obtidos com os serviços *Explore* e *Predict* (IBM, 2015a).

O serviço que está relacionado a este trabalho é o Explorer. O fluxo de trabalho desse serviço é: (1) o usuário carrega os dados, (2) são apresentadas possíveis perguntas que o usuário possa querer responder com a visualização que será criada e (3) o usuário escolhe a pergunta (ou cria uma nova) e (4) por fim, a visualização é renderizada na tela. Cada pergunta tem uma ou mais visualizações associadas e recomendadas (IBM, 2015a). As técnicas de visualização disponíveis no serviço Explorer são: gráfico de colunas/barras, gráfico de colunas/barras agrupadas, gráfico colunas/barras empilhadas, gráfico de linhas, gráfico de área, gráfico de pizza, mapa geográfico, treemap, gráfico de bolhas, mapa de calor, gráfico de barras radial, tabela, visualização categórica (é uma técnica de visualização que agrupa os dados e usa formatos, como quadrado, pontos, para exibir os dados. Também é utilizado o recurso de saturação de cores).

Na ferramenta VizAssist as visualizações são criadas com base em três critérios: (1) nos dados do usuário, (2) em algumas informações sobre os dados (como a importância de cada atributo) e (3) nos objetivos do usuário (BOUALI *et al.*, 2015). Após coletar essas informações, a ferramenta apresenta uma lista de visualizações candidatas e o usuário escolhe uma ou mais visualizações para realizar a tarefa de exploração do conjunto de dados no contexto de mineração visual de dados. Para cada visualização selecionada o usuário pode fornecer um *feedback* com a finalidade de melhorar o mapeamento visual dos dados. A ferramenta utiliza um algoritmo genético para fazer as recomendações (BOUALI *et al.*, 2015).

A ferramenta Voyager auxilia a exploração em profundidade (*breadth-oriented exploration*) de conjuntos de dados, ou seja, o usuário tem acesso a uma galeria de visualizações para realizar a tarefa de análise exploratória e conhecer todo o conjunto de dados (WONGSUPHASAWAT *et al.*, 2016). Essa ferramenta utiliza a *engine* de recomendação *Compass* que adotou uma abordagem similar a VizQL (usada na ferramenta Tableau) para enumerar, agrupar e ordenar as visualizações com base nas propriedades dos dados e em princípios perceptuais (WONGSUPHASAWAT *et al.*, 2016).

Após o usuário selecionar (ou carregar) o conjunto de dados, a ferramenta Voyager já exibe visualizações para cada variável do conjunto. O usuário ainda tem a opção de selecionar as variáveis que deseja analisar, então a ferramenta apresenta a visualização recomendada e também exibe outras visualizações das variáveis selecionadas combinadas com outra variável do conjunto de dados (WONGSUPHASAWAT *et al.*, 2016).

O diferencial da ferramenta RVis é a abordagem utilizada para a recomendação da visualização que é baseada em regras para a classificação das técnicas de visualização. Outra característica é a automatização da recomendação com o uso de técnicas de aprendizado de máquina, especificamente da árvore de decisão (modelo de classificação) que combina critérios baseados nos dados e no contexto (como a tarefa que o usuário deseja realizar) para a definição das regras.

## 2.3 Aprendizado de Máquina

Segundo SHI (1992), aprendizado de máquina (*machine learning*) é o estudo de técnicas que permitam à máquina adquirir novos conhecimentos, novas habilidades e organizar o conhecimento existente. De acordo com NGUYEN & ARMITAGE (2008), a aquisição de conhecimento ocorre de forma automática a partir de experiências, refinamentos

e melhorias na base de conhecimento que a máquina possui. De modo geral, o aprendizado pode ocorrer de duas maneiras: supervisionado ou não-supervisionado (TAN *et al.*, 2005).

No aprendizado supervisionado é fornecido ao algoritmo um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido. Cada exemplo é descrito por um vetor com os valores das características (ou atributos) e com o rótulo da classe associada. O objetivo do algoritmo é construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados (REZENDE, 2003). A classificação e a regressão são tipos específicos do aprendizado supervisionado (NGUYEN & ARMITAGE, 2008).

O processo de aprendizado supervisionado inicia com a divisão dos dados em dois subconjuntos: conjunto de treinamento e conjunto de testes (HAN *et al.*, 2011b). Em seguida, o algoritmo é aplicado no conjunto de treinamento e com isso é obtido um modelo treinado que representa o conhecimento extraído dos dados (NGUYEN & ARMITAGE, 2008). Por fim, o modelo treinado é aplicado ao conjunto de testes. Como o conjunto de testes está previamente rotulado, é possível medir a taxa de acerto do modelo, comparando o resultado com os rótulos do conjunto de testes (NGUYEN & ARMITAGE, 2008).

Segundo HAN *et al.* (2011b), uma das técnicas utilizadas para dividir os dados é denominada validação cruzada (*cross-validation*). Essa técnica consiste em dividir os dados em k partes (*folds*). Destas, k-1 partes são utilizadas para o treinamento e uma parte é utilizada nos testes. Essa divisão ocorre k vezes, de forma que cada parte seja usada uma vez como conjunto de testes (HAN *et al.*, 2011b).

No aprendizado não-supervisionado, o algoritmo analisa o conjunto de dados, ou parte dele, e busca definir grupos, onde os elementos dentro dos grupos apresentam alguma similaridade, sendo dissimilares se comparados aos elementos que compõem os outros

grupos. Após a definição dos agrupamentos, normalmente, é necessário determinar o que cada agrupamento significa no contexto do problema que está sendo analisado (REZENDE, 2003). Agrupamento é um tipo específico de aprendizado não-supervisionado (NGUYEN & ARMITAGE, 2008).

O foco deste trabalho são os algoritmos de aprendizado supervisionado que realizam a tarefa de classificação. De maneira geral, o processo de classificação inicia a partir de um conjunto de exemplos pré-classificados (ou pré-rotulados), que são utilizados para construir um conjunto de regras, ou seja, um modelo. Esse modelo é utilizado na classificação de exemplos não vistos (ou não rotulados) (HAN *et al.*, 2011b) (NGUYEN & ARMITAGE, 2008).

Segundo HAN *et al.* (2011b), diversos modelos de classificação de dados são definidos, tais como: árvore de decisão, Naïve Bayes (ELKAN, 1997), rede neural artificial, k vizinhos mais próximos, algoritmos genéticos e raciocínio baseado em casos. A seguir são descritos com mais detalhes os algoritmos C4.5 (árvore de decisão) (QUINLAN, 1993), Naïve Bayes (ELKAN, 1997), kNN (k-*Nearest Neighbors*) (HAN *et al.*, 2011c) e rede neural do tipo MLP (*Multilayer Perceptron*) (HAN *et al.*, 2011a) que são os algoritmos usados na etapa de testes do conjunto de regras deste trabalho, conforme apresentado na seção 3.3.

O algoritmo C4.5 (sucessor do algoritmo ID3) (QUINLAN, 1993) implementa a técnica árvore de decisão. A árvore de decisão (conforme mostra a Figura 6) é um diagrama na estrutura de árvore, onde cada nó interno (não folha) representa um teste de um atributo e cada ramo representa uma saída do teste. Em cada nó folha fica o rótulo da classe (HAN *et al.*, 2011b).

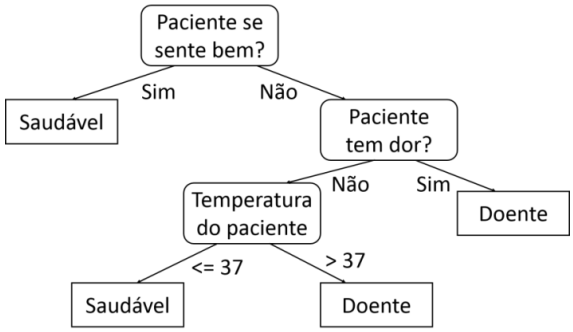


Figura 6. Exemplo de árvore de decisão. Adaptado de REZENDE (2003).

Com a árvore de decisão, a tarefa[9] de classificação é executada da seguinte maneira: dada uma tupla X, para a qual não é conhecido o rótulo da classe, o valor de cada atributo da tupla é testado na árvore de decisão. Um caminho é traçado a partir do nó raiz até o nó folha, o qual armazena o rótulo da classe para aquela tupla (HAN *et al.*, 2011b).

Por exemplo, considere um conjunto (hipotético) de dados de pacientes, conforme apresentado na Tabela 5. Para descobrir a classe do ID = 1 e do ID = 2 (ou seja, o diagnóstico dos pacientes), usando a árvore de decisão da Figura 6, basta iniciar pela raiz seguindo cada teste até que um nó folha seja alcançado. Nesse caso, a classe do ID = 1 e do ID = 2 é, respectivamente, saudável e doente.

Tabela 5. Conjunto (hipotético) de dados de pacientes.

| **ID** | **Paciente se sente bem?** | **Paciente tem dor?** | **Temperatura do paciente** |
|---|---|---|---|
| **1** | Não | Não | <=37 |
| **2** | Não | Sim | - |

O algoritmo Naïve Bayes (ELKAN, 1997) é baseado no Teorema de Bayes e é denominado ingênuo (*naive*) por assumir que os atributos são independentes, ou seja, o

[9] Tarefa no contexto de aprendizado de máquina pode ser entendida como um tipo de problema de descoberta de conhecimento a ser solucionado (DIAS, 2008).

algoritmo assume que o valor de um atributo em uma determinada classe não é influenciado pelos valores de outros atributos (HAN *et al.*, 2011b), (WITTEN *et al.*, 2011). O Teorema de Bayes trabalha com probabilidade condicional, ou seja, permite calcular a probabilidade de um evento ocorrer dada uma condição (HAN *et al.*, 2011b). O algoritmo Naïve Bayes (ELKAN, 1997) utiliza esse teorema para determinar a classe mais provável de uma dada tupla X maximizando a probabilidade a posteriori (WITTEN *et al.*, 2011).

O algoritmo kNN (HAN *et al.*, 2011c) é baseado em aprendizado por analogia, ou seja, compara uma determinada tupla de teste com as tuplas de treinamento similares (HAN *et al.*, 2011b). As tuplas de treinamento são descritas por n atributos. Cada tupla representa um ponto em um espaço n-dimensional (WITTEN *et al.*, 2011). De acordo com HAN *et al.* (2011b), a classificação ocorre da seguinte maneira: dada uma tupla desconhecida X, o algoritmo kNN (HAN *et al.*, 2011c) procura no espaço n-dimensional por k tuplas de treinamento que são similares à tupla X, ou seja, o algoritmo calcula a distância de X a todas as tuplas de treinamento e considera apenas as *k* tuplas mais próximas de X. A classe de X é determinada com base na classe mais frequente dentre essas *k* tuplas (HAN *et al.*, 2011b).

A rede neural do tipo MLP (HAN *et al.*, 2011a) é um algoritmo supervisionado treinado através do algoritmo de retropropagação (*backpropagation*) (WITTEN *et al.*, 2011). Uma rede neural é formada por: uma camada de entrada, uma ou mais camadas escondidas e uma camada de saída. As entradas da rede correspondem aos atributos de cada tupla de treinamento (HAN *et al.*, 2011b). A camada de entrada é responsável pela recepção e propagação das informações de entrada para a camada seguinte. As camadas ocultas são compostas por nós e são camadas que transmitem as informações por meio de conexões entre a camada de entrada e a de saída. A cada conexão está associado um peso, formando um conjunto de pesos que pondera as entradas permitindo gerar um valor para a função de

ativação de cada neurônio, que produz uma saída. O conjunto de pesos de uma rede neural representa o conhecimento adquirido pelo modelo. Por fim a camada de saída é composta por neurônios que recebem as informações das camadas ocultas e fornecem a resposta (WITTEN *et al.*, 2011).

De acordo com HAN *et al.* (2011b), o algoritmo *backpropagation* que implementa uma rede neural, aprende ao processar iterativamente um conjunto de tuplas desconhecidas comparando a predição da rede com a classe desejada. Se a predição estiver errada, o erro é calculado e os valores são retropropagados da camada de saída até a camada de entrada. Conforme as iterações ocorrem, os pesos são ajustados para minimizar o erro e o processamento é feito novamente até que se obtenha a resposta desejada (HAN *et al.*, 2011b).

# 3. Conjunto de regras

Nesta seção são descritas as etapas da metodologia usada para elaborar o conjunto de regras para recomendação de técnicas de visualização de informação. A Figura 7 mostra essas etapas.

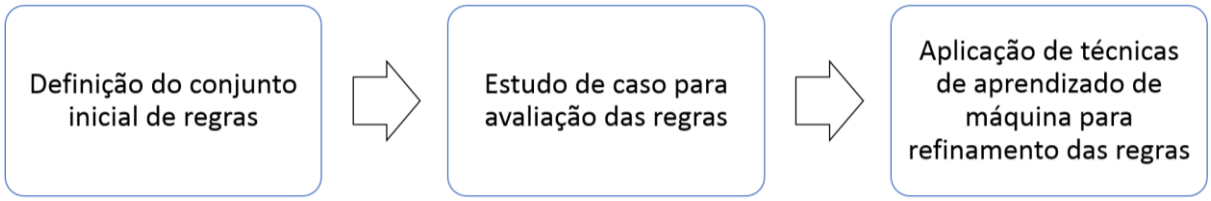


Figura 7. Metodologia usada para elaborar o conjunto de regras para recomendação de visualização.

## 3.1 Definição do conjunto inicial de regras

Após a análise da literatura, foi elaborado um conjunto inicial contendo 93 regras que foram definidas com base nas características do conjunto de dados e nas informações de contexto, como a tarefa que o usuário pode realizar com a visualização. Os dados considerados estão no formato tabular, conforme exemplificado na Tabela 6. Esta delimitação do escopo pode ser justificada pela relativa frequência com que esses dados são gerados. Em MUNZNER (2014), a autora afirma que muitos conjuntos de dados estão no formato de tabelas compostas por linhas e colunas. Além disso, não foram considerados conjuntos de dados que possuem apenas texto, como: notícias, discursos, conteúdo de página web, texto em geral.

Na representação tabular dos dados, denominada por HOFFMAN & GRINSTEIN (2002) como *Table Visualizations*, as linhas são os itens de dados e as colunas são as dimensões, também conhecidas como atributos (DE OLIVEIRA & LEVKOWITZ, 2003) (MUNZNER, 2014). Por exemplo, no conjunto de dados da Tabela 6, as linhas (itens de

dados) representam as pessoas e as colunas (atributos) representam o nome, a idade, o tamanho de camisa e a fruta favorita. Esta nomenclatura, atributo para coluna e itens de dados para linhas, também foi adotada por DE OLIVEIRA & LEVKOWITZ (2003).

Quanto às técnicas de visualização consideradas, faz parte do escopo deste:

- Gráficos em 2D.
- Gráficos estatísticos.
- Gráficos para representação de hierarquias.

Por outro lado, não faz parte do escopo deste trabalho:

- Gráficos em 3D.
- Gráficos para representação de dados geográficos (latitude e longitude).
- Gráficos para representar dados multidimensionais.

Tabela 6. Exemplo de representação tabular do conjunto de dados aceito na ferramenta RVis. Adaptado de MUNZNER (2014).

| **Nome** | **Idade** | **Tamanho de camisa** | **Fruta favorita** |
|---|---|---|---|
| Amy | 8 | P | Maçã |
| Clara | 9 | M | Pêra |
| Ernest | 12 | G | Pêssego |
| George | 9 | M | Laranja |
| Hector | 8 | G | Melancia |

A cada regra, são especificadas as características dos dados e a tarefa a ser realizada. A partir destas informações, é determinada a técnica de visualização mais apropriada conforme a estratégia de recomendação definida. Em relação às características do conjunto de dados, as seguintes variáveis foram consideradas na formulação das regras:

- **Quantidade de unidades de medida**: representa quantas unidades de medida tem o conjunto de dados, especificamente as colunas com valores quantitativos (máximo 2 unidades de medida diferentes).
- **Quantidade de itens de dados**: representa a quantidade de itens de dados (linhas) do conjunto de dados. Foram definidos os seguintes intervalos de valores: [1, 5], [1, 15], [1, 20], [1, 30], [16, 30], [20, 30].
- **Tem valor negativo**: indica se o conjunto de dados possui valor negativo. Os valores adotados são: verdadeiro ou falso.
- **Quantidade de atributos**: informa quantos atributos (colunas) tem o conjunto de dados. O mínimo é 2 atributos e o máximo é 4.
- **Quantidade de atributos quantitativos**: informa quantos atributos são do tipo quantitativo. Considerando que dados puramente textuais e gráficos multidimensionais não fazem parte do escopo deste trabalho, é necessário, no mínimo, 1 atributo quantitativo e no máximo 4.
- **Quantidade de atributos categóricos**: informa quantos atributos são do tipo categórico. Os limites para esta variável são: no mínimo 1 atributo categórico e no máximo 3 atributos. Estas definições foram necessárias pois técnicas multidimensionais não fazem parte do escopo do trabalho.
- **Tipo de atributo categórico**: indica o tipo do atributo categórico que pode ser nominal, ordinal ou intervalar.

Quanto à **tarefa**, foram consideradas as seguintes opções: comparação nominal, correlação, distribuição, parte-todo, ranqueamento, série temporal e hierarquia. Conforme descrito no início desta seção, a escolha destas tarefas foi baseada na análise dos trabalhos encontrados na literatura. Depois de analisadas as sugestões apresentadas pelos autores,

foram identificadas algumas sobreposições que, nem sempre, estavam explícitas. Alguns autores, por exemplo, apresentam nomes diferentes para tarefas conceitualmente semelhantes, como é o caso da tarefa "tendência" definida por RIBECCA (2016) que se refere à tarefa "série temporal" apresentada em FEW (2012). Nestes casos, as duplicatas foram eliminadas.

A Tabela 7 mostra dois exemplos de regras criadas com base nas variáveis apresentadas anteriormente.

Tabela 7. Exemplos de regras.

| **Quantidade de unidade de medida** | **Quantidade de itens de dados** | **Tem valor negativo** | **Quantidade de atributos** | **Quantidade de atributos quantitativos** | **Quantidade de atributos categóricos** | **Tipo de atributo categórico** | **Tarefa** | **Classe** |
|---|---|---|---|---|---|---|---|---|
| 1 | <=5 | Falso | 2 | 1 | 1 | Nominal | Parte-todo | Gráfico de Pizza |
| 2 | <=30 | Falso | 3 | 2 | 1 | Nominal | Correlação | Gráfico de Dispersão |

O conjunto inicial tinha 96 regras, 17 técnicas de visualização e 7 tarefas. A Tabela 8 mostra o número de regras para cada tarefa. A Tabela 9 apresenta as técnicas consideradas e o número de regras estabelecido para cada técnica.

Tabela 8. Número de regras (por tarefa) da primeira versão.

| **Tarefa** | **Regras da primeira versão** |
|---|---|
| Parte-todo | 23 |
| Série Temporal | 22 |
| Ranqueamento | 12 |
| Distribuição | 12 |
| Comparação Nominal | 8 |
| Correlação | 8 |

| | |
|---|---|
| Hierarquia | 5 |

Tabela 9. Técnicas de visualização consideradas e o número de regras para cada técnica da primeira versão.

| **Técnica de Visualização** | **Regras da primeira versão** |
|---|---|
| Gráfico de Linhas | 14 |
| Múltiplos Gráficos de Barras | 11 |
| Gráfico de Barras | 9 |
| Gráfico de Áreas | 8 |
| Múltiplos Gráficos de Linhas | 7 |
| Gráfico de Colunas | 6 |
| Gráfico de Dispersão | 6 |
| Gráfico de Bolhas | 6 |
| Gráfico de Colunas Empilhadas | 5 |
| Treemap | 5 |
| Gráfico de Barras Empilhadas | 4 |
| Múltiplos Gráficos de Colunas | 3 |
| *Spiral Plot* | 3 |
| *Circle Packing* | 2 |
| Histograma | 2 |
| Gráfico de Pizza | 1 |
| Múltiplos Histogramas | 1 |

## 3.2 Avaliação do conjunto de regras

Após definir o conjunto inicial das regras[10] (v1), foram identificados alguns pontos de questionamentos que exigiram uma análise complementar. A estratégia definida para

[10] Disponível em: http://rvis-fvis.rhcloud.com/rvis/regras/regras_rvis_v1.xlsx

avaliação foi apresentar as regras relacionadas a estes pontos de verificação a um grupo de usuários de forma que as recomendações apresentadas pudessem ser avaliadas.

Diferentes situações geraram estes pontos de questionamento:

- Falta de consenso na literatura sobre a adequação de algumas técnicas de visualização. Por exemplo, o gráfico de pizza é um gráfico recomendado por ILIINSKY & STEELE (2011), RIBECCA (2016) e HARDIN *et al.* (2012) para representar valores quantitativos que revelem a porção que cada valor representa em um todo (ou seja, o relacionamento parte-todo). No entanto FEW (2012) não recomenda o uso desse gráfico.
- Imprecisão em relação à quantidade de itens de dados suportada por cada técnica de visualização. Segundo SKAU (2012) a quantidade de itens de dados exibidos pelo gráfico depende da área reservada para a visualização. Para KOSARA (2010), por exemplo, o uso do gráfico de pizza deve ser limitado a 5 "fatias". Já para SKAU (2012), o limite é 7, para HARDIN *et al.* (2012), o limite é 6 e, para CAIRO (2012), o gráfico de pizza deve representar até 3 itens de dados. ILIINSKY & STEELE (2011) não indicam um valor específico para o limite, mas afirmam que o gráfico de pizza deve ser usado com poucas fatias relevantes. Mesmo não havendo consenso na quantidade de dados exibida no gráfico de pizza, todos esses autores concordam que o gráfico de colunas é o mais indicado para substituir o gráfico de pizza quando a quantidade de dados ultrapassar o limite definido (ILIINSKY & STEELE, 2011), (KOSARA, 2010), (SKAU, 2012), (HARDIN *et al.*, 2012).
- Possibilidade de uso de recursos interativos e outros elementos complementares do gráfico.

A partir destes pontos de questionamento, foram definidas as hipóteses que seriam analisadas com o estudo de caso. Essas hipóteses foram relacionadas, principalmente, à característica “quantidade de itens de dados”. É importante deixar claro que os valores adotados como limite da quantidade de itens de dados podem ser dependentes da área de exibição do gráfico. No estudo de caso deste trabalho, a área de exibição foi definida com 1.000px de largura e, no mínimo, 600px de altura. Além disso, em função do tamanho da área de exibição foi definido que o limite máximo de itens de dados considerado seria 30 linhas. Esse valor foi definido em função de testes iniciais.

As hipóteses foram:

H1. A quantidade máxima de itens de dados adequada para o gráfico de pizza é 5, ou seja, o limite máximo de “fatias” é 5.

H2. A quantidade máxima de itens de dados adequada para o gráfico de colunas é 20, ou seja, o limite máximo são 20 colunas.

H3. Gráfico de barras é a opção indicada para exibir mais de 20 itens de dados.

H4. A quantidade máxima de itens de dados adequada para o gráfico de colunas agrupadas é 16 com até 4 grupos.

H5. O rótulo prejudica a análise dos dados quando há mais de 2 atributos quantitativos representados no gráfico.

H6. A quantidade máxima de itens de dados adequada para o gráfico de linhas é 20.

H7. Quando há variáveis com unidades de medidas diferentes e a tarefa não é correlação, recomenda-se *small multiples*, ou seja, vários gráficos (sendo um gráfico para cada variável).

H8. Quando há variáveis com unidades de medidas diferentes e a tarefa é correlação, recomenda-se o gráfico de dispersão.

H9.Para dados com relacionamento hierárquico, usar o treemap para representar conjuntos com no máximo 2 atributos categóricos.

O estudo de caso foi realizado através de uma aplicação web[11] e dividido em duas etapas. Na primeira etapa os participantes foram informados sobre o objetivo do estudo de caso e responderam duas perguntas relacionadas ao perfil, conforme mostra a Figura 8. A terceira pergunta só era exibida se o participante marcasse "sim" para a segunda pergunta.

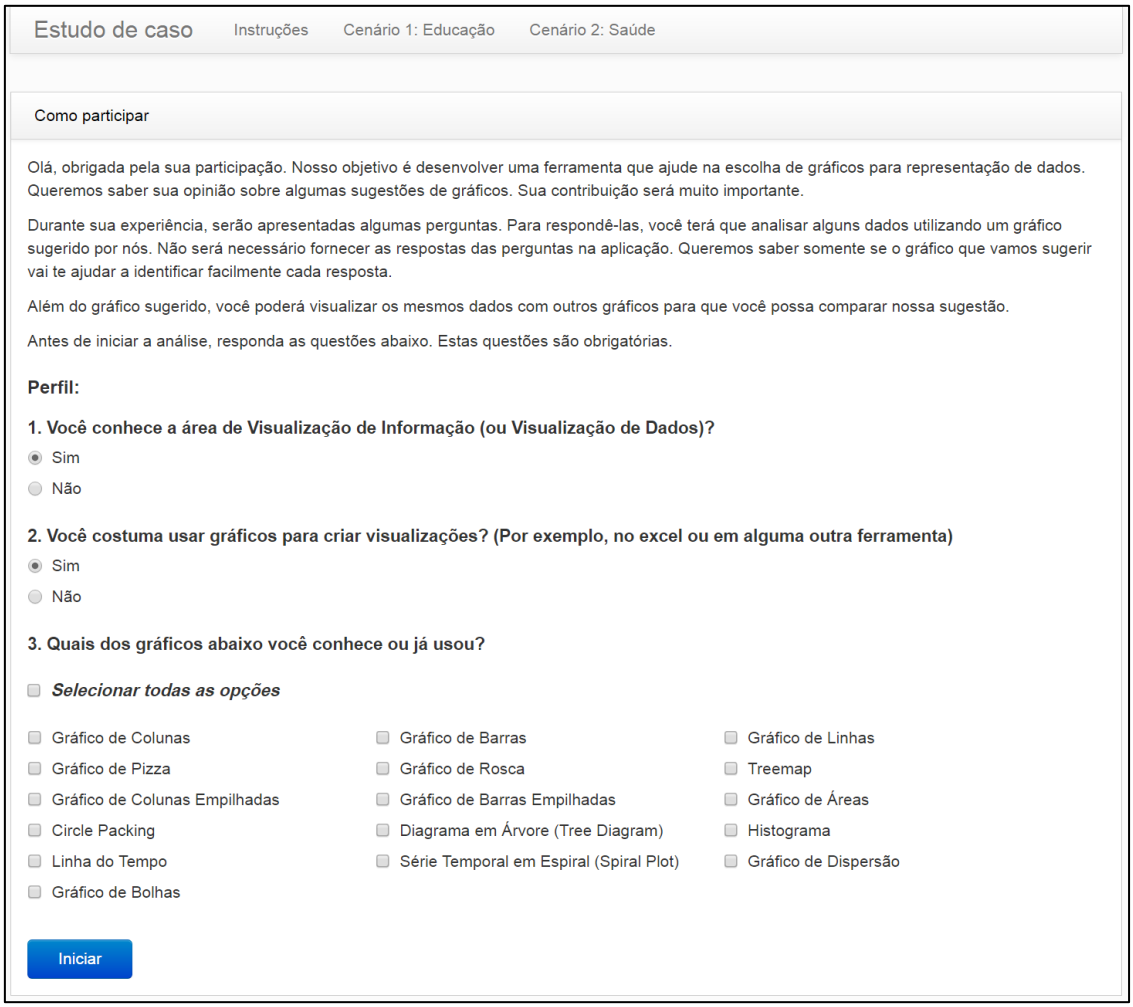


Figura 8. Tela inicial da aplicação web do estudo de caso.

[11] Disponível em http://rvis-fvis.rhcloud.com/rvis/estudo-caso.xhtml

Na segunda etapa, foram elaborados 2 cenários. Para cada cenário, foram apresentadas tarefas que deveriam ser executadas pelo participante. Conforme mostra a Figura 9, essas tarefas eram perguntas sobre o cenário descrito que deveriam ser respondidas a partir da análise da visualização sugerida pela aplicação. Além da técnica sugerida (Figura 10), o participante poderia visualizar os mesmos dados com outras técnicas de modo que fosse possível comparar a recomendação (Figura 12). As respostas das perguntas relacionadas às tarefas não foram coletadas.

Os dados analisados representaram a opinião dos participantes em relação à sugestão da técnica e foram coletados com a aplicação de um questionário. Basicamente, o usuário deveria responder se concordava com a recomendação da ferramenta. Conforme mostra a Figura 11, as respostas para cada questão foram definidas baseadas na escala Likert de 5 pontos: “concordo fortemente”, “concordo”, “neutro”, “discordo” e “discordo fortemente”. Também havia um espaço para o participante colocar comentários.

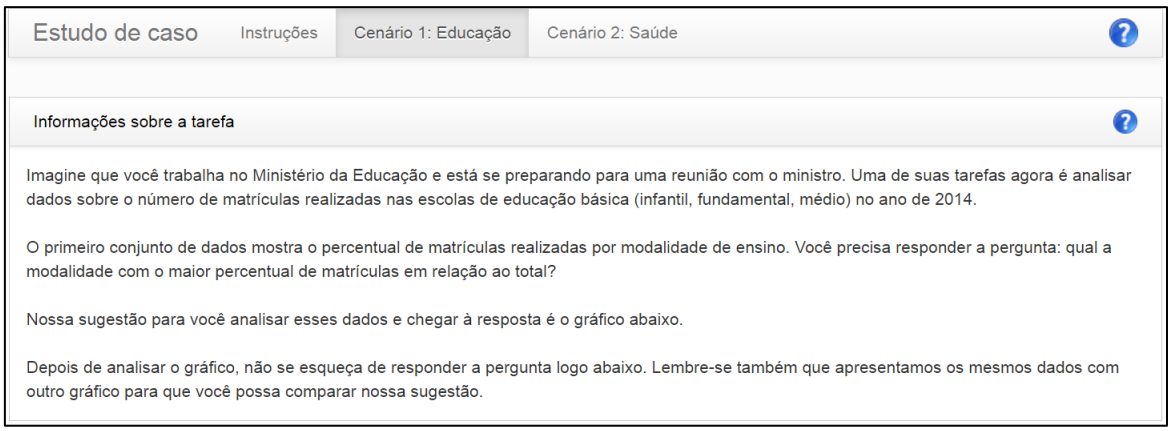


Figura 9. Tela da aplicação: área com informações sobre a tarefa.

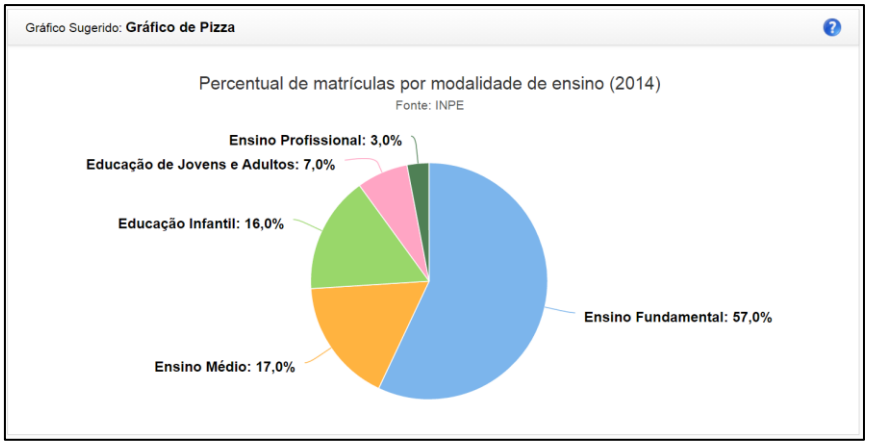


Figura 10. Tela da aplicação: área com o gráfico sugerido.

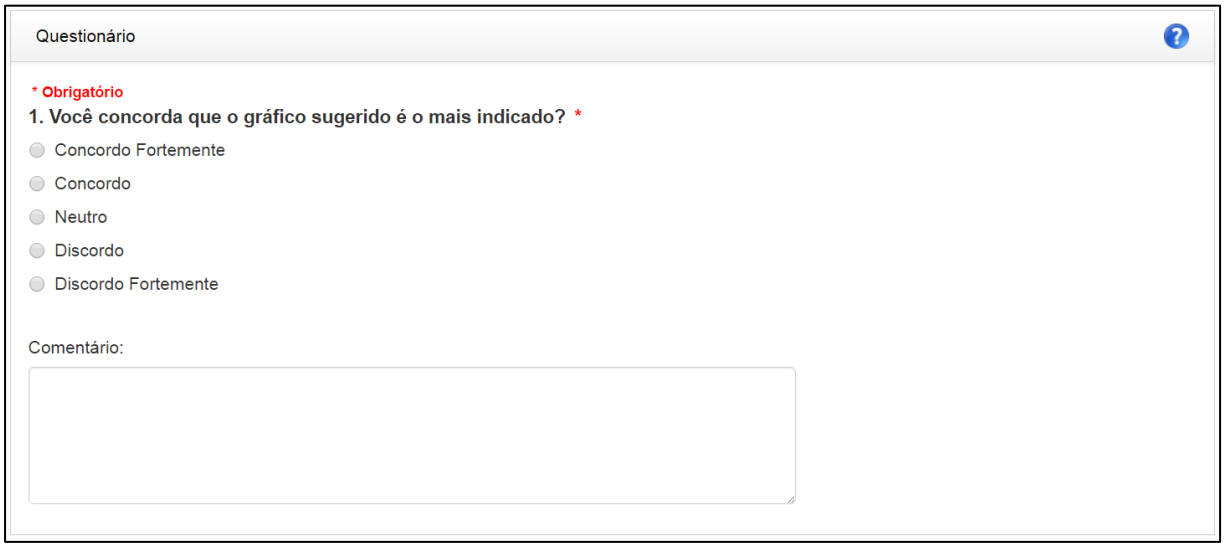


Figura 11. Tela da aplicação: área com o questionário.

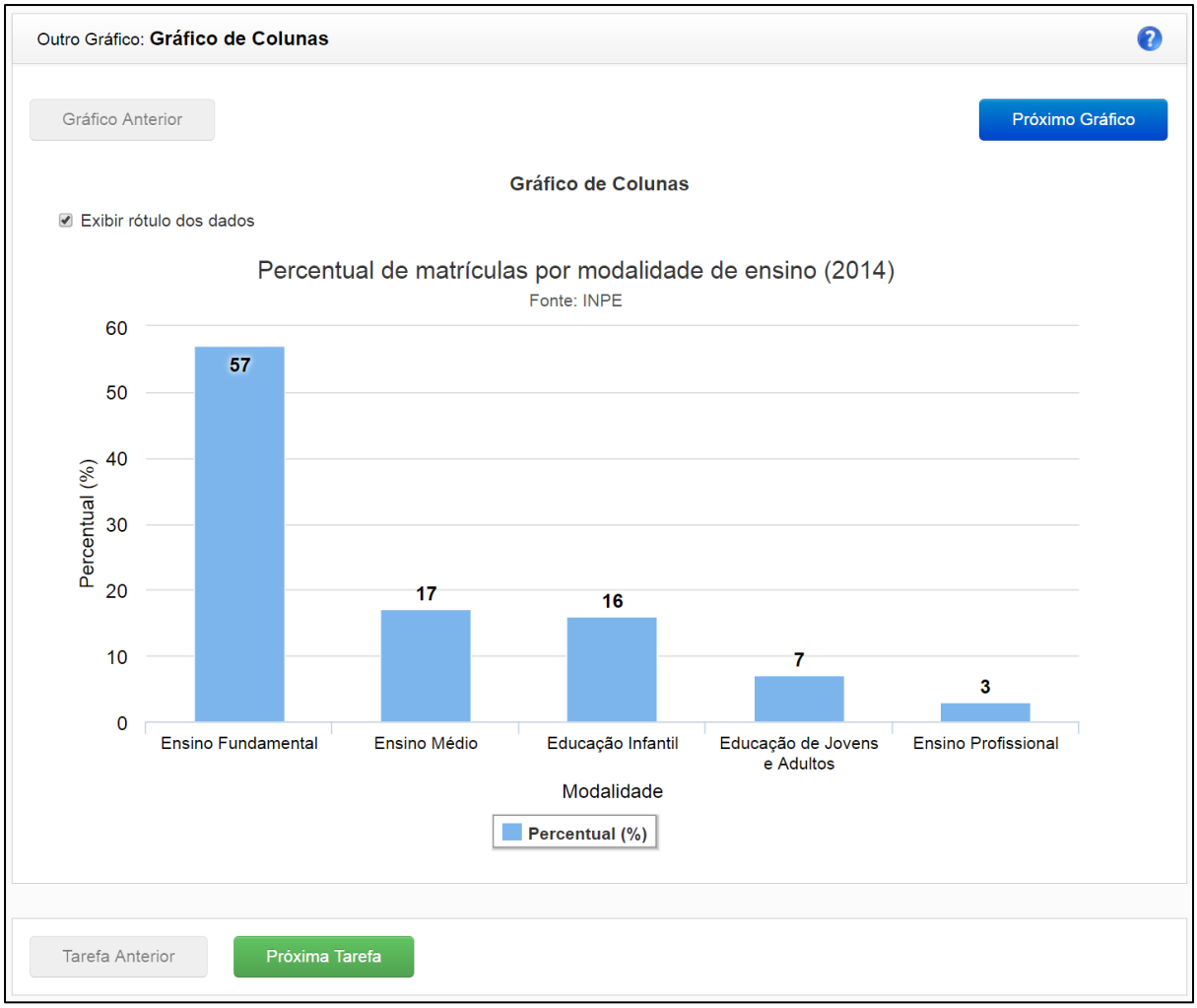


Figura 12. Tela da aplicação: área com outros gráficos.

### 3.2.1 Análise dos resultados

O estudo de caso teve um total de 92 participantes. Cerca de 55% declaram possuir algum conhecimento sobre a área de Visualização de Informação (ou Visualização de Dados) e em torno de 91% dos participantes afirmaram que costumavam usar gráficos para criar visualizações através de alguma ferramenta (por exemplo, o Excel). Para esses participantes, foi solicitado que informassem quais as técnicas de viusalização que os mesmos conheciam ou já utilizaram na criação dessas visualizações. O objetivo desta pergunta foi verificar se os participantes conheciam técnicas de visualização como treemap, *circle packing* e *spiral plot*. A Figura 13 apresenta o ranking dos gráficos conhecidos ou utilizados pelos participantes. Os

gráficos estatísticos como gráfico de pizza, gráfico de barras, gráfico de colunas e gráfico de linhas ficaram nas primeiras posições do ranking. Já os gráficos treemap, *circle packing* e *spiral plot* estão nas últimas posições indicando que eram conhecidos ou utilizados por poucos participantes.

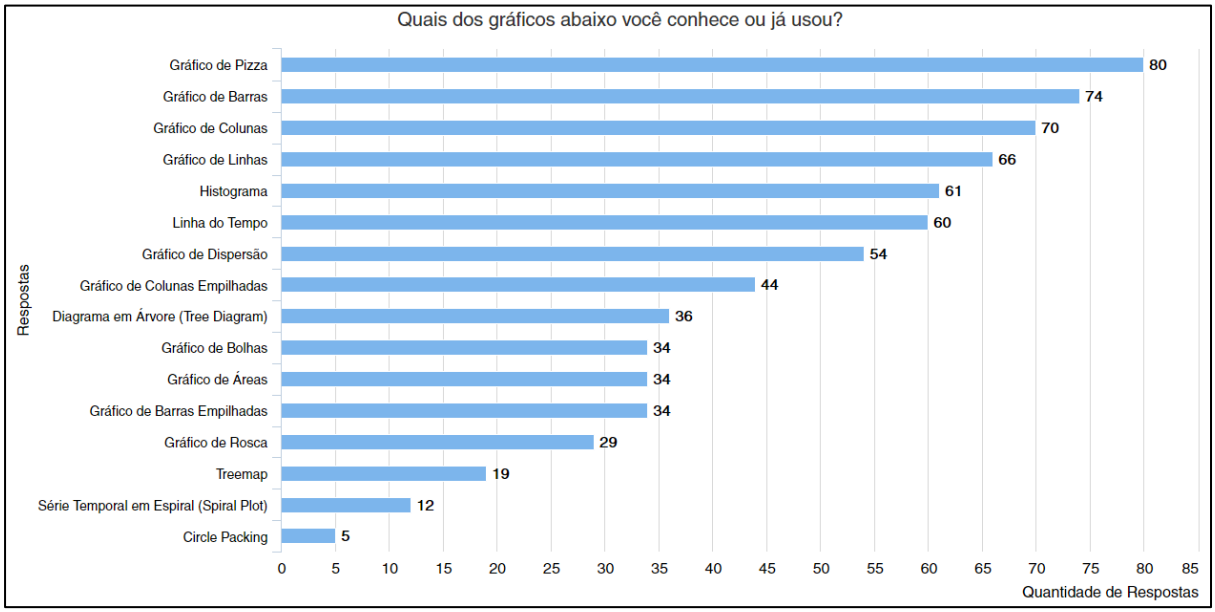


Figura 13. Ranking das técnicas de visualização conhecidas ou utilizadas pelos participantes do estudo de caso.

A seguir, os resultados serão apresentados de acordo com a hipótese testada. Vale ressaltar que, a estratégia de solução adotada foi: a cada hipótese analisada, caso a porcentagem de aceitação fosse maior que a porcentagem de rejeição, a hipótese seria considerada como verdadeira ainda que esses resultados não garantissem o rigor científico para comprovação destas hipóteses. Isso só foi possível ou viável, porque o objetivo foi elaborar um conjunto inicial de regras que ainda passaria por mais ciclos de avaliação.

H1. A quantidade máxima de itens de dados adequada para o gráfico de pizza é 5, ou seja, o limite máximo de “fatias” é 5.

Para realizar essa avaliação, foi apresentado o gráfico de pizza da Figura 14 com 5 itens de dados e o participante deveria identificar a modalidade com o maior percentual de

matrículas em relação ao total (tarefa era parte-todo). Cerca de 89% dos participantes concordaram que o gráfico de pizza foi adequado para a realização da tarefa.

Em seguida, foi apresentado um gráfico de colunas com os dados de matrículas por estado. A tarefa foi identificar em qual estado houve o maior número de matrículas em relação ao total (todo-parte). Na seção "Outro Gráfico", o primeiro gráfico apresentado foi o gráfico de pizza conforme mostra a Figura 15. Ao analisar a imagem, é possível perceber que o volume de dados é incompatível com a capacidade de representação da técnica, o que pode prejudicar a análise. Em torno de 65% dos participantes concordaram com a sugestão do gráfico de colunas para a realização da tarefa parte-todo com mais de 5 itens de dados. Em torno de 21% dos participantes preferiram o gráfico de pizza da Figura 15 (com mais de 5 itens de dados).

Os resultados indicam que H1 é verdadeira. Desta forma, para as regras relacionadas à tarefa parte-todo foram estabelecidas as seguintes definições para a característica "quantidade de itens de dados":

- Se tarefa = 'parte-todo' e quantidade-tens-dados <=5 então classe = 'gráfico de pizza'
- Se tarefa = 'parte-todo' e quantidade-tens-dados > 5 então classe = 'gráfico de colunas'

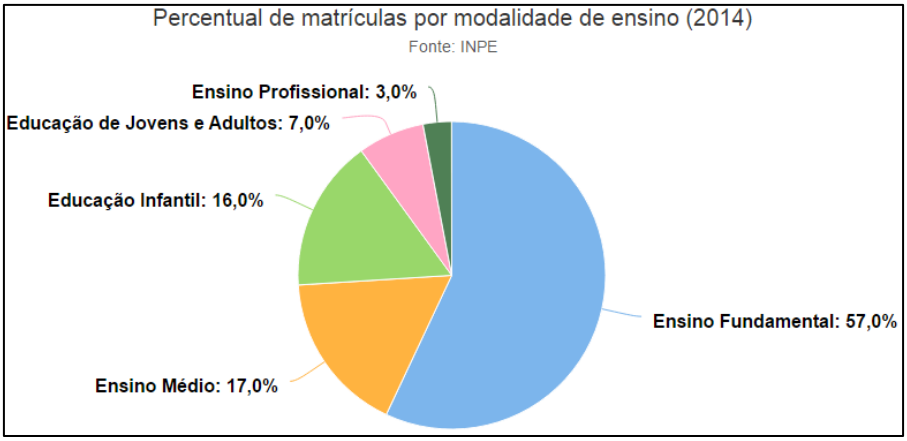

Figura 14. Gráfico de pizza com 5 itens de dados (H1).

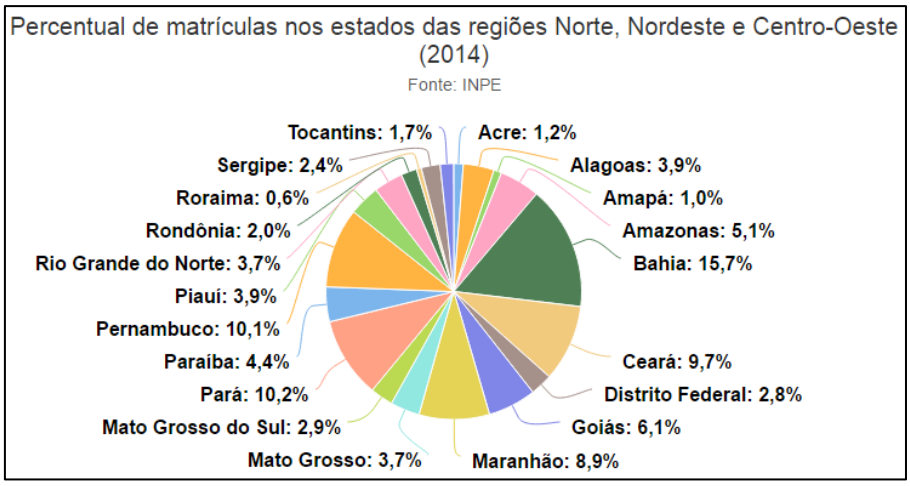


Figura 15. Gráfico de pizza com excesso de dados (H1).

H2. A quantidade máxima de itens de dados adequada para o gráfico de colunas é 20, ou seja, o limite máximo são 20 colunas.

H3. Gráfico de barras é a opção indicada para exibir mais de 20 itens de dados.

Para confirmar essas hipóteses, primeiro foi apresentado o gráfico de colunas da Figura 16 com 20 itens de dados. Em seguida, foi apresentado o gráfico de barras da Figura 17 com mais de 20 itens de dados. Em ambos os casos, o participante precisava comparar os valores.

Cerca de 65% dos participantes concordaram com a recomendação do gráfico de colunas da Figura 16 com até 20 itens de dados. Em torno de 4% foram neutros e cerca de 31% discordaram dessa recomendação. Dos participantes que discordaram, em torno de 53% afirmaram que foi mais fácil comparar os valores com o gráfico de barras (um dos gráficos sugeridos). Esse percentual de discordância pode ser explicado pelo fato de que, no gráfico de barras, os valores são apresentados ordenados (de forma decrescente), o que facilita a comparação dos valores. Além disso, em torno de 83% dos participantes concordaram com a recomendação do gráfico de barras para mais de 20 itens de dados. Esses percentuais indicam que a H3 é verdadeira.

Ainda que a porcentagem de rejeição do gráfico de coluna tenha sido considerável, os resultados indicam que H2 é verdadeira. Logo, para as regras relacionadas à tarefa ‘comparação nominal’ foram estabelecidas as seguintes definições para a característica “quantidade de itens de dados”:

- Se tarefa = ‘comparação-nominal’ e quantidade-tens-dados <=20 então classe = ‘gráfico de colunas’.
- Se tarefa = ‘comparação-nominal’ e quantidade-tens-dados > 20 então classe = ‘gráfico de barras’.

Além disso, quando a tarefa for comparação, a recomendação é adicionar um recurso interativo para ordenar os dados de forma a facilitar a análise.

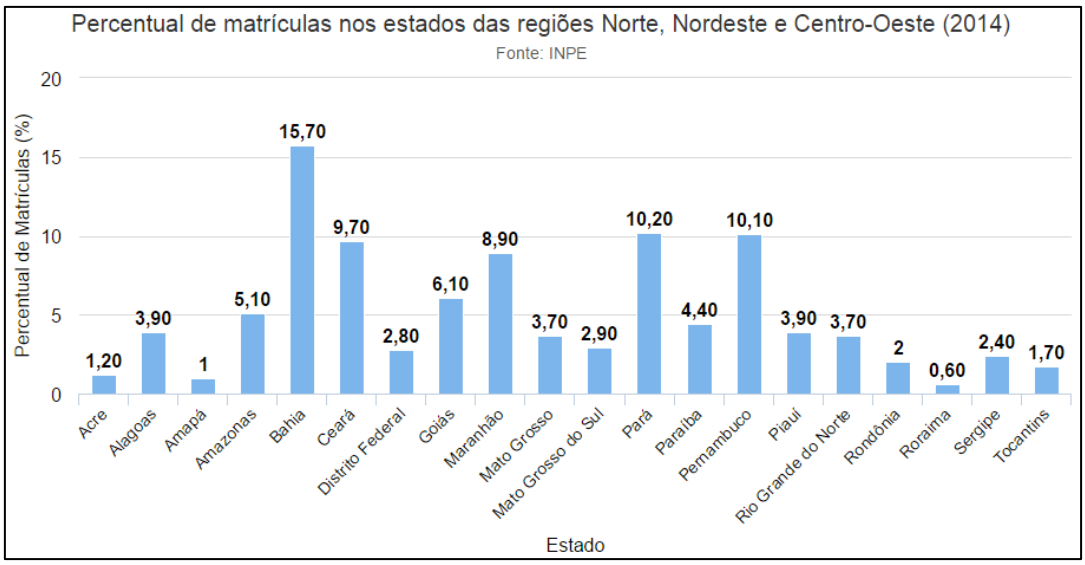


Figura 16. Gráfico de colunas com até 20 itens de dados (H2).

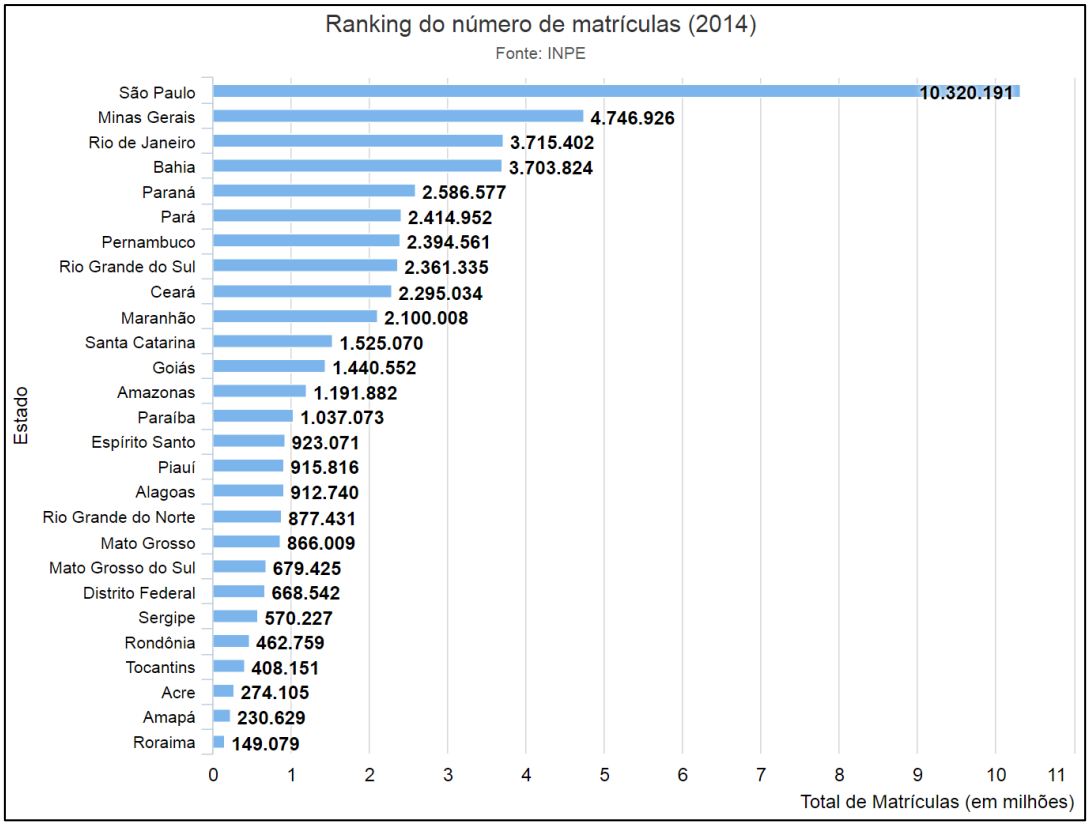


Figura 17. Gráfico de Barras com mais de 20 itens de dados (H3).

H4. A quantidade máxima de itens de dados adequada para o gráfico de colunas agrupadas é 16 com até 4 grupos.

Para avaliar a quantidade de grupos (que corresponde à quantidade de atributos quantitativos nas regras), foi apresentado o gráfico de colunas agrupadas da Figura 18 com 16 itens de dados e 4 atributos quantitativos.

A tarefa dos participantes, ao analisar o gráfico da Figura 18, era identificar em qual estado houve o maior número de matrículas do tipo federal. Cerca de 49% dos participantes discordaram que o gráfico de colunas agrupadas era a técnicas de visualização mais indicada para a realização dessa tarefa. Em torno de 40% concordaram com a recomendação e 11% se mantiveram neutros em relação ao gráfico sugerido. Esse percentual de discordância dos participantes em relação ao gráfico sugerido se justifica devido às características do conjunto

de dados que possuía escalas diferentes. Havia dados na escala de milhões e de mil. Essa diferença na escala e a quantidade de colunas para cada estado pode ter dificultado a análise dos dados.

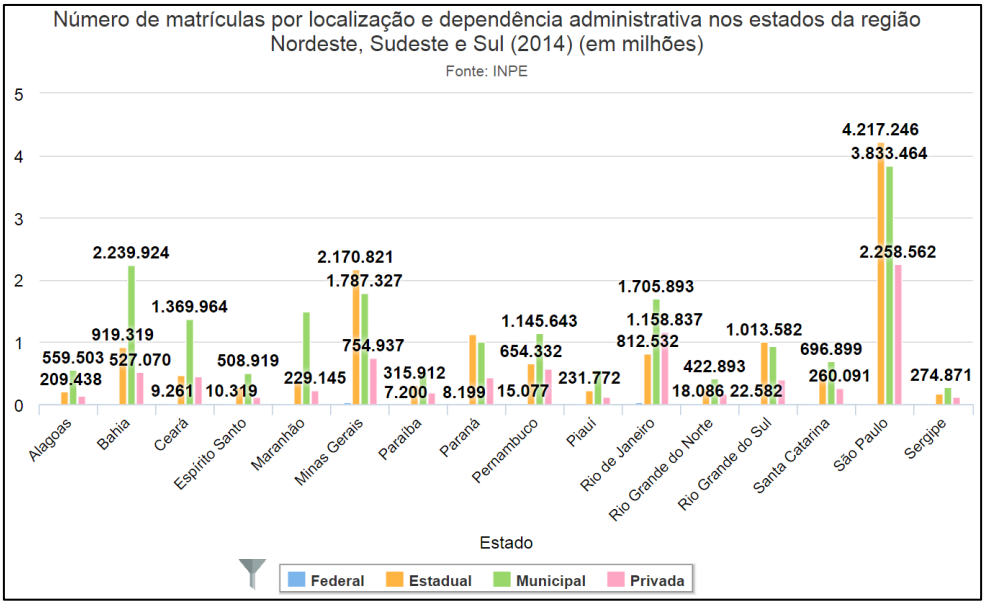


Figura 18. Gráfico de colunas agrupadas com 16 itens de dados e 4 atributos quantitativos (H4 e H5).

Dos participantes que discordaram da recomendação do gráfico da Figura 18, cerca de 62% indicaram o gráfico da Figura 19 como o gráfico mais apropriado para a análise da tarefa proposta. Esse gráfico usa o conceito de "múltiplos pequenos" (do inglês "*small multiples*") proposto por TUFTE (2001). Essa solução envolve um conjunto de gráficos menores, todos dispostos de forma que possam ser vistos simultaneamente (TUFTE, 2001). No gráfico da Figura 19, são vários gráficos de colunas, sendo um gráfico de colunas para cada tipo: federal, estadual, municipal e privada.

Mesmo com o percentual de discordância (em torno de 49%), H4 foi confirmada com algumas ressalvas: será adotada uma quantidade menor de grupos (3), mas a quantidade máxima de itens de dados continuará até 16.

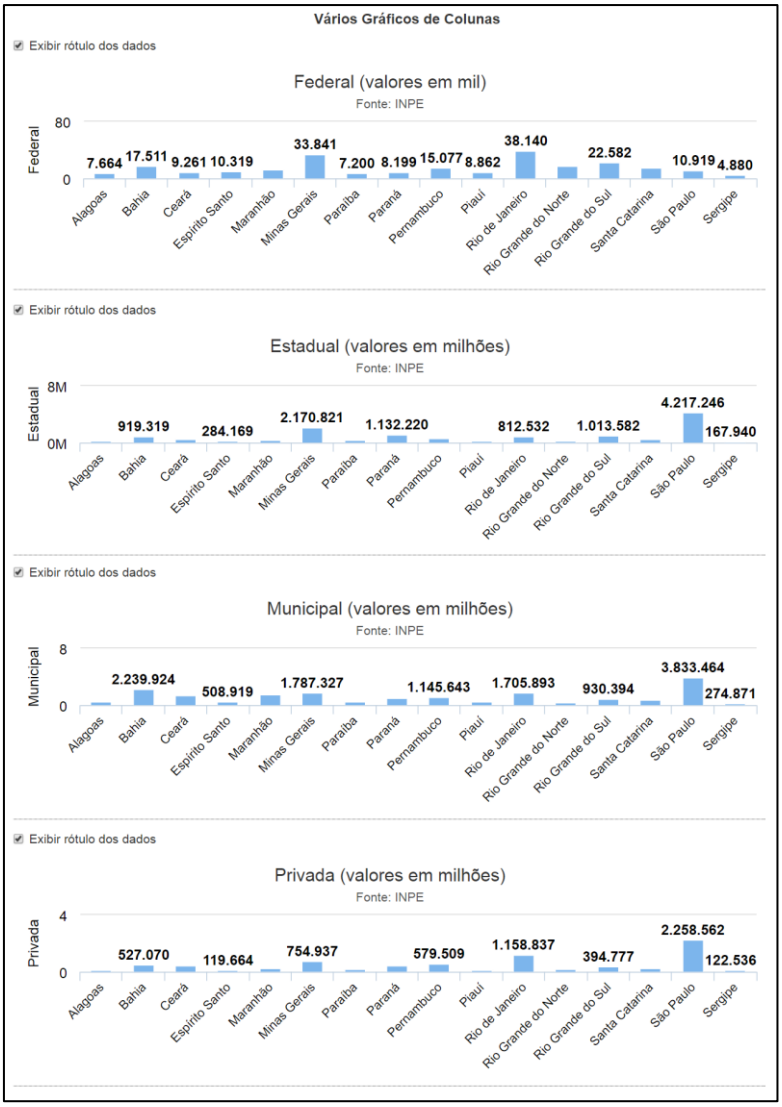


Figura 19. Vários gráficos de colunas como alternativa à H4.

H5. O rótulo prejudica a análise dos dados quando há mais de 2 atributos quantitativos representados no gráfico.

O gráfico da Figura 20 e o gráfico da Figura 18 (que tinham mais de 2 atributos quantitativos) foram usados para avaliar H5. Para 85% dos participantes, o rótulo dos dados contribuiu para realizar a tarefa, contrariando a hipótese. Logo, esse percentual indica que H5 não é verdadeira.

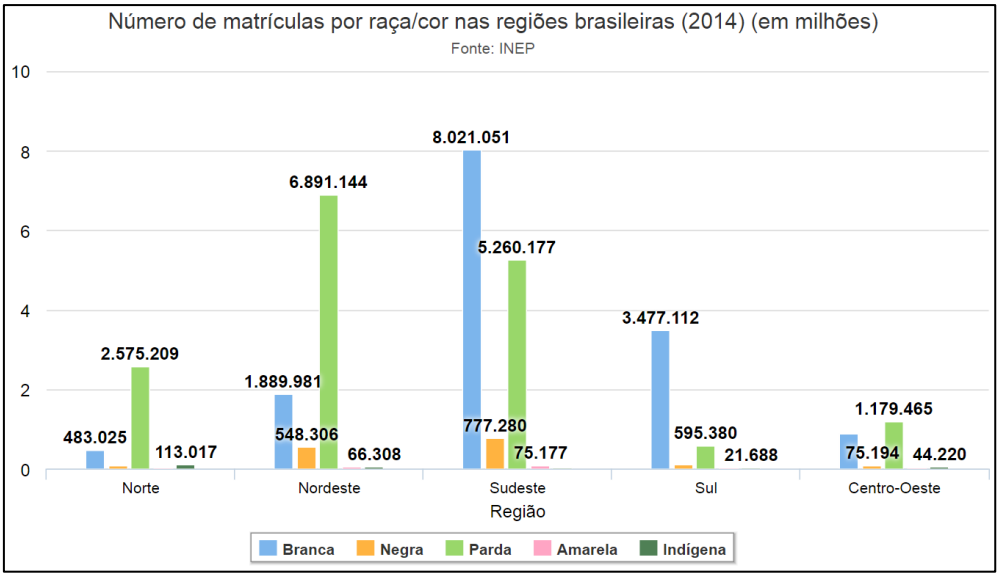


Figura 20. Gráfico de colunas agrupadas com 5 itens de dados e 5 atributos quantitativos (H5).

H6. A quantidade máxima de itens de dados adequada para o gráfico de linhas é 20.

Para avaliar essa hipótese, primeiro foi apresentado o gráfico da Figura 21 com 20 itens de dados. Com esse gráfico o participante deveria analisar o comportamento dos dados ao longo do tempo. Em torno de 81% dos participantes concordaram que o gráfico de linhas foi adequado para a realização da tarefa. Em seguida, o participante deveria analisar a distribuição de outro conjunto de dados temporal. Esse conjunto tinha mais de 20 itens de dados. Para 54% dos participantes o histograma da Figura 22 foi mais adequado na realização da tarefa do que o gráfico de linhas. Esses resultados sugerem que H6 é verdadeira.

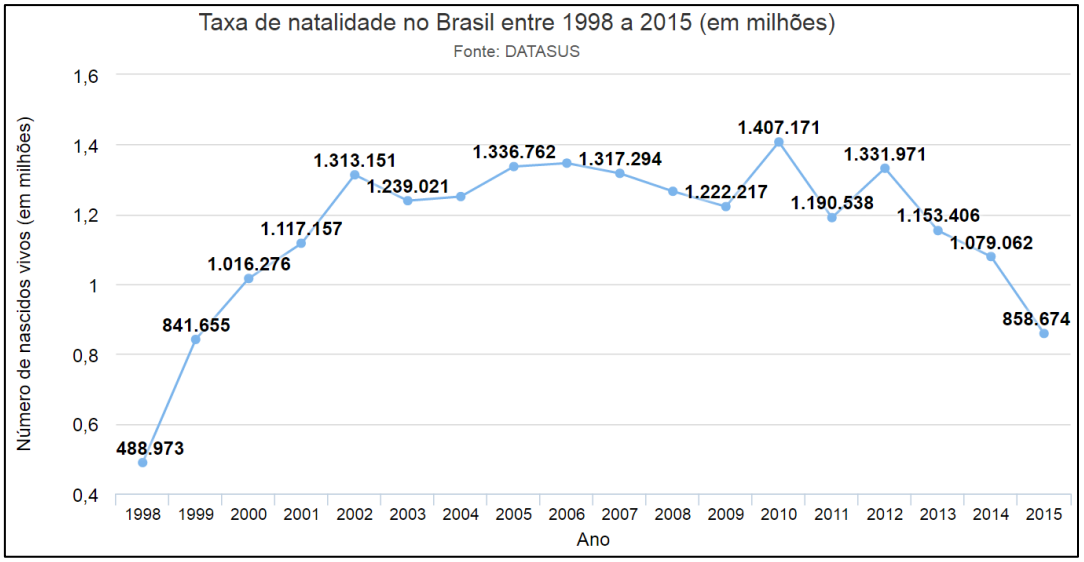


Figura 21. Gráfico de linhas com até 20 itens de dados (H6).

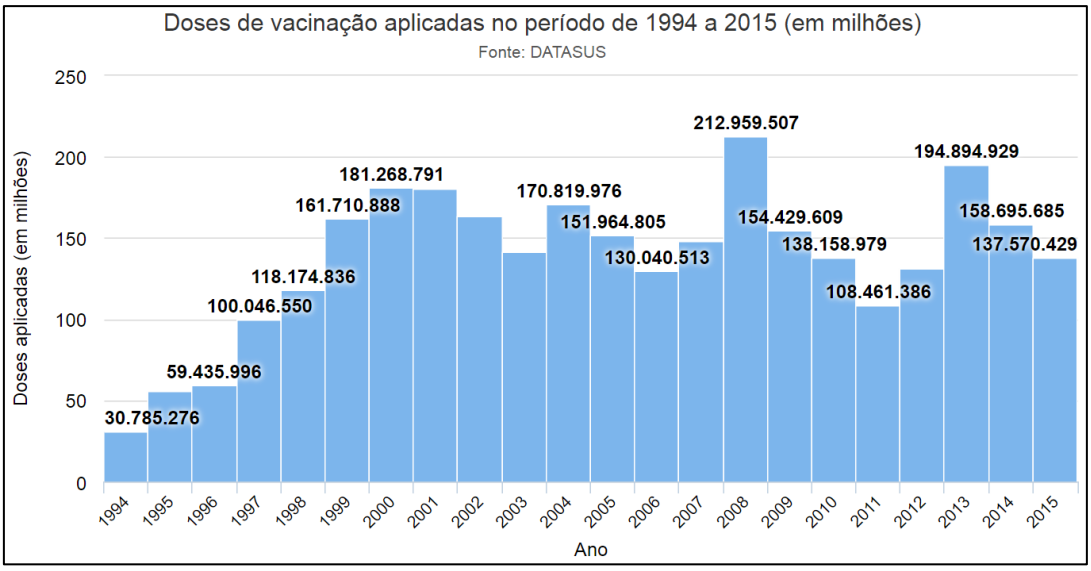


Figura 22. Histograma com mais de 20 itens de dados (H6).

H7. Quando há variáveis com unidades de medidas diferentes e a tarefa não é correlação, recomenda-se *small multiples*, ou seja, vários gráficos (sendo um gráfico para cada variável).

Para confirmar essa hipótese foi apresentado aos participantes o gráfico da Figura 23. Os participantes deveriam descobrir em qual estado brasileiro houve o maior número de internações pelo SUS (Sistema Único de Saúde) e qual estado recebeu a maior quantia em

dinheiro para pagar os gastos com as internações. Apenas 37% dos participantes concordaram que essa técnica de visualização era a mais indicada para realizar a tarefa proposta, 50% discordaram e 13% foram neutros em relação à recomendação.

Da quantidade de participantes que discordaram, cerca de 32% indicaram o gráfico de dispersão da Figura 24. Esse resultado pode ser explicado em função das características do conjunto de dados utilizado na avaliação de H7: os estados com maior número de internações também eram os estados que receberam a maior quantia. Logo, no gráfico de dispersão ficou mais fácil de identificar esse comportamento dos dados. Esse resultado sugere que H7 não é verdadeira.

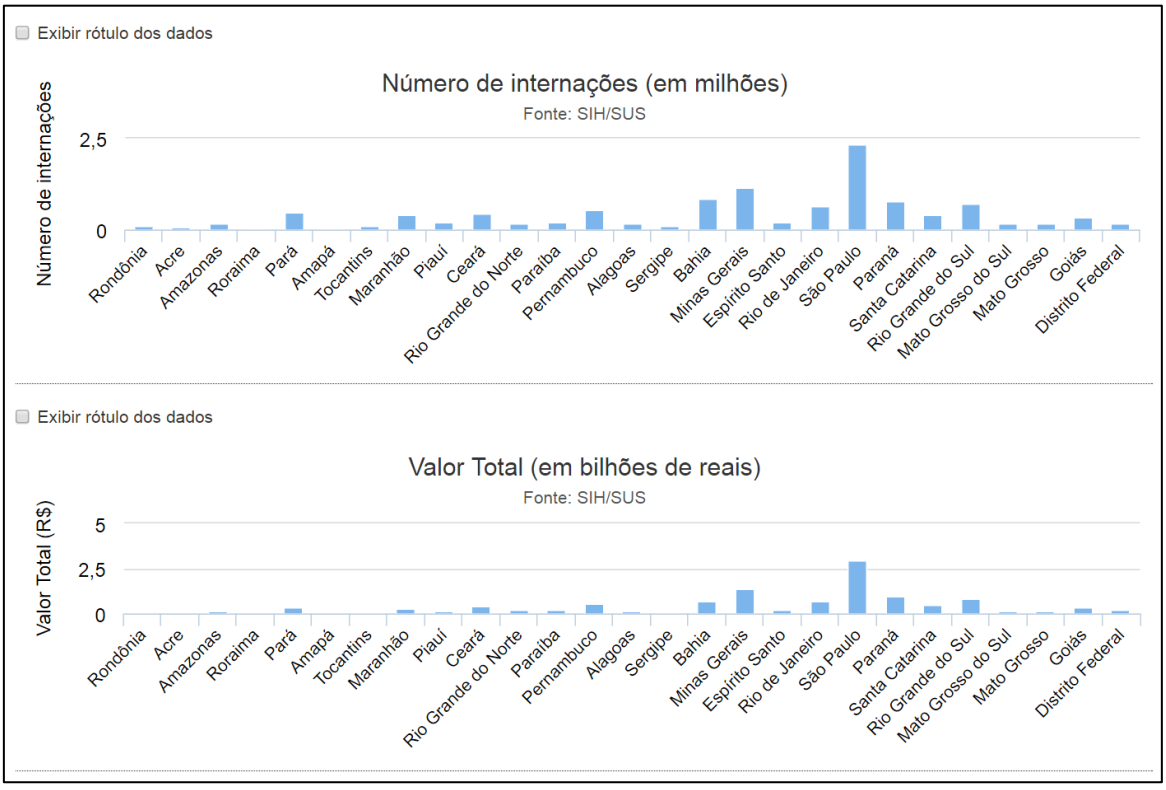


Figura 23. Um gráfico de colunas para cada variável do conjunto de dados (H7).

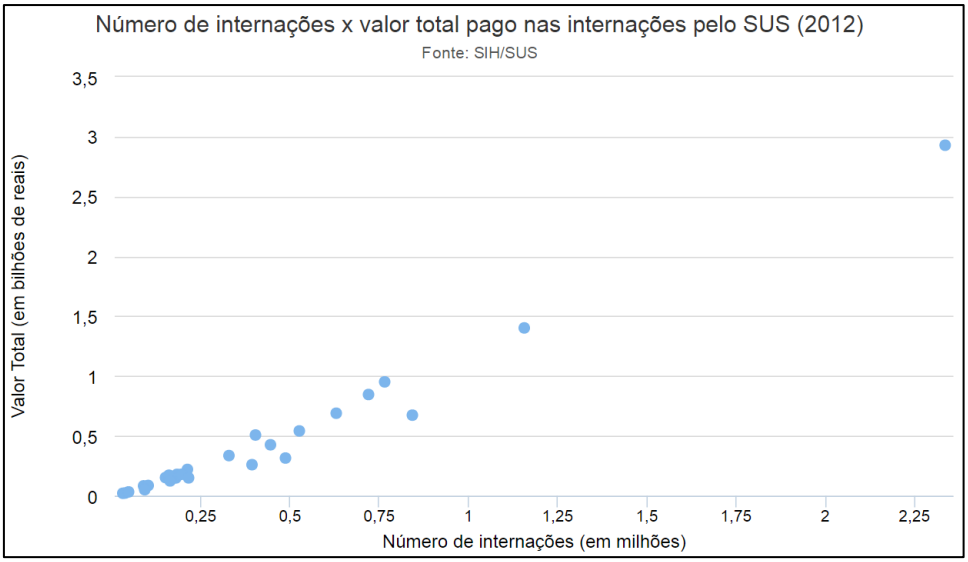


Figura 24. Gráfico de dispersão como alternativa à H7.

H8. Quando há variáveis com unidades de medidas diferentes e a tarefa é correlação, recomenda-se o gráfico de dispersão.

Para avaliar essa hipótese os participantes deveriam analisar a relação entre as taxas de natalidade e de mortalidade no Brasil usando o gráfico da Figura 25. Cerca de 49% dos participantes concordaram com essa recomendação. Mas em torno de 13% discordaram e preferiram o gráfico de colunas para realizar a tarefa proposta. Esse resultado indica que H8 é verdadeira.

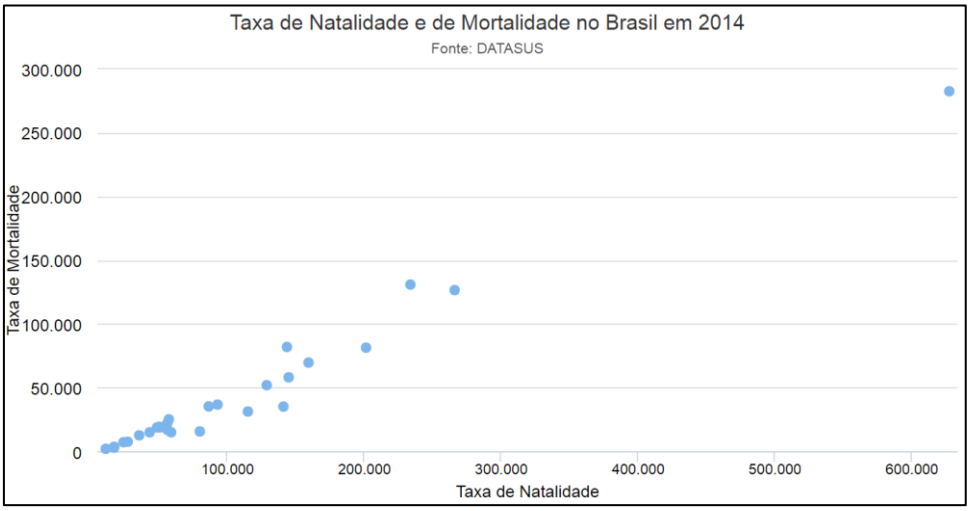


Figura 25. Gráfico de dispersão para representar variáveis com unidade de medidas diferente (H8).

H9. Para dados com relacionamento hierárquico, usar o treemap para representar conjuntos com no máximo 2 atributos categóricos.

Essa hipótese foi avaliada através do gráfico da Figura 26. Ao analisar esse gráfico, os participantes deveriam descobrir qual o tipo de vacina mais aplicada em 2015. Os dados eram caracterizados por um relacionamento hierárquico e estavam organizados por tipo de dose e por tipo de vacina (que representavam os atributos categóricos). Em torno de 61% dos participantes concordaram que o treemap é o mais indicado para a representação esses dados. Cerca de 25% dos participantes preferiram o *circle packing* para representar esses dados. Esse resultado indica que H9 é verdadeira.

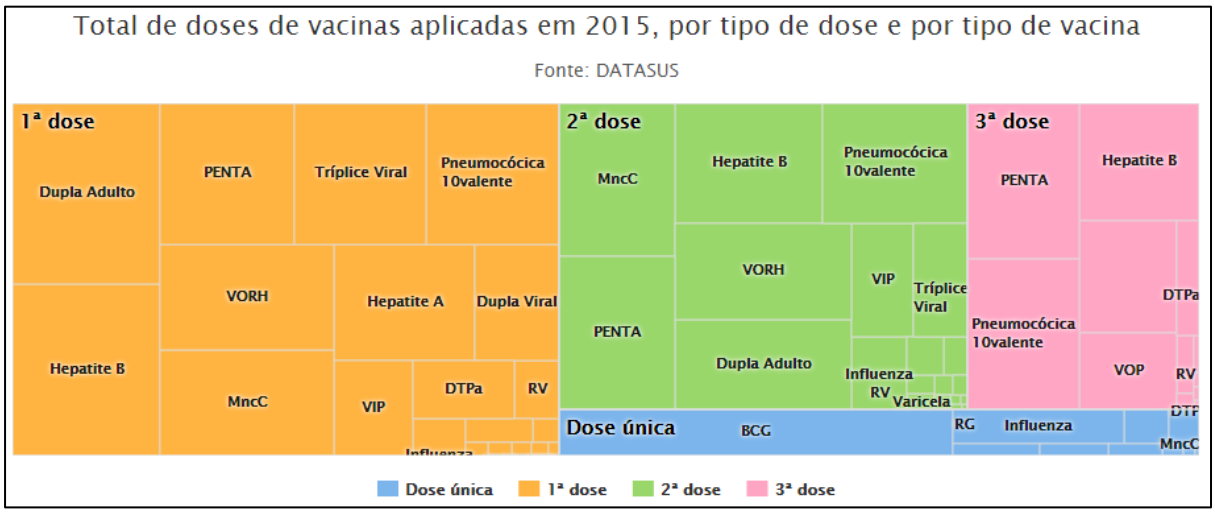


Figura 26. Treemap para representar dados com relacionamento hierárquico (H9).

Com base nesses resultados, alterou-se o conjunto inicial de regras (v1). As principais alterações ocorreram em função da característica "itens de dados". A segunda versão do conjunto (v2) ficou com 90 regras e 17 classes de gráfico[12]. Conforme mostra a Tabela 10, nessa nova versão: o gráfico de barras teve 7 regras associadas e o histograma 1 regra em função dos resultados de H4 e H6, respectivamente. Além disso, o número de regras do

[12] Disponível em: http://rvis-fvis.rhcloud.com/rvis/regras/regras_rvis_v2.xlsx

gráfico de linhas e do gráfico de colunas agrupadas diminuíram. Enquanto que o número de regras da técnica "múltiplos gráficos de colunas" aumentou de 3 para 5 regras. As outras classes de gráficos ficaram com o mesmo número de regras em comparação com a primeira versão do conjunto de regras.

Tabela 10. Número de regras (por técnica de visualização) da 2ª versão em comparação a 1ª versão.

| **Técnica de Visualização** | **Regras da segunda versão** | **Regras da primeira versão** |
|---|---|---|
| Gráfico de Linhas | 13 | 14 |
| Múltiplos Gráficos de Barras | 11 | 11 |
| Gráfico de Barras | 7 | 9 |
| Gráfico de Áreas | 8 | 8 |
| Múltiplos Gráficos de Linhas | 7 | 7 |
| Gráfico de Colunas | 6 | 6 |
| Gráfico de Dispersão | 6 | 6 |
| Gráfico de Bolhas | 6 | 6 |
| Gráfico de Colunas Empilhadas | 4 | 5 |
| Treemap | 5 | 5 |
| Gráfico de Barras Empilhadas | 4 | 4 |
| Múltiplos Gráficos de Colunas | 5 | 3 |
| *Spiral Plot* | 3 | 3 |
| *Circle Packing* | 2 | 2 |
| Histograma | 1 | 2 |
| Gráfico de Pizza | 1 | 1 |
| Múltiplos Histogramas | 1 | 1 |

Conforme mostra a Tabela 11, em relação ao número de regras por tarefa, nessa nova versão (v2), a tarefa "parte-todo" ficou com 23 regras e a tarefa "ranqueamento" com 12 regras. As outras tarefas mantiveram o mesmo número de regras em comparação com a primeira versão.

Tabela 11. Número de regras (por tarefa) da 2ª versão em comparação a 1ª versão.

| Tarefa | Regras da segunda versão | Regras da primeira versão |
|---|---|---|
| Parte-todo | 23 | 24 |
| Série Temporal | 22 | 22 |
| Ranqueamento | 12 | 14 |
| Distribuição | 12 | 12 |
| Comparação Nominal | 8 | 8 |
| Correlação | 8 | 8 |
| Hierarquia | 5 | 5 |

## 3.3 Aplicação de técnicas de aprendizado de máquina para refinamento das regras

Técnicas de aprendizado de máquina foram aplicadas na segunda versão do conjunto de regras com o objetivo de refinar essas regras. Com o auxílio da ferramenta Weka (versão 3.6), foram realizados testes de acurácia e desempenho nos seguintes modelos de classificação: árvore de decisão, kNN (HAN *et al.*, 2011c), Naïve Bayes (ELKAN, 1997) e rede neural do tipo MLP (HAN *et al.*, 2011a). Esses modelos de classificação foram selecionados por serem classificadores largamente empregados na literatura (NGUYEN & ARMITAGE, 2008), (AGARWAL & MITTAL, 2014), (DREISEITL *et al.*, 2001).

As configurações adotadas para os testes no Weka foram:

- Validação cruzada (*cross-validation*) com 10 *folds*.
- Parâmetro "*Random seed for XVal / % Split*" testado com os valores 1, 2 e 3. Esse parâmetro é responsável por gerar *folds* com regras aleatórias. No texto, o parâmetro "*Random seed for XVal / % Split*" será denominado *random seed* apenas para facilitar a leitura.

- Árvore de decisão com o algoritmo C4.5 (QUINLAN, 1993) que no Weka é implementado pelo algoritmo J48. Nos testes foram mantidos os parâmetros fornecidos pelo Weka como padrão ao J48[13].
- kNN (HAN *et al.*, 2011c) com o algoritmo iBk[14], sendo k = 3.
- Rede neural do tipo MLP[15] (HAN *et al.*, 2011a) e Naïve Bayes[16](ELKAN, 1997), mantendo os parâmetros fornecidos pelo Weka como padrão.

Os testes no Weka ocorreram na seguinte ordem:

1. Entrada do arquivo de treinamento (ou seja, o conjunto de regras). Vale ressaltar que as técnicas de visualização (ou seja, os gráficos) são as classes do conjunto de treinamento.
2. Escolha do algoritmo e alteração dos parâmetros (quando necessário).
3. Treinamento e teste do algoritmo usando *cross-validation*. Essa etapa foi realizada para cada valor do parâmetro *random seed* (1, 2 e 3).
4. Coleta e análise dos resultados.

A Figura 27 apresenta o resultado dos testes na segunda versão do conjunto de regras. O resultado é apresentado em função da porcentagem de instâncias (nesse caso, regras) classificadas corretamente. Conforme ilustra a Figura 27, a rede neural do tipo MLP (HAN *et al.*, 2011a) classificou corretamente em torno de 83% das regras, independentemente do valor do parâmetro *random seed*. Esse algoritmo teve o melhor desempenho. Já o algoritmo C4.5

[13] Configuração do algoritmo J48 (árvore de decisão): weka.classifiers.trees.J48 -C 0.25 -M 2
[14] Configuração do algoritmo iBk (kNN): weka.classifiers.lazy.IBk -K 3 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
[15] Configuração do algoritmo MLP: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
[16] Configuração do algoritmo Naïve Bayes: weka.classifiers.bayes.NaiveBayes

(QUINLAN, 1993) (J48 no Weka) teve o pior desempenho e classificou corretamente em torno de 45% das regras.

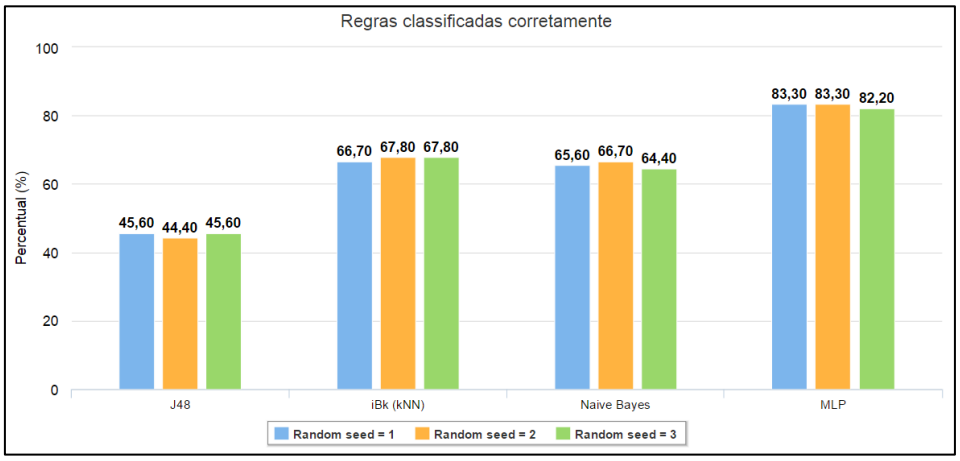


Figura 27. Representação gráfica do resultado dos testes no conjunto de regras v2.

Além da porcentagem de regras classificadas corretamente, também foram analisadas a matriz de confusão e as regras classificadas incorretamente por cada algoritmo. Através dessa análise, foram identificadas 4 regras que possuíam as mesmas características mas recomendavam classes distintas de gráficos. Em função dessa ambiguidade, foi necessário rever o conjunto de regras. Com base nas sugestões encontradas na literatura, 2 regras foram retiradas do conjunto original.

Através da matriz de confusão também foi possível observar que as regras das classes "histograma" e "vários histogramas" não foram classificadas corretamente em nenhum dos testes realizados. Segundo FEW (2012), o histograma é um tipo de gráfico de colunas. Desta forma, essas classes foram substituídas, respectivamente, pelas classes "gráfico de colunas" e "múltiplos gráficos de colunas".

Outra alteração no conjunto de regras foi em relação às classes "gráfico de áreas", "gráfico de colunas" e "gráfico de linhas". Em relação às recomendações do gráfico de áreas e do gráfico de colunas para a tarefa "série temporal", com exceção do C4.5 (QUINLAN,

1993) (J48 no Weka), os demais algoritmos sempre sugeriam o gráfico de linhas (para essas duas classes). A viabilidade de usar o gráfico de linha neste caso foi avaliada e as regras da tarefa "série temporal" foram alteradas para recomendar o gráfico de linhas. Esta alteração possibilitou obter regras que te fato traduzam na máquina o conhecimento e a interpretação humana.

Com essas alterações foi criada mais uma versão do conjunto de regras (v3) com 88 regras e 15 classes de gráfico. Conforme mostra a Tabela 12, em comparação com a segunda versão do conjunto (v2), nessa nova versão (v3), as classes "gráfico de linhas", "treemap" e "vários gráficos de colunas" tiveram mais regras associadas, enquanto que para as classes "gráfico de áreas" e "gráfico de colunas empilhadas" o número de regras associadas diminuiu. As outras classes de gráficos ficaram com a mesma quantidade de regras associadas.

Tabela 12. Número de regras (por técnica de visualização) da 3ª versão em comparação a 2ª versão.

| **Técnica de Visualização** | **Regras da terceira versão** | **Regras da segunda versão** |
|---|---|---|
| Gráfico de Linhas | 16 | 13 |
| Múltiplos Gráficos de Barras | 11 | 11 |
| Gráfico de Barras | 7 | 7 |
| Gráfico de Áreas | 4 | 8 |
| Múltiplos Gráficos de Linhas | 7 | 7 |
| Gráfico de Colunas | 6 | 6 |
| Gráfico de Dispersão | 6 | 6 |
| Gráfico de Bolhas | 6 | 6 |
| Gráfico de Colunas Empilhadas | 4 | 4 |
| Treemap | 5 | 5 |
| Gráfico de Barras Empilhadas | 4 | 4 |
| Múltiplos Gráficos de Colunas | 6 | 5 |
| *Spiral Plot* | 3 | 3 |

| | | |
|---|---|---|
| *Circle Packing* | 2 | 2 |
| Histograma | - | 1 |
| Gráfico de Pizza | 1 | 1 |
| Múltiplos Histogramas | - | 1 |

Quanto à quantidade de regras por tarefa, conforme mostra a Tabela 13, em comparação com a primeira versão (v2), nessa nova versão (v3), apenas para a tarefa "série temporal" diminui o número de regras associadas (20 regras). As outras tarefas ficaram com a mesma quantidade de regras de v2.

Tabela 13. Número de regras (por tarefa) da 3ª versão em comparação a 2ª versão.

| **Tarefa** | **Regras da terceira versão** | **Regras da segunda versão** |
|---|---|---|
| Parte-todo | 23 | 23 |
| Série Temporal | 20 | 22 |
| Ranqueamento | 12 | 12 |
| Distribuição | 12 | 12 |
| Comparação Nominal | 8 | 8 |
| Correlação | 8 | 8 |
| Hierarquia | 5 | 5 |

Os mesmos testes realizados com a versão anterior (v2) também foram realizados com esta nova versão (v3). Conforme os resultados apresentados na Figura 28, o algoritmo da rede neural do tipo MLP (HAN *et al.*, 2011a) continuou com o melhor desempenho na classificação correta das regras e os outros algoritmos melhoraram o desempenho.

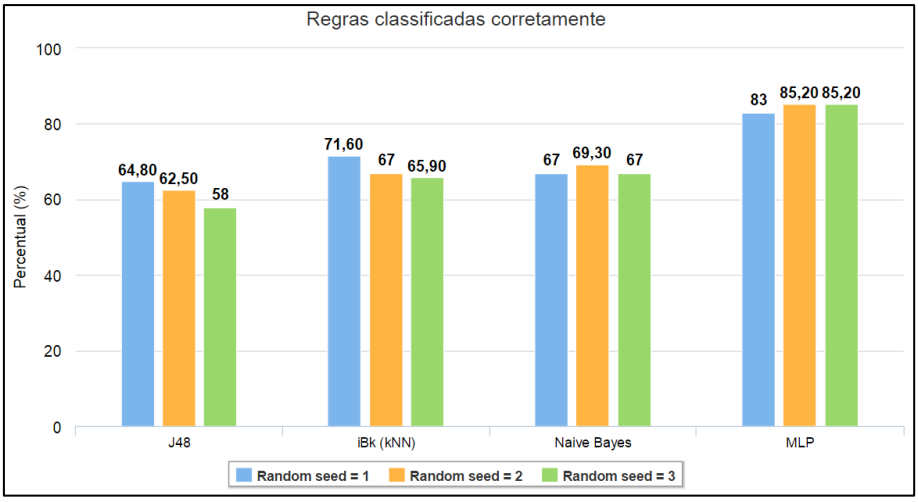


Figura 28. Representação gráfica do resultado dos testes no conjunto de regras v3.

Em relação às alterações nas regras que recomendavam as classes “histograma” e “vários histogramas” e que passaram a recomendar, respectivamente, as classes “gráfico de colunas” e “vários gráficos de colunas”. Todos os algoritmos erraram a classificação dessas regras e recomendaram as classes “gráfico de linhas” e “vários gráficos de linhas”. Isso ocorreu, pois, apenas duas características dessas regras diferenciam as classes: “gráfico de colunas” e “vários gráficos de colunas” das classes “gráfico de linhas” e “vários gráficos de linhas”. Logo, essa alteração não contribui para melhorar a representação das regras e consequentemente melhorar a taxa de acerto dos algoritmos de aprendizado.

Em relação às alterações nas regras da tarefa “série temporal”, apenas o algoritmo Naïve Bayes (ELKAN, 1997) (com *random seed* = 1) errou a classificação de uma regra, recomendando a classe “gráfico de áreas” ao invés da classe “gráfico de linhas”. Logo, conclui-se que essa alteração contribui para melhorar o desempenho dos algoritmos.

Ao analisar a matriz de confusão e as regras classificadas incorretamente pelos algoritmos C4.5 (QUINLAN, 1993) (J48 no Weka), kNN (HAN *et al.*, 2011c), Naïve Bayes (ELKAN, 1997) e a rede neural do tipo MLP (HAN *et al.*, 2011a), foi possível verificar que todos os classificadores erraram a regra relacionada ao gráfico de pizza. Dependendo do valor

do parâmetro *random seed* e do algoritmo, as classes de gráfico sugerida foram: gráfico de colunas, gráfico de barras ou treemap. Esse erro ocorreu porque a regra do gráfico de pizza era uma especialização das regras do gráfico de colunas e do gráfico de barras, por exemplo. Ou seja, algumas características das regras desses três gráficos eram idênticas, como: "unidade de medida", "quantidade de itens de dados", "quantidade de atributos" e "tipo de atributo categórico".

Então foi adicionada ao conjunto de regras uma nova característica ("quantidade mínima de itens de dados") que poderia diferenciar essas regras e, consequentemente, poderia fazer com que os algoritmos acertassem a regra do gráfico de pizza. Por causa dessa nova característica, a característica "quantidade de itens de dados" foi renomeada para "quantidade máxima de itens de dados".

Após essas alterações, a nova versão do conjunto (v4) passou a ter 224 regras e 15 classes de gráfico[17]. A Tabela 14 mostra o número de regras por tarefas e a Tabela 15 mostra o número de regras para cada técnica de visualização da nova versão (v4).

Tabela 14. Número de regras (por tarefa) da 4ª versão em comparação a 3ª versão.

| **Tarefa** | **Regras da quarta versão** | **Regras da terceira versão** |
|---|---|---|
| Parte-todo | 40 | 23 |
| Série Temporal | 64 | 20 |
| Ranqueamento | 24 | 12 |
| Distribuição | 24 | 12 |
| Comparação Nominal | 32 | 8 |
| Correlação | 24 | 8 |
| Hierarquia | 16 | 5 |

[17] Disponível em: http://rvis-fvis.rhcloud.com/rvis/regras/regras_rvis_v4.xlsx

Tabela 15. Número de regras (por técnica de visualização) da 4ª versão em comparação a 3ª versão.

| **Técnica de Visualização** | **Número de regras de v4** | **Número de regras de v3** |
|---|---|---|
| Gráfico de Linhas | 38 | 16 |
| Múltiplos Gráficos de Barras | 20 | 11 |
| Gráfico de Barras | 12 | 7 |
| Gráfico de Áreas | 8 | 4 |
| Múltiplos Gráficos de Linhas | 21 | 7 |
| Gráfico de Colunas | 25 | 6 |
| Gráfico de Dispersão | 12 | 6 |
| Gráfico de Bolhas | 20 | 6 |
| Gráfico de Colunas Empilhadas | 12 | 4 |
| Treemap | 20 | 5 |
| Gráfico de Barras Empilhadas | 4 | 4 |
| Múltiplos Gráficos de Colunas | 15 | 6 |
| *Spiral Plot* | 12 | 3 |
| *Circle Packing* | 4 | 2 |
| Gráfico de Pizza | 1 | 1 |

Os mesmos testes foram realizados nessa nova versão. Conforme o resultado apresentado na Figura 29, nota-se que houve uma melhoria significativa na acurácia e desempenho do algoritmo C4.5 (QUINLAN, 1993) (J48 no Weka) com as alterações realizadas. Esse algoritmo acertou em torno de 90% das regras. A rede neural do tipo MLP (HAN *et al.*, 2011a) passou a acertar em torno de 95% das regras. O Naïve Bayes (ELKAN, 1997) e o iBk (kNN (HAN *et al.*, 2011c)) também melhoraram a acurácia com as alterações realizadas.

Ao analisar a matriz de confusão dos testes, foi possível identificar que nenhum algoritmo acertou a regra do gráfico de pizza. Dependendo do valor de *random seed*, os algoritmos classificaram a regra do gráfico de pizza como sendo ou uma regra do gráfico de

colunas ou do gráfico de colunas empilhadas ou do gráfico de barras. Esse resultado indica que a regra do gráfico de pizza ainda não possui as características necessárias para diferenciá-la das outras regras e assim melhorar a representação dessa regra para o algoritmo de classificação. Logo, esse é um aspecto do conjunto de regras que necessita ser avaliado.

De modo geral, a alteração no conjunto de regras em relação às características "quantidade mínima de itens de dados" e "quantidade máxima de itens de dados" foi fundamental para que os algoritmos (com exceção do iBK – kNN (HAN *et al.*, 2011c)) melhorassem a classificação das regras. Dentre os 4 algoritmos testados, o J48 e a rede neural do tipo MLP (HAN *et al.*, 2011a) obtiveram o melhor desempenho na classificação das regras.

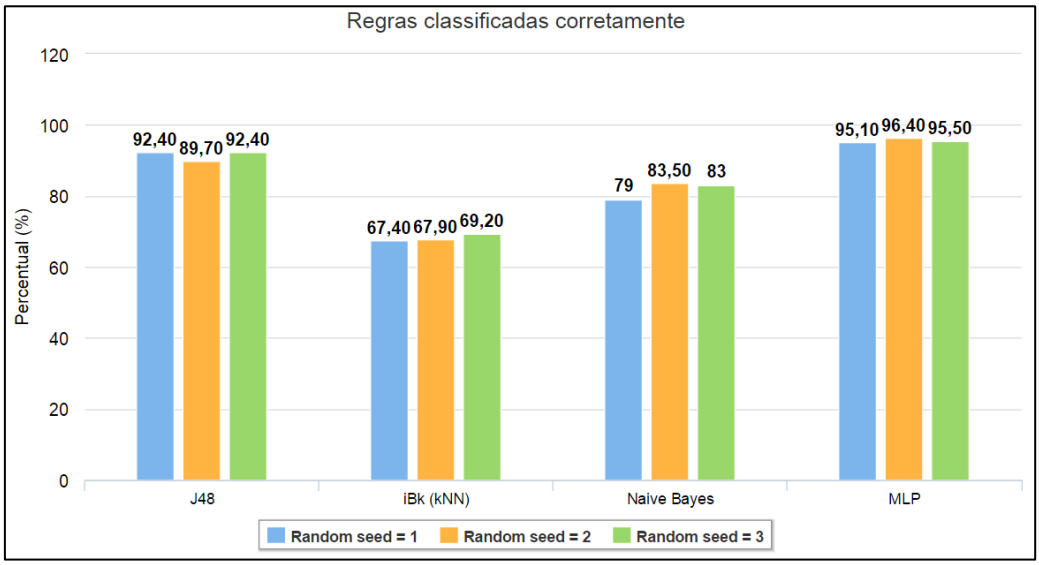


Figura 29. Representação gráfica do resultado dos testes no conjunto de regras v4.

Vale ressaltar que, o conjunto final de regras (v4, bem como as versões anteriores) não está completo, ou seja, não contém todas as possíveis combinações de valores das características do conjunto. Esse conjunto não está completo, pois algumas características são interdependentes, logo não é possível considerar todas as possíveis combinações dos valores dessas características. Por exemplo, não é possível combinar todos valores das características

“quantidade mínima de itens de dados” e “quantidade máxima de itens de dados”, pois os mesmos devem ser definidos em função dos intervalos, que nesse caso são 4 combinações possíveis: [1, 5], [6, 15], [16, 20] e [21, 30].

Além disso, para algumas tarefas, alguns valores de determinadas características não foram utilizados para compor as regras. Por exemplo, não foi criada nenhuma regra para a tarefa parte-todo que considere que a característica “valor negativo” receba valor 1 (verdadeiro), que indica que há valores negativos no conjunto de dados. Pois para esse tipo de tarefa, o conjunto de dados não pode ter valores negativos.

Conforme apresentado anteriormente, a árvore de decisão e a rede neural do tipo MLP (HAN *et al.*, 2011a) tiveram o melhor desempenho nos testes realizados. Sendo assim, a escolha de qual algoritmo implementar na ferramenta RVis ficou entre esses dois algoritmos. KOTSIANTIS *et al.* (2007) apresentaram um estudo comparativo (com base em diversos estudos teóricos e empíricos) de características relacionadas a performance dos seguintes algoritmos de classificação: árvore de decisão, rede neural, kNN (HAN *et al.*, 2011c), Naïve Bayes (ELKAN, 1997), SVM e regras de associação, conforme mostra a Tabela 16.

O estudo comparativo de KOTSIANTIS *et al.* (2007) foi utilizado para escolher entre os algoritmos C4.5 (QUINLAN, 1993) e a rede neural do tipo MLP (HAN *et al.*, 2011a) como o algoritmo a ser implementado na ferramenta desenvolvida neste trabalho. Conforme mostra a Tabela 16, a árvore de decisão tem melhor performance com as características: 2, 4, 5, 8, 10, 12 e 13. Já a rede neural tem melhor performance com as características: 1, 7 e 11. Para as demais características (3, 6 e 9), os dois algoritmos têm performance parecida.

Tabela 16. Comparação dos algoritmos de aprendizado supervisionado. Adaptado de KOTSIANTIS et al. (2007). **** representa a melhor performance e * representa a pior performance.

| ID | Características | Árvore de Decisão | Rede Neural |
|---|---|---|---|
| 1 | Acurácia em geral | ** | *** |
| 2 | Velocidade de aprendizado com relação ao número de atributos e ao número de instâncias | *** | * |
| 3 | Velocidade de classificação | **** | **** |
| 4 | Tolerância a valores faltosos (*missing*) | *** | * |
| 5 | Tolerância a atributos irrelevantes | *** | * |
| 6 | Tolerância a atributos redundantes | ** | ** |
| 7 | Tolerância a atributos altamente interdependentes (exemplo: problemas de paridade) | ** | *** |
| 8 | Lida com atributos discretos/binários/contínuos | **** | ***[18] |
| 9 | Tolerância a ruídos | ** | ** |
| 10 | Lida com o perigo de *overfitting* | ** | * |
| 11 | Tentativas para aprendizado incremental | ** | *** |
| 12 | Fácil interpretação do funcionamento do algoritmo | **** | * |
| 13 | Manipulação dos parâmetros do modelo | *** | * |

No contexto do conjunto de regras deste trabalho, não são todas as características da Tabela 16 que foram consideradas na escolha dos algoritmos. Por exemplo, a característica "tolerância a valores faltosos", pois o conjunto de regras não tem valores faltosos, todas as regras são completas. A característica "tolerância a ruídos" também não precisa ser considerada, pois o conjunto de regras não possui ruídos.

[18] Exceto atributos discretos.

Excluindo essas duas características, a árvore de decisão continua com melhor performance em 6 características enquanto que a rede neural tem performance melhor em 5 características (incluindo as características com performance semelhante). Então com base nesse estudo, o algoritmo C4.5 (QUINLAN, 1993) da árvore de decisão (J48 no Weka) foi escolhido como *engine* de recomendação da ferramenta RVis.

# 4. RVis: Ferramenta para Recomendação de Visualização

A ferramenta RVis foi construída como uma aplicação web[19], conforme ilustra a Figura 30. Para a ferramenta sugerir uma visualização ao usuário, primeiramente é necessário fazer o upload do conjunto de dados. O arquivo do conjunto de dados deve estar no seguinte formato tabular: a primeira linha contém o nome (cabeçalho) das colunas e as demais linhas do arquivo contêm os itens de dados. A ferramenta disponibiliza arquivos de exemplo desse formato nas extensões TXT, CSV e XLS(X).

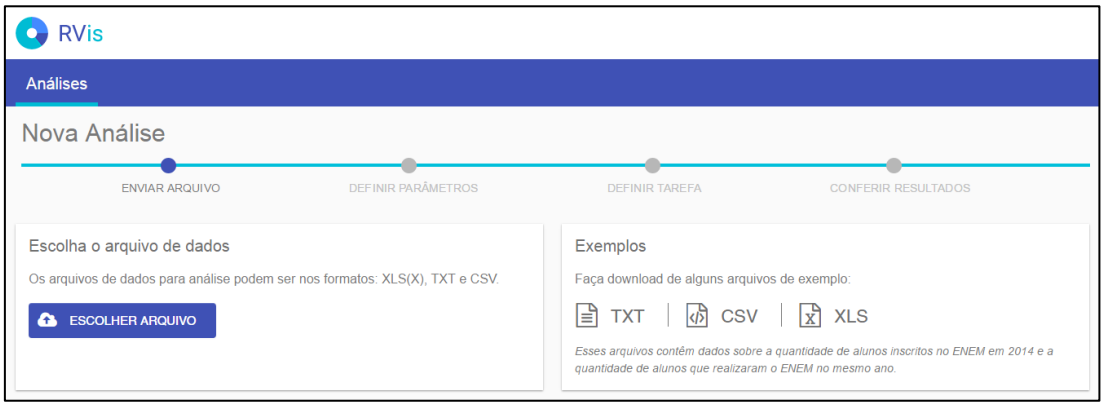


Figura 30. Página inicial da ferramenta RVis.

Se o arquivo de dados estiver no formato correto, a ferramenta exibe para o usuário a tela da Figura 31, na qual o usuário informa o tipo de dado de cada coluna do conjunto carregado. Para auxiliar o usuário na definição desses parâmetros, a ferramenta disponibiliza a explicação dos tipos de dados (nominal, ordinal, intervalar e numérico). Além disso, também exibe parte dos dados, caso o usuário esqueça o conteúdo de cada coluna. Para

[19] A aplicação web RVis está disponível em http://rvis-fvis.rhcloud.com/rvis/.

visualizar todos os dados do conjunto, o usuário deve clicar no *link* correspondente. A etapa seguinte à definição dos parâmetros, é a escolha dos dados a serem visualizados e da tarefa, conforme mostra a Figura 32.

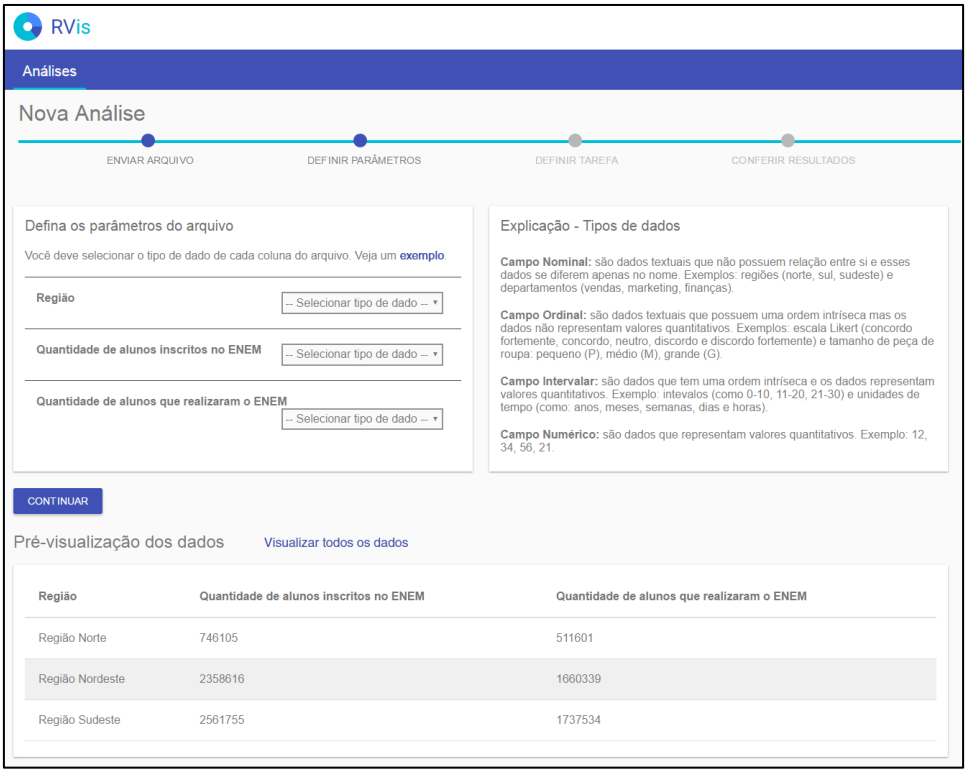


Figura 31. Tela para usuário definir parâmetros da visualização.

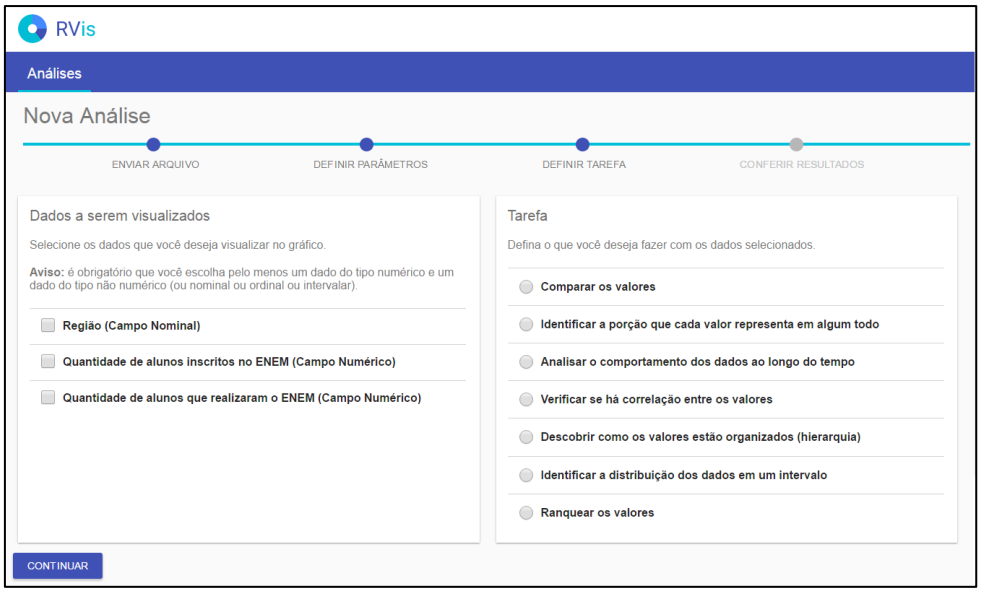


Figura 32. Tela para usuário escolher os dados a serem visualizados e definir a tarefa.

Conforme mostra a Figura 33, na última etapa, a ferramenta exibe a visualização recomendada com base nas escolhas do usuário. As visualizações foram implementadas com o auxílio das bibliotecas HighCharts[20] e D3js[21]. O usuário pode customizar o gráfico inserindo informações de título e subtítulo do gráfico e também o título dos eixos X e Y. A ferramenta também permite imprimir e exportar o gráfico em PNG e PDF.

Ainda na tela da etapa "Conferir Resultados", a ferramenta exibe outros gráficos, conforme mostra a Figura 33. A definição desses outros gráficos é realizada em função apenas dos dados selecionados, não considerando a tarefa. Por exemplo, ao selecionar os três campos da Figura 32, independente da tarefa, e não repetindo o gráfico sugerido, a ferramenta exibe na área "Outros Gráficos", as seguintes opções de visualização: gráfico de barras, gráfico de linhas, gráfico de áreas, gráfico de colunas empilhadas, gráfico de barras empilhadas, gráfico de dispersão, treemap, *circle packing*, vários gráficos de colunas, vários gráficos de barras, vários gráficos de linhas e vários gráficos de áreas.

Vale ressaltar que, dependendo dos dados selecionados pelo usuário para serem exibidos na visualização, algumas técnicas não são consideradas na seção "Outros Gráficos", pois não é possível gerar a visualização com os dados escolhidos. Por exemplo, o gráfico de bolhas não está na lista de outros gráficos da Figura 33 devido às características dos dados selecionados (1 campo nominal e 2 campos numéricos). Para exibir o gráfico de bolhas é necessário que o usuário selecione dados com as seguintes características: 1 campo nominal e 3 campos numéricos.

Ao lado da área "Outros Gráficos" está disponível um mecanismo para coletar o feedback do usuário em relação à recomendação feita pela ferramenta RVis. Caso o usuário

[20] http://www.highcharts.com/
[21] https://d3js.org/

discorde da sugestão da visualização, esse usuário pode indicar um ou mais gráficos que representaram melhor os dados selecionados. Essa indicação do usuário é enviada para o banco de dados que guarda a nova regra sugerida pelo usuário. Essas novas regras podem ser usadas para melhorar o algoritmo de recomendação da ferramenta.

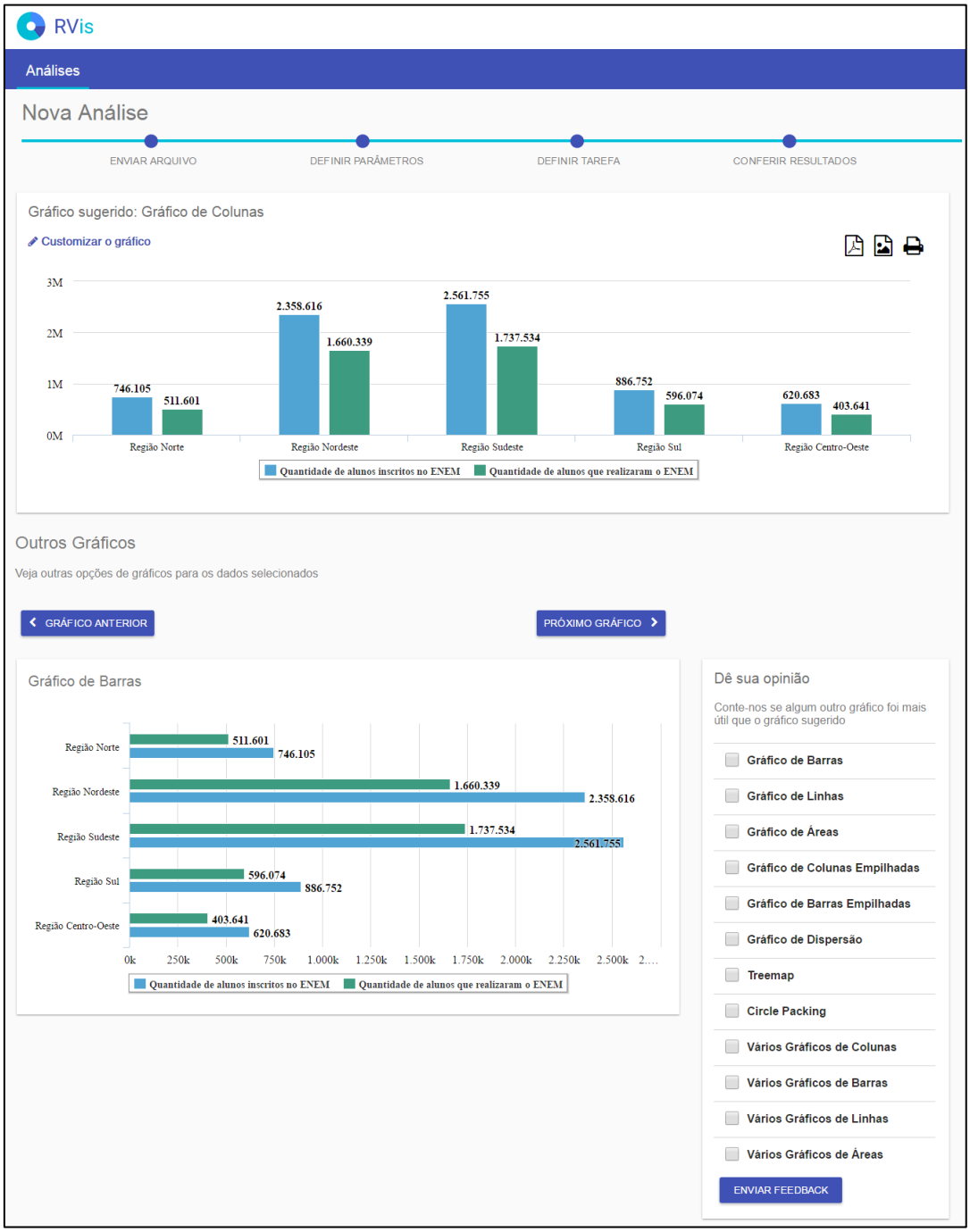


Figura 33. Tela que exibe a visualização recomendada pela ferramenta RVis.

# 5. Avaliação da Ferramenta RVis

Esta seção descreve a metodologia utilizada na avaliação da ferramenta RVis, bem como os resultados obtidos. O planejamento da avaliação foi baseado no framework DECIDE (ROGERS *et al.*, 2011) que tem por objetivo auxiliar avaliadores inexperientes a planejarem e realizarem avaliações de IHC (Interface Humano-Computador) em geral (PRATES & BARBOSA, 2003). De acordo com esse framework, as avaliações de IHC devem ocorrer em 6 etapas, que são (ROGERS *et al.*, 2011), (PRATES & BARBOSA, 2003):

1. **D**etermine: determinar os objetivos gerais da avaliação.
2. **E**xplore: explorar questões mais específicas a serem respondidas, ou seja, decompor os objetivos em perguntas específicas considerando os usuários-alvo e as atividades.
3. **C**hoose: escolher as técnicas de avaliação para responder as questões.
4. **I**dentify: identificar questões práticas, como a seleção dos participantes.
5. **D**ecide: decidir como lidar com questões éticas, por exemplo, o anonimato dos participantes.
6. **E**valuate: avaliar, interpretar e apresentar os dados.

Foi realizado um estudo de caso para avaliar a usabilidade e a recomendação da ferramenta RVis. Esse estudo foi planejado e executado de acordo com cada etapa do framework DECIDE (ROGERS *et al.*, 2011), conforme apresentado a seguir.

**1. Determinar os objetivos gerais da avaliação**

O objetivo geral foi avaliar a utilidade das recomendações apresentadas pela ferramenta RVis, ou seja, verificar se os gráficos sugeridos atenderam à expectativa dos usuários. Além disso, também foi avaliada a usabilidade e a funcionalidade da ferramenta.

**2. Explorar questões específicas a serem respondidas**

Para atingir o objetivo foram definidas questões específicas como:

- Os usuários encontraram dificuldades para usar a ferramenta?
- Os gráficos sugeridos pela ferramenta atenderam à expectativa dos usuários?
- A visualização sugerida pela ferramenta permitiu ao usuário entender os dados e obter *insights* (novos conhecimentos sobre os dados)?
- Alguma outra visualização facilitou o entendimento dos dados?

**3. Escolher o paradigma e as técnicas de avaliação**

A técnica de avaliação escolhida foi o estudo de caso com usuários finais que avaliaram a usabilidade e as recomendações da ferramenta. A coleta dos dados dessa avaliação foi realizada através do questionário apresentado na Tabela 17.

Tabela 17. Questionário para avaliação da ferramenta RVis.

| **Perfil do Participante** |
|---|
| 1. Você conhece a área de Visualização de Informação (ou Visualização de Dados)?<br>( ) Sim ( ) Não<br><br>2. Você costuma usar gráficos para criar visualizações? (Por exemplo, no Excel ou em alguma outra ferramenta)<br>( ) Sim ( ) Não<br><br>3. Você conhece outras ferramentas que auxiliam na seleção e criação de visualizações?<br>( ) Sim ( ) Não |
| **Comparação com outras ferramentas** |
| 4. Com que frequência você utiliza essas ferramentas?<br>1 – Raramente e 5 – Frequentemente<br>( ) 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 |

5. Conte-nos quais são essas outras ferramentas que você conhece e/ou utiliza que auxiliam na seleção e criação de visualizações

[ ] Watson Analytics

[ ] Tableau

[ ] QlikView

[ ] Excel

[ ] Outros _____

**Questões gerais**

6. Foi fácil navegar pela interface da ferramenta?

( ) Sim ( ) Não

7. O número de etapas (4) para atingir o resultado final (a criação da visualização) estava adequado?

( ) Sim ( ) Não

8. Os recursos de ajuda da ferramenta (como: explicação sobre o tipo de dado, exemplo do formato de arquivo) foram úteis durante o processo de seleção e criação de visualizações?

( ) Sim ( ) Não

9. Foi fácil entender quais eram as tarefas?

( ) Sim ( ) Não

10. Você concordou com os gráficos sugeridos?

( ) Sim ( ) Não

11. A exibição de outros gráficos (além do gráfico sugerido pela ferramenta) auxiliaram no entendimento dos dados?

( ) Sim ( ) Não

**Comentários / Sugestões**

| |
|---|
| 12. Você autoriza a reprodução parcial ou total dos seus comentários e sugestões?<br>*Se você não escreveu comentário e/ou sugestões, selecione a opção "Não"*<br>( ) Sim ( ) Não |

**4. Identificar questões práticas**

Nesta etapa do framework DECIDE são considerados fatores como: perfil de participantes; ambiente para realizar a avaliação; seleção das tarefas; planejamento e preparação do material; alocação de pessoal, recursos e equipamentos (PRATES & BARBOSA, 2003).

No estudo de caso da ferramenta RVis foi definido que não haveria restrição ao perfil dos usuários. O objetivo das perguntas relacionadas ao perfil do participante (no questionário do Tabela 17) era identificar se o usuário conhecia a área de Visualização de Informação, se criava gráficos e se conhecia ou não outras ferramentas de criação de visualizações. Se algum participante respondesse "não" às perguntas, o mesmo poderia continuar respondendo às perguntas relacionadas à usabilidade e à recomendação da ferramenta RVis.

A tarefa do participante era construir visualizações com dados disponíveis na ferramenta RVis ou com os próprios dados e avaliar a usabilidade e funcionalidade da ferramenta respondendo um questionário. O participante realizou a avaliação de usabilidade e da recomendação da ferramenta RVis no seu próprio computador, especificamente em um navegador web, não sendo necessário alocar equipamentos ou algum local (como laboratório) para realizar a avaliação.

**5. Decidir como lidar com questões éticas**

ROGERS *et al.* (2011) propõem algumas diretrizes para assegurar que as avaliações de usabilidade sejam feitas de forma ética e para garantir a proteção dos direitos dos

participantes das avaliações. Essas diretrizes são (ROGERS *et al.*, 2011) (PRATES & BARBOSA, 2003):

- Informar aos participantes os objetivos do estudo e quais serão as atividades executadas.
- Deixar claro o tipo de dado coletado e como serão analisados.
- Informar aos participantes que dados particulares não serão divulgados.
- Certificar-se que os participantes saibam que podem interromper a avaliação a qualquer momento.
- Evitar incluir citações ou descrições que revelem a identidade dos participantes. Caso use alguma citação dos participantes, peça autorização prévia e de preferência mostre o trecho a ser divulgado.

No caso da avaliação da ferramenta RVis, no questionário (da Tabela 17) aplicado ao final do estudo de caso não foram incluídas perguntas que poderiam identificar o participante, como nome, endereço, sexo, idade. Na descrição do estudo de caso foi informado que a participação era anônima, conforme apresentado no Anexo A.

**6. Avaliar, interpretar e apresentar os dados**

Os dados coletados podem ser analisados de 3 formas: preditiva, interpretativa ou experimental (PRATES & BARBOSA, 2003). Na análise preditiva, com base em dados coletados de especialistas, os avaliadores tentam prever os tipos de problemas que os usuários enfrentarão ao utilizar determinada ferramenta. A análise interpretativa é realizada quando os avaliadores procuram explicar os fenômenos que ocorreram durante a interação do usuário com o sistema. Essa interação não ocorre dentro de ambientes controlados. A análise experimental ocorre quando dados coletados em ambientes controlados, como laboratórios,

precisam ser analisados em função de variáveis observadas e conhecidas (PRATES & BARBOSA, 2003).

Os dados coletados neste estudo de caso para avaliação da ferramenta RVis foram analisados de forma interpretativa, pois a interação do usuário com a ferramenta não foi realizada em ambiente controlado. Os participantes interagiram com a ferramenta através do navegador web do próprio computador.

O estudo de caso teve um total de 64 participantes. Sendo que em torno de 56% disseram que conheciam a área de Visualização de Informação (ou Visualização de Dados) e cerca de 89% dos participantes costumavam usar gráficos para gerar visualizações. Em torno de 41% conheciam outras ferramentas que auxiliassem na seleção e criação de visualizações. Do número de participantes que conheciam outras ferramentas, em torno de 46% utilizavam essas ferramentas frequentemente. A ferramenta mais conhecida e/ou utilizada foi o Excel (81%), conforme mostra a Figura 34. Em torno de 35% dos participantes conheciam outras ferramentas (além das ferramentas citadas no questionário) que eram: Sigma Plot[22], SPSS[23], Origin[24], LibreOffice Calc[25], Chart Tool[26], Matlab[27], Scilab[28], SciDAVis[29], MatlabPlot[30] e Splunk[31].

[22] https://systatsoftware.com/products/sigmaplot/
[23] http://www-03.ibm.com/software/products/pt/spss-statistics
[24] http://www.originlab.com/
[25] https://pt-br.libreoffice.org/descubra/calc/
[26] http://www.onlinecharttool.com/
[27] http://www.mathworks.com/products/matlab/
[28] http://www.scilab.org/
[29] http://scidavis.sourceforge.net/
[30] http://matplotlib.org/
[31] http://www.splunk.com/pt_br/products/splunk-enterprise.html

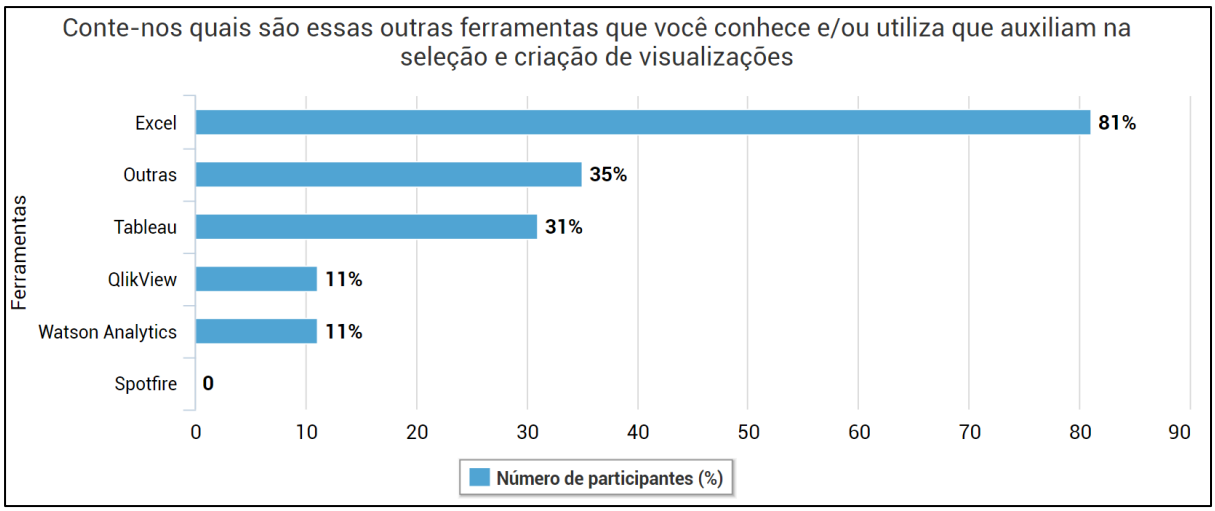


Figura 34. Ranking das ferramentas conhecidas e/ou utilizadas pelos participantes do primeiro estudo de caso.

A seguir serão apresentados e discutidos os resultados do estudo de caso relacionados à usabilidade e à recomendação da ferramenta RVis. Para analisar o resultado das questões relacionadas à usabilidade foram consideradas as heurísticas da Tabela 8.

Tabela 18. Heurísticas para avaliação do resultado das questões relacionadas à usabilidade.

| **Questão do Questionário da Tabela 17** | **Heurísticas** |
|---|---|
| 6 | O sistema deve sempre manter os usuários informados sobre o que está acontecendo através de feedback apropriado e no tempo certo (NIELSEN (1994).<br>Os elementos da interface não podem ser sutis e precisam ser destacados através de alguma distinção visual. |
| 7 | Número de ações mínimas para atingir uma meta ou tarefa (SCAPIN & BASTIEN (1997). |
| 8 e 9 | O sistema deve "falar" a linguagem do usuário, ou seja, apresentar palavras, frases e conceitos familiares e não técnicos (NIELSEN (1994). |

Segundo KRUG (2006), um dos objetivos da navegação em uma aplicação web é informar onde o usuário está. Além disso, uma das heurísticas de usabilidade proposta por NIELSEN (1994) afirma que o sistema deve sempre manter os usuários informados sobre o que está acontecendo através de *feedback* apropriado e no tempo certo. Na ferramenta RVis, conforme mostra a Figura 35, a barra de navegação tem a função de mostrar ao usuário em qual etapa o mesmo se encontra no processo de seleção e criação de visualizações. Conforme o usuário avança, a ferramenta destaca a etapa alcançada. KRUG (2006) afirma que os elementos da interface não podem ser sutis e precisam ser destacados através de alguma distinção visual, por exemplo, uma cor diferente e um texto em negrito (como ocorre na ferramenta, conforme mostra a Figura 35).

Na interface da ferramenta também foram inseridos botões para o usuário avançar nas etapas. Esses botões foram destacados dos demais elementos nas páginas (com azul de fundo e letra branca), conforme mostra a Figura 35. A interface da ferramenta foi projetada para facilitar a navegação dos usuários. Os resultados do estudo de caso comprovam que esse objetivo foi alcançado, pois em torno de 89% dos participantes disseram que foi fácil navegar pela interface da ferramenta, conforme mostra a Figura 36.

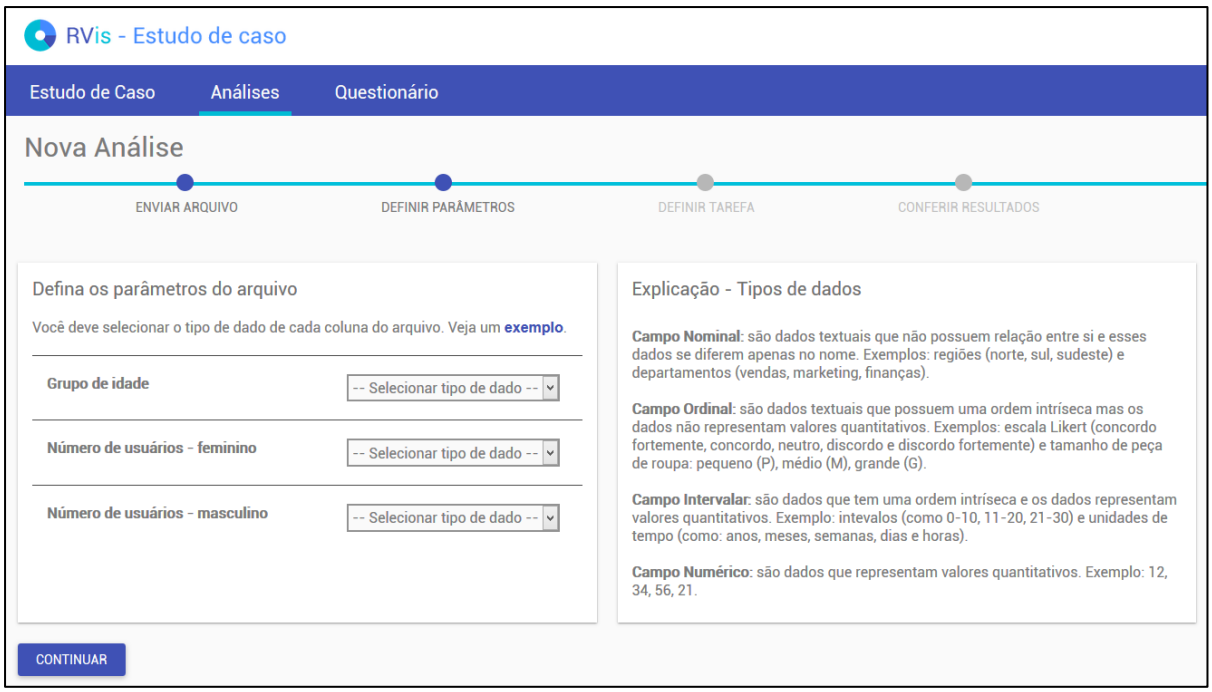


Figura 35. Elementos de destaque na interface da ferramenta RVis.

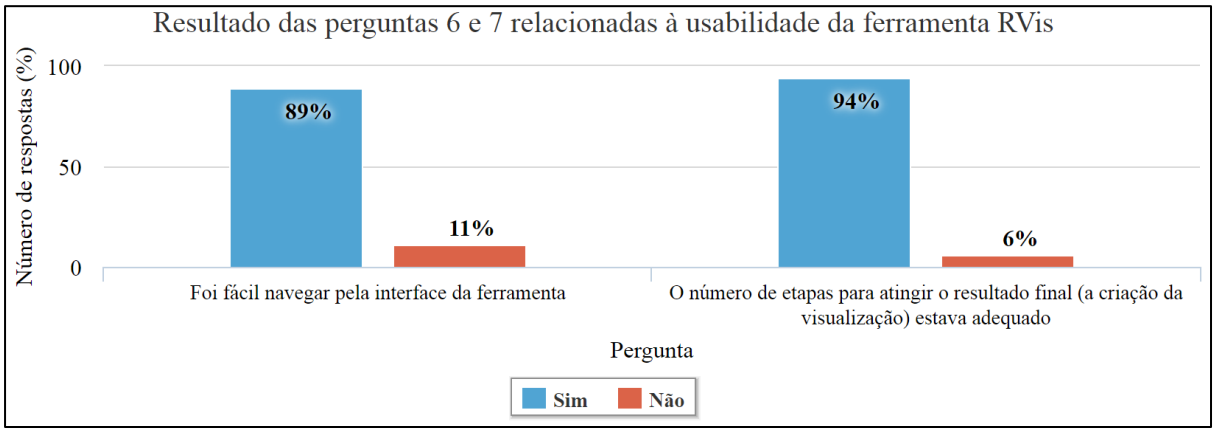


Figura 36. Resultado das perguntas 6 e 7relacionadas à usabilidade da ferramenta RVis.

Durante o planejamento da interface ficou definido que o processo de seleção e criação de visualizações seria realizado em 4 etapas. Essa definição está alinhada à heurística de ações mínimas proposta por SCAPIN & BASTIEN (1997). Essa heurística está relacionada ao número de ações para atingir uma meta ou tarefa (FORSELL & JOHANSSON, 2010). Conforme mostra a Figura 36, em torno de 94% dos participantes acharam adequado colocar 4 etapas para chegar ao resultado final que era a criação da visualização.

Vale ressaltar que a interface da ferramenta RVis foi alterada para o estudo de caso. Em relação às imagens apresentadas no capítulo 4, foram adicionados dois itens no menu principal, conforme mostra a Figura 37. O item “Estudo de Caso” exibia as instruções para o participante avaliar a ferramenta. O item “Questionário” direcionava o participante para a página do questionário. Além disso, a Figura 38 mostra que na página “Análises” foram adicionados exemplos de conjuntos de dados que o participante poderia utilizar durante o estudo de caso.

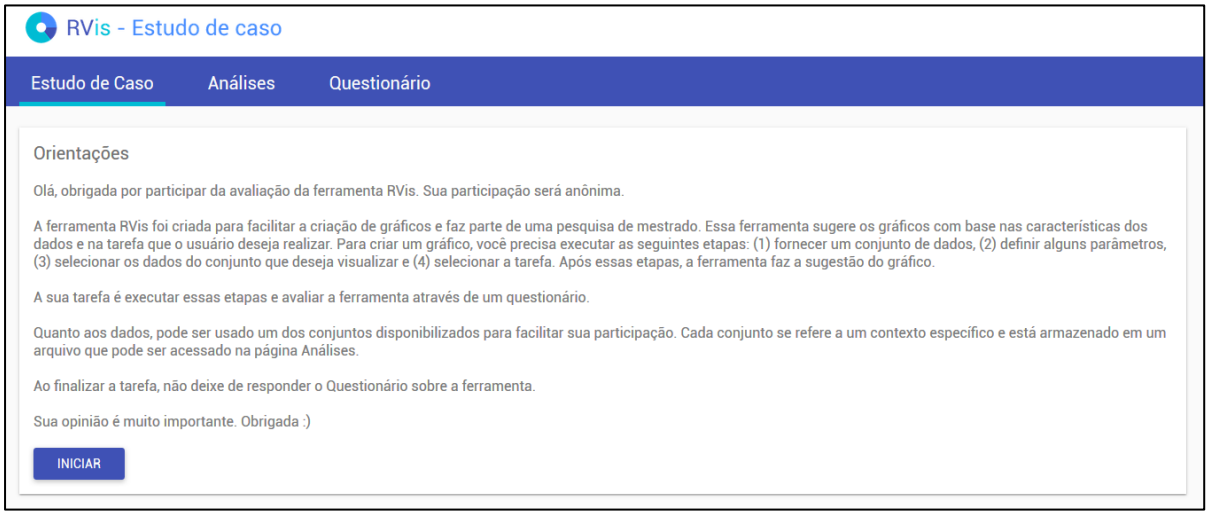


Figura 37. Interface da ferramenta RVis adaptada para o estudo de caso.

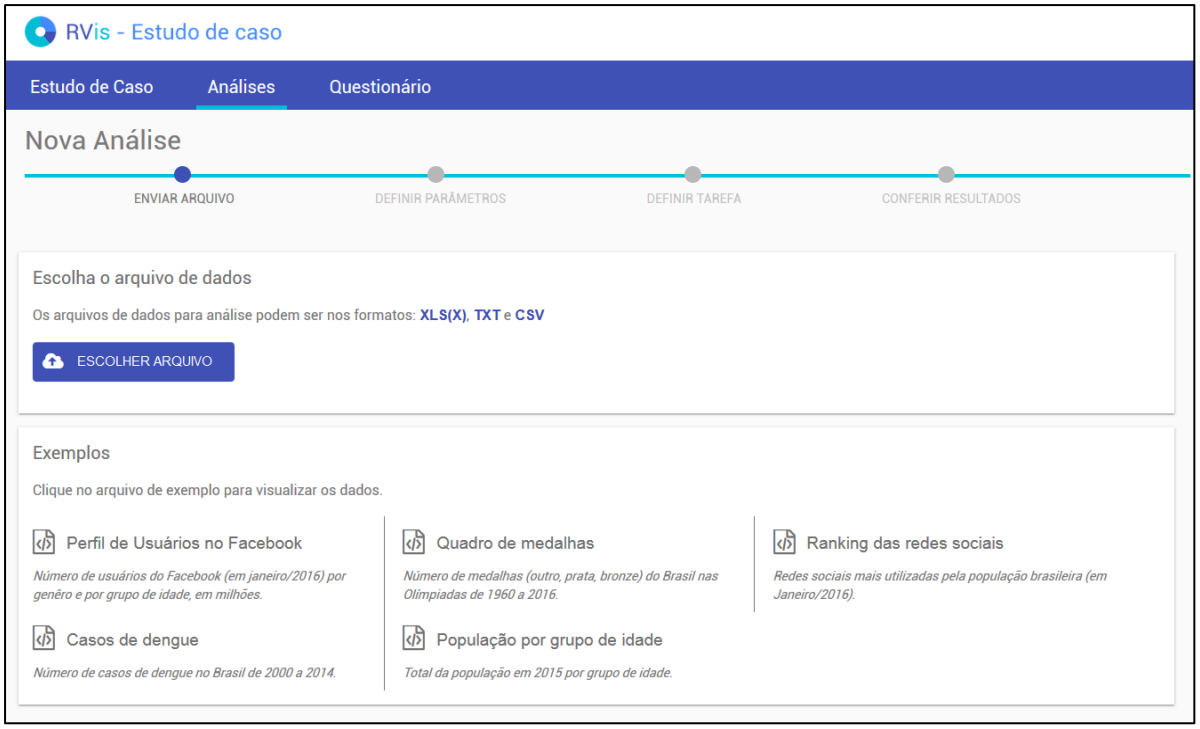


Figura 38. Página de análise da ferramenta RVis com exemplos de conjuntos de dados para o estudo de caso.

Outra heurística de usabilidade afirma que o sistema deve "falar" a linguagem do usuário, ou seja, o sistema deve apresentar palavras, frases e conceitos familiares e não técnicos (NIELSEN, 1994). Na segunda etapa do processo de seleção e criação de visualizações na ferramenta RVis, o usuário precisa definir o tipo de informação de cada coluna do conjunto de dados. Os tipos de informação são: campo nominal, ordinal, intervalar e numérico (conforme explicados na seção 2.2.1). Esses termos são técnicos para muitos usuários, no entanto, a ferramenta apresenta a definição e exemplos para facilitar o entendimento desses termos, conforme mostra a Figura 35. Outro recurso de ajuda da ferramenta é deixar disponível para *download*, na primeira etapa (envio do arquivo de dados), exemplos do formato de arquivo que a ferramenta aceita. Conforme mostra a Figura 39, para 92% dos participantes esses recursos de ajuda da ferramenta foram úteis durante a utilização da ferramenta.

Ainda em relação à heurística do sistema “falar” a linguagem do usuário, na página da terceira etapa (Definir tarefa), as tarefas que o usuário escolhia também foram descritas para facilitar o entendimento, pois muitos usuários poderiam não estar familiarizados com os termos. Cerca de 86% dos participantes afirmaram que foi fácil entender as tarefas, conforme mostra a Figura 39.

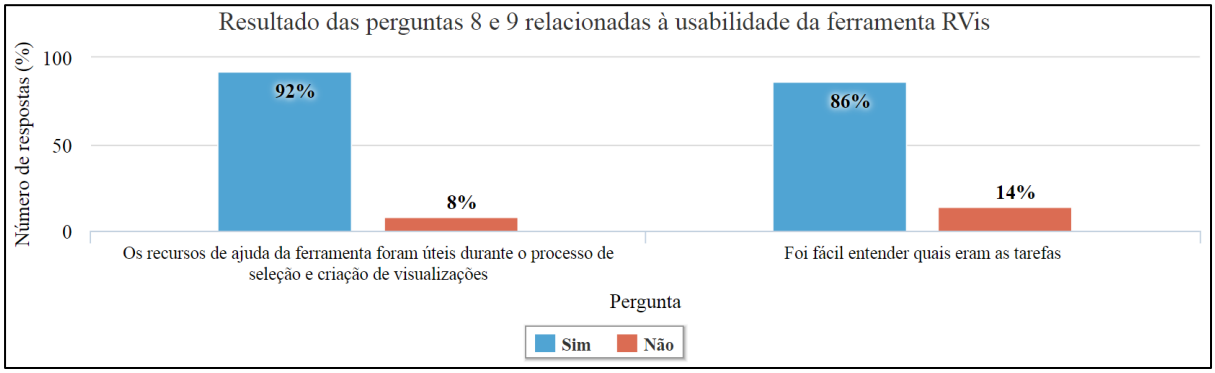


Figura 39. Resultado das perguntas 8 e 9 relacionadas à usabilidade da ferramenta RVis.

Haviam duas perguntas no questionário relacionadas à recomendação dos gráficos feita pela ferramenta. A finalidade da primeira pergunta foi verificar se o usuário concordava com o gráfico sugerido, ou seja, se o objetivo principal da ferramenta foi atingindo. Em torno de 91% dos participantes concordaram com os gráficos sugeridos, conforme mostra a Figura 40.

A finalidade da segunda pergunta era verificar se a apresentação de outros gráficos (além do gráfico sugerido) também auxiliava na análise dos dados. Conforme mostra a Figura 40, cerca de 86% dos participantes concordaram que os outros gráficos contribuíram no entendimento dos dados.

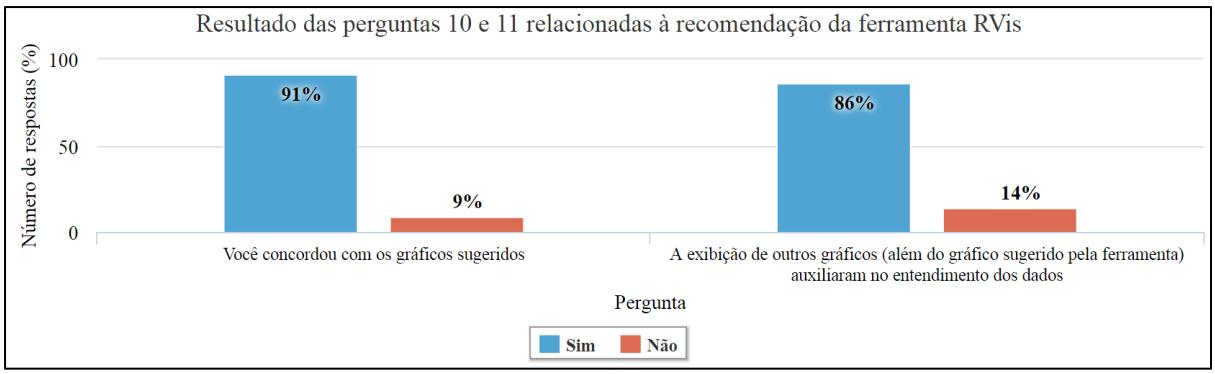


Figura 40. Resultado das perguntas 10 e 11 relacionadas à recomendação da ferramenta RVis.

Os participantes que discordavam da recomendação da ferramenta tinham a opção de sugerir outras técnicas. Foi criada uma nova regra para cada gráfico sugerido pelos participantes. Ao final do estudo de caso foram registradas 39 novas regras, ou seja, 39 sugestões dos participantes. Vale ressaltar que essas regras não foram inseridas no conjunto de regras da recomendação, apenas foram armazenadas no banco de dados para análises futuras.

Foi realizada uma comparação entre as regras sugeridas pelos participantes e as regras que não foram classificadas corretamente pelo algoritmo durante a etapa de aprendizado. O objetivo foi verificar se alguma regra formada a partir da sugestão do participante fazia parte do conjunto inicial de regras. O Anexo B apresenta as regras sugeridas pelos participantes.

Durante a comparação dessas regras, não foi encontrada nenhuma regra criada a partir do gráfico sugerido pelo usuário que já estivesse no conjunto de regras inicial. Uma evolução da ferramenta RVis é a utilização dessas regras para treinar e melhorar o desempenho do algoritmo, conforme descrito no capítulo a seguir.

Um segundo estudo de caso foi realizado com o objetivo de comparar a recomendação da ferramenta RVis com as ferramentas Microsoft Excel, Tableau, VizAssist, Voyager e Watson Analytics. As ferramentas ViSC e Exploration Views (citadas na seção 2.2.1) não foram avaliadas, pois não foi possível acessá-las. Além disso, essas ferramentas não tem a

opção para o usuário fazer upload de conjunto de dados. O usuário tem apenas a opção de usar conjunto de dados pré-definidos. O Many Eyes não foi testado, pois foi descontinuado.

Os dados do segundo estudo de caso foram coletados através de um questionário. Foi realizada uma análise interpretativa desses dados, pois a interação dos participantes com as 6 ferramentas não foi realizada em ambiente controlado. O Anexo C apresenta as instruções enviadas aos participantes (por *email*).

O estudo de caso teve um total de 7 participantes, sendo que todos conheciam outra ferramenta de criação de gráficos. Conforme mostra a Figura 41, todos os participantes já usaram o Excel. O Tableau já foi utilizado por cerca de 43% dos participantes e o Watson Analytics por cerca de 29%. A ferramenta citada no campo "Outros" foi o QlikView.

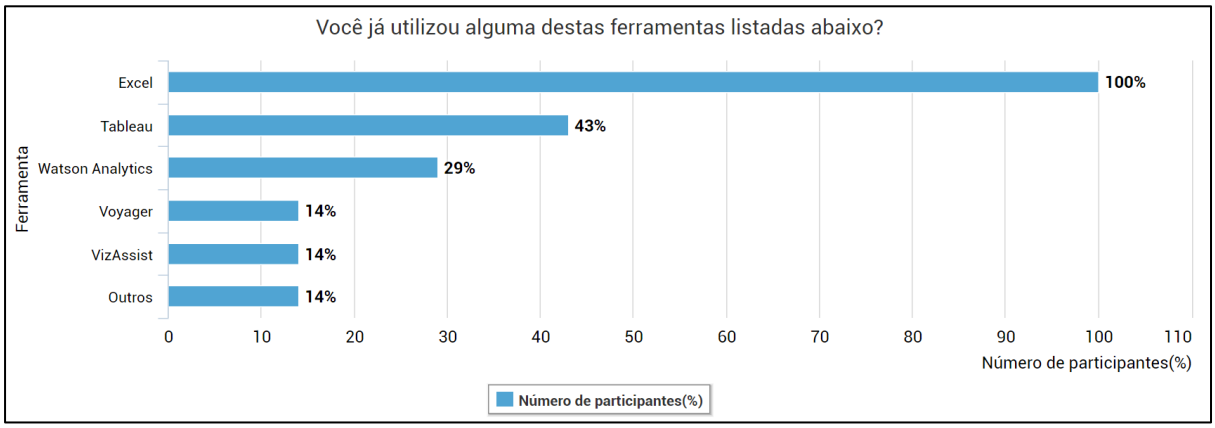


Figura 41. Ranking das ferramentas já utilizadas pelos participantes do segundo estudo de caso.

Neste estudo de caso, os participantes deveriam analisar as recomendações sugeridas pelas 6 ferramentas em duas tarefas e identificar se alguma ferramenta sugeriu o mesmo gráfico que a ferramenta RVis. A Figura 42 apresenta esse resultado, as recomendações que mais coincidiram com a sugestão da RVis foram aquelas elaboradas pelo Tableau na tarefa 1 (57%) e as sugeridas pelo VizAssist na tarefa 2 (57%).

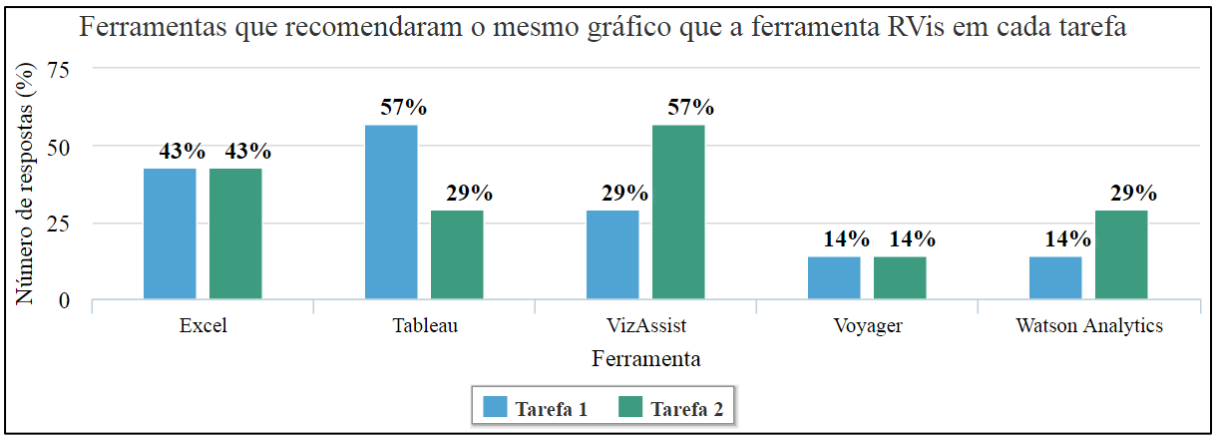


Figura 42. Ferramentas com recomendações iguais em cada tarefa.

Para os casos em que as recomendações foram idênticas quando comparadas as sugestões do RVis e de outra ferramenta, foi investigado o aspecto relacionado à facilidade de utilização. Em torno de 86% dos participantes afirmaram que foi mais fácil utilizar a RVis para realizar a tarefa 1, enquanto que para 57% a RVis foi mais fácil de utilizar para a tarefa 2. Estes resultados indicam que a RVis apresenta alguma vantagem em relação às outras ferramentas analisadas seja pela adequação da recomendação ou pela facilidade de uso.

# 6. Conclusão e Trabalhos Futuros

## 6.1 Considerações Finais

O aumento do volume de dados que estão sendo gerados exige mecanismos que possibilitem o processamento de conjuntos de dados cada vez maiores. Além disso, há uma demanda por métodos que sejam capazes de auxiliar a descoberta de conhecimento. As técnicas de Visualização de Informação, por exemplo, representam um desses mecanismos que auxiliam na representação dos dados e na descoberta de novos conhecimentos.

Os sistemas de recomendação podem ampliar essas possibilidades ao realizar a combinação não trivial das informações, o que pode retornar para o usuário informação nova, adequada, útil, que não seria alcançada através da avaliação humana dos dados. Em meio à dificuldade dos usuários selecionarem a melhor opção dentre uma grande variedade de alternativas que solucionam seu problema, sistemas de recomendação podem auxiliar nessa tarefa.

Esta dissertação está alinhada a este contexto e os objetivos foram: definir um conjunto de regras para a classificação e recomendação de técnicas de visualização de informação e elaborar uma ferramenta para apoiar a seleção e criação de visualizações através da recomendação. O mecanismo definido para a recomendação foi baseado na classificação das técnicas de visualização considerando os aspectos que influenciam essa escolha, como a tarefa e as características do conjunto de dados.

Para avaliar a usabilidade e a recomendação da ferramenta RVis, foi realizado um estudo de caso que mostrou que a abordagem de recomendação através da classificação das

técnicas de visualização pode auxiliar os usuários na criação do gráfico mais adequado para seus dados e para a tarefa que deseja realizar.

## 6.2 Contribuições

A principal contribuição deste trabalho foi empregar técnicas de aprendizado de máquina, especificamente os modelos de classificação, para refinar as regras que melhor representam a interpretação humana no contexto de classificação de técnicas de visualização de informação. As regras foram definidas a partir de um estudo da literatura. Os resultados dos testes ao aplicar as técnicas de aprendizado de máquina indicaram que a árvore de decisão (isto é, o algoritmo C4.5 (QUINLAN, 1993)) e a rede neural do tipo MLP (HAN *et al.*, 2011a) obtiveram o melhor desempenho na classificação das regras. As alterações realizadas durante os testes permitiram melhorar a representação das regras de recomendação e isso refletiu na taxa de acerto dos algoritmos.

O algoritmo escolhido para classificar e recomendar as técnicas de visualização foi a árvore de decisão (especificamente o algoritmo C4.5 (QUINLAN, 1993)). O uso desse algoritmo é um diferencial da ferramenta RVis em relação às ferramentas que também sugerem gráficos. A taxa de acerto desse algoritmo permitiu que a ferramenta RVis atendesse às expectativas dos usuários em relação às visualizações sugeridas, conforme mostram os resultados do primeiro estudo de caso.

## 6.3 Limitações

Uma das limitações deste trabalho se refere ao escopo definido, no qual não são consideradas as técnicas multidimensionais e as técnicas para representação de dados

geográficos (latitude e longitude) e de dados puramente textuais. Essa limitação poderá ser resolvida com a criação de regras para recomendação e classificação dessas técnicas.

## 6.4 Trabalhos Futuros

Uma vertente de evolução será analisar as técnicas sugeridas pelos usuários no estudo de caso para inferir novas regras e melhorar o aprendizado do algoritmo. A análise das novas regras deverá ser feita com especialistas na área e/ou com outro estudo de caso para saber a opinião de outros usuários, se concordam ou não com as regras. A partir dos resultados obtidos, as novas regras serão incorporadas ao conjunto de treinamento para realizar o novo aprendizado do algoritmo. Nesse aprendizado será verificado se a taxa de acerto continua igual ou superior à taxa de acerto no treinamento realizado anteriormente. O novo conjunto de regras também poderá ser testado com outros algoritmos e dependendo do resultado, a ferramenta terá outro algoritmo para recomendação.

Outra evolução do trabalho é considerar outras informações de contexto na definição das regras para recomendação das visualizações. É necessário investigar, além das tarefas executadas, outros elementos do contexto relevantes para classificação das técnicas de visualização. Um desses elementos pode ser o perfil do usuário da visualização, ou seja, as características do público-alvo. Outro elemento que pode auxiliar na recomendação é o objetivo da visualização que pode variar entre exploratório (para descoberta do conjunto de dados) ou explanatório (para comunicação do conjunto de dados). Para visualizações exploratórias é importante disponibilizar recursos para o usuário interagir com a visualização, como: filtros, exibição do histórico de ações executadas na visualização e seleção de subconjuntos dos dados originais.

# 7. Bibliografia


AGARWAL, B., MITTAL, N., 2014. "Text Classification Using Machine Learning Methods-A Survey". In: BABU, B. V., NAGAR, Atulya, DEEP, Kusum, PANT, Millie, BANSAL, Jagdish Chand, RAY, Kanad & GUPTA, Umesh (eds.), Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. S.l.: Springer India. Advances in Intelligent Systems and Computing, 236. pp. 701–709.

AIGNER, W., MIKSCH, S., SCHUMANN, H., et al., 2011. "Survey of Visualization Techniques". In: Visualization of Time-Oriented Data. S.l.: Springer London. Human-Computer Interaction Series. pp. 147–254.

AMAR, R., EAGAN, J., STASKO, J., 2005. "Low-level components of analytic activity in information visualization". In: IEEE Symposium on Information Visualization, 2005. INFOVIS 2005. S.l.: s.n. Outubro 2005. pp. 111–117.

BORKIN, M., VO, A., BYLINSKII, Z., et al., 2013, "What Makes a Visualization Memorable?". In: IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis 2013).

BOUALI, F., GUETTALA, A., VENTURINI, G., 2015, "VizAssist: an interactive user assistant for visual data mining". In: The Visual Computer. pp. 1–17.

CAIRO, A., 2012, The Functional Art: An introduction to information graphics and visualization. . S.l., New Riders.

CARD, S., 2003. "Information visualization". In: JACKO, Julie A. & SEARS, Andrew (eds.), The Human-computer Interaction Handbook. Hillsdale, NJ, USA: L. Erlbaum Associates Inc. pp. 544–582.

CARD, S.K., MACKINLAY, J., 1997. "The structure of the information visualization design space". In: Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis '97). Washington, DC, USA: IEEE Computer Society. 1997. pp. 92–.

CARD, Stuart K., MACKINLAY, Jock D. & SHNEIDERMAN, Ben (eds.), 1999, Readings in information visualization: using vision to think. . San Francisco, CA, USA, Morgan Kaufmann Publishers Inc.

CASNER, S.M., 1991, "Task-analytic Approach to the Automated Design of Graphic Presentations". In: ACM Trans. Graph. v. 10, pp. 111–151.

DIAS, M.M., 2008, "Parâmetros na escolha de técnicas e ferramentas de mineração de dados". In: Acta Scientiarum. Technology. v. 24, pp. 1715–1725.

DICIO, 2016. DICIO. Disponível em: <http://www.dicio.com.br/>. Acessado em: 21 Abril 2016.

DIX, A., 2013. "Introduction to Information Visualisation". In: AGOSTI, Maristella, FERRO, Nicola, FORNER, Pamela, MÜLLER, Henning & SANTUCCI, Giuseppe (eds.), Information Retrieval Meets Information Visualization. S.l.: Springer Berlin Heidelberg. Lecture Notes in Computer Science, 7757. pp. 1–27.

DO, P., 2015. Disponível em: <https://vida.io/documents/ZCzewTza4ZSzMWSBG>. Acessado em: 15 Abril 2016.

DO NASCIMENTO, H.A., FERREIRA, C.B., 2005. "Visualização de Informações–uma abordagem prática". In: XXV Congresso da Sociedade Brasileira de Computação, XXIV JAI. UNISINOS, S. Leopoldo–RS. S.l.: s.n. 2005.

DREISEITL, S., OHNO-MACHADO, L., KITTLER, H., et al., 2001, "A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions". In: Journal of Biomedical Informatics. v. 34, pp. 28–36.

ELIAS, M., BEZERIANOS, A., 2011. "Exploration Views: Understanding Dashboard Creation and Customization for Visualization Novices". In: CAMPOS, Pedro, GRAHAM, Nicholas, JORGE, Joaquim, NUNES, Nuno, PALANQUE, Philippe & WINCKLER, Marco (eds.), Human-Computer Interaction – INTERACT 2011. S.l.: Springer Berlin Heidelberg. Lecture Notes in Computer Science, 6949. pp. 274–291.

ELKAN, C., 1997. Boosting And Naive Bayesian Learning. S.l.

FEW, S., 2012, Show Me the Numbers: Designing Tables and Graphs to Enlighten. . Second edition. Burlingame, Calif., Analytics Press.

FORSELL, C., JOHANSSON, J., 2010. "An heuristic set for evaluation in information visualization". In: Proceedings of the International Conference on Advanced Visual Interfaces. New York, NY, USA: ACM. 2010. pp. 199–206.

FREITAS, C.M.D.S., CHUBACHI, O.M., LUZZARDI, P.R.G., et al., 2001, "Introdução à visualização de informações". In: .

FREYNE, J., SMYTH, B., 2010. "Creating Visualizations: A Case-Based Reasoning Perspective". In: COYLE, Lorcan & FREYNE, Jill (eds.), Artificial Intelligence and Cognitive Science. S.l.: Springer Berlin Heidelberg. Lecture Notes in Computer Science, 6206. pp. 82–91.

GOTZ, D., WEN, Z., 2009. "Behavior-driven Visualization Recommendation". In: Proceedings of the 14th International Conference on Intelligent User Interfaces. New York, NY, USA: ACM. 2009. pp. 315–324.

HAN, J., KAMBER, M., PEI, J., 2011a. "Classification by Backpropagation". In: Data Mining: Concepts and Techniques. 3rd. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. pp. 327–336.

HAN, J., KAMBER, M., PEI, J., 2011b, Data Mining: Concepts and Techniques. . 3rd. San Francisco, CA, USA, Morgan Kaufmann Publishers Inc.

HAN, J., KAMBER, M., PEI, J., 2011c. "k-Nearest-Neighbor Classifiers". In: Data Mining: Concepts and Techniques. 3rd. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. pp. 348–350.

HANRAHAN, P., 2006. "VizQL: A Language for Query, Analysis and Visualization". In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. New York, NY, USA: ACM. 2006. pp. 721–721.

HARDIN, M., HOM, D., PEREZ, R., et al., 2012. "Which chart or graph is right for you?". In: VizWiz. Disponível em: <http://vizwiz.blogspot.com.br/2012/02/which-chart-or-graph-is-right-for-you.html>. Acessado em: 23 Junho 2014.

HOFFMAN, P.E., GRINSTEIN, G.G., 2002. "Information Visualization in Data Mining and Knowledge Discovery". In: FAYYAD, Usama, GRINSTEIN, Georges G. & WIERSE, Andreas (eds.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. pp. 47–82.

IBM, 2015a. Disponível em: <https://watson.analytics.ibmcloud.com/>. Acessado em: 8 Dezembro 2015.

IBM, 2015b. Disponível em: <http://www-969.ibm.com/software/analytics/manyeyes/>. Acessado em: 8 Dezembro 2015.

ILIINSKY, N., STEELE, J., 2011, Designing Data Visualizations: Intentional Communication from Data to Display. . S.l., s.n. Acessado em: 3 Dezembro 2015.

JELEN, B., 2013, Excel 2013 Charts and Graphs. . S.l., Que Publishing.

KEIM, D.A., 2002, "Information visualization and visual data mining". In: IEEE Transactions on Visualization and Computer Graphics. v. 8, pp. 1–8.

KEIM, D.A., MANSMANN, F., SCHNEIDEWIND, J., et al., 2006. "Challenges in Visual Data Analysis". In: Tenth International Conference on Information Visualization, 2006. IV 2006. S.l.: s.n. 2006. pp. 9–16.

KEY, A., HOWE, B., PERRY, D., et al., 2012. "VizDeck: Self-organizing Dashboards for Visual Analytics". In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. New York, NY, USA: ACM. 2012. pp. 681–684.

KOSARA, R., 2010. "Understanding Pie Charts". In: eagereyes. Disponível em: <https://eagereyes.org/techniques/pie-charts>. Acessado em: 25 Janeiro 2016.

KOTSIANTIS, S.B., ZAHARAKIS, I.D., PINTELAS, P.E., 2007, "Machine learning: a review of classification and combining techniques". In: Artificial Intelligence Review. v. 26, pp. 159–190.

KRUG, S., 2006, Não me faça pensar!: uma abordagem de bom senso à usabilidade na web. . S.l., Alta Books.

LENGLER, R., EPPLER, M.J., 2007a. "Towards a periodic table of visualization methods of management". In: Proceedings of the IASTED International Conference on Graphics and Visualization in Engineering. Anaheim, CA, USA: ACTA Press. 2007. pp. 83–88.

LENGLER, R., EPPLER, M.J., 2007b. VISUAL LITERACY: AN E-LEARNING TUTORIAL ON VISUALIZATION FOR COMMUNICATION, ENGINEERING AND BUSINESS. Disponível em: <http://www.visual-literacy.org/periodic_table/periodic_table.html>. Acessado em: 6 Novembro 2013.

LUZZARDI, P.R.G., 2003. Critérios de avaliação de técnicas de visualização de informações hierárquicas. . Tese de Doutorado. Porto Alegre: Universidade Federal do Rio Grande do Sul.

MACKINLAY, J., 1986, "Automating the design of graphical presentations of relational information". In: ACM Trans. Graph. v. 5, pp. 110–141.

MACKINLAY, J., HANRAHAN, P., STOLTE, C., 2007, "Show Me: Automatic Presentation for Visual Analysis". In: IEEE Transactions on Visualization and Computer Graphics. v. 13, pp. 1137–1144.

MUNZNER, T., 2014, Visualization Analysis and Design. . S.l., CRC Press.

NGUYEN, T.T.T., ARMITAGE, G., 2008, "A survey of techniques for internet traffic classification using machine learning". In: IEEE Communications Surveys Tutorials. v. 10, pp. 56–76.

NIELSEN, J., 1994. "Usability Inspection Methods". In: Conference Companion on Human Factors in Computing Systems. New York, NY, USA: ACM. 1994. pp. 413–414.

DE OLIVEIRA, M.C.F., LEVKOWITZ, H., 2003, "From visual data exploration to visual data mining: a survey". In: IEEE Transactions on Visualization and Computer Graphics. v. 9, pp. 378–394.

PRATES, R.O., BARBOSA, S.D.J., 2003, "Avaliação de Interfaces de Usuário – Conceitos e Métodos". In: XXIII Congresso Nacional da Sociedade Brasileira de Computação. XXII Jornadas de Atualização em Informática (JAI).

PURCHASE, H.C., ANDRIENKO, N., JANKUN-KELLY, T.J., et al., 2008. "Theoretical Foundations of Information Visualization". In: KERREN, Andreas, STASKO, John T., FEKETE, Jean-Daniel & NORTH, Chris (eds.), Information Visualization. S.l.: Springer Berlin Heidelberg. Lecture Notes in Computer Science, 4950. pp. 46–64.

QUINLAN, J.R., 1993, C4.5: Programs for Machine Learning. . San Francisco, CA, USA, Morgan Kaufmann Publishers Inc.

REZENDE, S.O., 2003, Sistemas inteligentes: fundamentos e aplicações. . S.l., Editora Manole Ltda.

RIBECCA, S., 2016. Disponível em: <http://www.datavizcatalogue.com/index.html>. Acessado em: 15 Outubro 2014.

RIBEIRO, F.C., CAETANO, B.P., PAULA, M.M.V. DE, et al., 2016. "Keep calm and visualize your data: minicurso de Visualização de Dados". In: Tópicos em Sistemas de Informação: Minicursos SBSI 2016. S.l.: s.n. pp. 31–52.

ROGERS, Y., SHARP, H., PREECE, J., 2011, Interaction Design: Beyond Human-Computer Interaction. . Chichester, West Sussex, U.K, John Wiley & Sons.

ROTH, S.F., KOLOJEJCHICK, J., MATTIS, J., et al., 1994. "Interactive Graphic Design Using Automatic Presentation Knowledge". In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM. 1994. pp. 112–117.

SCAPIN, D.L., BASTIEN, J.M.C., 1997, "Ergonomic criteria for evaluating the ergonomic quality of interactive systems". In: Behaviour & Information Technology. v. 16, pp. 220–231.

SHI, Z., 1992, Principles of Machine Learning. . 2. Beijing, International Academic Publishers.

SHNEIDERMAN, B., 1996. "The eyes have it: a task by data type taxonomy for information visualizations". In: , IEEE Symposium on Visual Languages, 1996. Proceedings. S.l.: s.n. 1996. pp. 336–343.

SKAU, D., 2012. "Best Practices: Maximum Elements For Different Visualization Types". In: Visually Blog. Disponível em: <http://blog.visual.ly/maximum-elements-for-visualization-types/>. Acessado em: 22 Janeiro 2016.

DE SOUSA, T.A.F., 2013. Sistema de recomendação para apoiar a construção de gráficos com dados estatísticos. . Dissertação de Mestrado. S.l.: PUC-RJ.

DE SOUSA, T.A.F., BARBOSA, S.D.J., 2013. "Sistema de recomendação para apoiar a construção de gráficos com dados estatísticos". In: Proceedings of the 12th Brazilian Symposium on Human Factors in Computing Systems. Porto Alegre, Brazil, Brazil: Brazilian Computer Society. 2013. pp. 168–177.

STOLTE, C., TANG, D., HANRAHAN, P., 2002, "Polaris: a system for query, analysis, and visualization of multidimensional relational databases". In: IEEE Transactions on Visualization and Computer Graphics. v. 8, pp. 52–65.

TABLEAU, 2016. TABLEAU SOFTWARE. Disponível em: <http://www.tableau.com/pt-br>. Acessado em: 20 Fevereiro 2016.

TABLEAU SOFTWARE, 2016. Disponível em: <https://onlinehelp.tableau.com/current/pro/desktop/en-us/help.html#buildauto_showme.html>. Acessado em: 30 Agosto 2016.

TAN, P.-N., STEINBACH, M., KUMAR, V., 2005, Introduction to Data Mining, (First Edition). . Boston, MA, USA, Addison-Wesley Longman Publishing Co., Inc.

TUFTE, E.R., 2001, The Visual Display of Quantitative Information. . 2nd edition. Cheshire, Conn, Graphics Pr.

VALIATI, E.R. DE A., 2008. Avaliação de usabilidade de técnicas de visualização de informações multidimensionais. . Tese de Doutorado. Porto Alegre: Universidade Federal do Rio Grande do Sul.

VIEGAS, F.B., WATTENBERG, M., VAN HAM, F., et al., 2007, "ManyEyes: A Site for Visualization at Internet Scale". In: IEEE Transactions on Visualization and Computer Graphics. v. 13, pp. 1121–1128.

VOIGT, M., PIETSCHMANN, S., GRAMMEL, L., et al., 2012. "Context-aware Recommendation of Visualization Components". In: eKNOW 2012, The Fourth International Conference on Information, Process, and Knowledge Management. S.l.: s.n. 30 Janeiro 2012. pp. 101–109.

WARE, C., 2004, Information visualization: perception for design. . 2. San Francisco, CA, Morgan Kaufman.

WITTEN, I.H., FRANK, E., HALL, M.A., 2011, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition. . 3 edition. Burlington, MA, Morgan Kaufmann.

WONGSUPHASAWAT, K., MORITZ, D., ANAND, A., et al., 2016, "Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations". In: IEEE Transactions on Visualization and Computer Graphics. v. 22, pp. 649–658.

YAMAGUCHI, J.K., 2010. Diretrizes para a escolha de técnicas de visualização aplicadas no processo de extração do conhecimento. . Dissertação de Mestrado. Universidade Estadual de Maringá (UEM) - Paraná: Programa de Pós-Graduação em Ciência da Computação.

YANG, H., LI, Y., ZHOU, M.X., 2014a, "Understand Users' Comprehension and Preferences for Composing Information Visualizations". In: ACM Trans. Comput.-Hum. Interact. v. 21, pp. 6:1–6:30.

YANG, H., LI, Y., ZHOU, M.X., 2014b, "Understand Users' Comprehension and Preferences for Composing Information Visualizations". In: ACM Trans. Comput.-Hum. Interact. v. 21, pp. 6:1–6:30.

ZHOU, M.X., CHEN, M., 2003. "Automated Generation of Graphic Sketches by Example". In: Proceedings of the 18th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 2003. pp. 65–71.

# 9. Anexos

## Anexo A

Texto descritivo sobre o estudo de caso que avaliou a ferramenta RVis.

| **Texto explicativo sobre o primeiro estudo de caso** |
| --- |
| Olá, obrigada por participar da avaliação da ferramenta RVis. Sua participação é anônima.<br>A ferramenta RVis foi criada para auxiliar os usuários na escolha e criação de gráficos e faz parte de uma pesquisa de mestrado. Essa ferramenta sugere os gráficos com base nas características dos dados e na tarefa que o usuário deseja realizar com esses dados. Para chegar ao resultado final, ou seja, na criação do gráfico, o usuário precisa executar as seguintes etapas: (1) fazer upload de um conjunto de dados, (2) definir alguns parâmetros, (3) selecionar os dados que deseja visualizar e a tarefa que deseja realizar com os dados. Após essas 3 etapas, a ferramenta faz a sugestão do gráfico.<br>A sua tarefa é executar as etapas descritas anteriormente, ou seja, utilizar a ferramenta para criar um gráfico e avaliar a usabilidade e a funcionalidade da ferramenta. Em relação ao conjunto de dados para upload, são fornecidos alguns conjuntos de exemplo. Clique no menu "Exemplos de Conjunto de Dados" para ter acesso a esses exemplos.<br>Você pode utilizar outros conjuntos de dados. Mas esses conjuntos precisam estar no formato aceito pela ferramenta. Na página principal tem exemplos de como seu conjunto de dados precisa estar configurado para que a ferramenta consiga ler os dados.<br>Ao finalizar a análise dos dados, clique em "Questionário" para responder algumas perguntas sobre a ferramenta. Sua opinião é muito importante. Obrigada :) |

# Anexo B

A Tabela 19 mostra as regras criadas a partir dos gráficos sugeridos pelos participantes do estudo de caso, conforme descrito no capítulo.

Tabela 19. Novas regras criadas a partir das sugestões dos participantes do estudo de caso.

| ID | Unidade de Medida | Quanti-dade mínima de dados | Quanti-dade máxima de dados | Valor negativo | Quanti-dade de atributos | Quantidade de atributos quantitativos | Quantidade de atributos categóricos | Tipo de atributo categórico | Tarefa | Gráfico sugerido |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | >=6 | <=15 | Falso | 3 | 2 | 1 | Nominal | Comparação Nominal | *Circle Packing* |
| 2 | 1 | >=6 | <=15 | Falso | 3 | 2 | 1 | Nominal | Parte-Todo | Gráfico de Barras |
| 3 | 1 | >=6 | <=15 | Falso | 2 | 1 | 1 | Nominal | Parte-Todo | Gráfico de Barras |
| 4 | 1 | >=6 | <=15 | Falso | 2 | 1 | 1 | Nominal | Parte-Todo | Gráfico de Pizza |
| 5 | 1 | >=6 | <=15 | Falso | 3 | 2 | 1 | Nominal | Correlação | Gráfico de Linhas |
| 6 | 1 | >=1 | <=5 | Falso | 3 | 2 | 1 | Nominal | Comparação | Gráfico de |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Nominal | Barras |
| 7 | 1 | >=1 | <=5 | Falso | 3 | 2 | 1 | Nominal | Comparação Nominal | Gráfico de Colunas Empilhadas |
| 8 | 1 | >=1 | <=5 | Falso | 3 | 2 | 1 | Nominal | Comparação Nominal | Treemap |
| 9 | 1 | >=6 | <=15 | Falso | 2 | 1 | 1 | Nominal | Série Temporal | Gráfico de Área |
| 10 | 1 | >=1 | <=5 | Falso | 3 | 2 | 1 | Nominal | Ranqueamento | Treemap |
| 11 | 1 | >=1 | <=5 | Falso | 3 | 2 | 1 | Nominal | Comparação Nominal | Vários Gráficos de Colunas |
| 12 | 1 | >=6 | <=15 | Falso | 2 | 1 | 1 | Intervalar | Distribuição | Gráfico de Barras |
| 13 | 1 | >=6 | <=15 | Falso | 3 | 2 | 1 | Nominal | Série Temporal | Gráfico de Barras |
| 14 | 1 | >=6 | <=15 | Falso | 3 | 2 | 1 | Nominal | Série Temporal | Gráfico de Colunas |
| 15 | 1 | >=6 | <=15 | Falso | 3 | 2 | 1 | Nominal | Série Temporal | Gráfico de Área |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 1 | >=6 | <=15 | Falso | 3 | 2 | 1 | Nominal | Série Temporal | Treemap |
| 17 | 1 | >=1 | <=5 | Falso | 2 | 1 | 1 | Nominal | Ranqueamento | Gráfico de Colunas |
| 18 | 1 | >=1 | <=5 | Falso | 2 | 1 | 1 | Ordinal | Parte-Todo | Gráfico de Pizza |
| 19 | 1 | >=1 | <=5 | Falso | 3 | 2 | 1 | Nominal | Comparação Nominal | Gráfico de Barras |
| 20 | 1 | >=1 | <=5 | Falso | 3 | 2 | 1 | Nominal | Comparação Nominal | Gráfico de Barras Empilhadas |
| 21 | 1 | >=1 | <=5 | Falso | 3 | 2 | 1 | Nominal | Ranqueamento | Gráfico de Colunas |
| 22 | 1 | >=1 | <=5 | Falso | 3 | 2 | 1 | Nominal | Ranqueamento | Vários Gráficos de Colunas |
| 23 | 1 | >=1 | <=5 | Falso | 2 | 1 | 1 | Intervalar | Comparação Nominal | Gráfico de Barras |
| 24 | 1 | >=6 | <=15 | Falso | 5 | 4 | 1 | Nominal | Parte-Todo | Gráfico de Colunas |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 25 | 1 | >=6 | <=15 | Falso | 2 | 1 | 1 | Nominal | Parte-Todo | Gráfico de Barras |
| 26 | 1 | >=6 | <=15 | Falso | 5 | 4 | 1 | Nominal | Série Temporal | Gráfico de Colunas |
| 27 | 1 | >=6 | <=15 | Falso | 2 | 1 | 1 | Nominal | Parte-Todo | Gráfico de Barras |
| 28 | 1 | >=6 | <=15 | Falso | 3 | 2 | 1 | Intervalar | Distribuição | Gráfico de Colunas |
| 29 | 1 | >=6 | <=15 | Falso | 3 | 2 | 1 | Nominal | Parte-Todo | Vários Gráficos de Colunas |
| 30 | 1 | >=16 | <=20 | Falso | 2 | 1 | 1 | Nominal | Parte-Todo | Gráfico de Barras |
| 31 | 1 | >=16 | <=20 | Falso | 2 | 1 | 1 | Nominal | Parte-Todo | Gráfico de Linhas |
| 32 | 1 | >=16 | <=20 | Falso | 2 | 1 | 1 | Nominal | Parte-Todo | Gráfico de Áreas |
| 33 | 1 | >=16 | <=20 | Falso | 2 | 1 | 1 | Nominal | Parte-Todo | Gráfico de Pizza |
| 34 | 1 | >=16 | <=20 | Falso | 2 | 1 | 1 | Nominal | Comparação | Gráfico de |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Nominal | Linhas |
| 35 | 1 | >=16 | <=20 | Falso | 2 | 1 | 1 | Nominal | Comparação Nominal | Gráfico de Áreas |
| 36 | 1 | >=16 | <=20 | Falso | 2 | 1 | 1 | Nominal | Comparação Nominal | Gráfico de Pizza |
| 37 | 1 | >=6 | <=15 | Falso | 5 | 4 | 1 | Intervalar | Parte-Todo | Gráfico de Colunas |
| 38 | 1 | >=6 | <=15 | Falso | 2 | 1 | 1 | Nominal | Série Temporal | Gráfico de Colunas |
| 39 | 1 | >=6 | <=15 | Falso | 2 | 1 | 1 | Intervalar | Parte-Todo | Gráfico de Barras |

# Anexo C

Texto descritivo sobre o segundo estudo de caso que comparou e avaliou 6 ferramentas: RVis, Microsoft Excel, Tableau, VizAssist, Voyager e Watson Analytics.

| **Texto explicativo sobre o segundo estudo de caso** |
|---|
| Olá, obrigada por participar do estudo de caso. Sua participação é anônima.<br>O objetivo é comparar a ferramenta RVis com as ferramentas VizAssist, Voyager, Watson Analytics, Excel e Tableau.<br>Sua tarefa é criar visualizações com o auxílio dessas ferramentas. Ao final, responda o questionário disponível em https://goo.gl/forms/dC9J0Rq0P5cT2TxH3.<br>Sua opinião é muito importante. Obrigada!<br><br>Para usar a ferramenta RVis, acesse http://rvis-fvis.rhcloud.com/rvis/index.xhtml<br>Para usar a ferramenta Watson Analytics, acesse https://watson.analytics.ibmcloud.com e informe um login e senha. Use os dados a seguir, caso não queira criar um usuário na ferramenta.<br>Login: rvis.tool@gmail.com e senha: rvistool<br>Para usar a ferramenta VizAssist, acesse http://vizassist.fr/<br>Para usar a ferramenta Voyager, acesse https://vega.github.io/voyager/<br>Para instalar a ferramenta Tableau Public, acesse https://public.tableau.com/s/<br>Verifique se o Excel está instalado em sua máquina.<br><br>Conjunto de dados: Você criará visualizações do conjunto cidades_digitais.csv[32] (enviado por email). Quando for criar as visualizações no Excel, utilize o arquivo cidades_digitais.xls[33] (também enviado por email).<br>Esse conjunto de dados tem informações sobre o andamento do programa Cidades Digitais, como quantidade de pontos atendidos no município, população, status da implantação de cada |

[32] Arquivo de dados disponível em http://www.mc.gov.br/documentos/cidades-digitais-lista-de-cidades-atendidas.csv

[33] Este arquivo (.xlsx) foi gerado com base no arquivo original (.csv)

cidade contemplada, valor total previsto para a implantação e quanto foi investido até o momento.

**Crie 1 visualização para cada tarefa:**

Tarefa 1: Comparação

O que deve ser feito: comparar o valor total previsto para a implantação do programa Cidades Digitais com o valor investido até o momento em cada cidade.

Tarefa 2: Ranqueamento

O que deve ser feito: verificar o ranking dos estados (UF) por investimentos recebidos até o momento. Ou seja, descobrir qual estado recebeu mais investimentos até o momento.

OBS: Para realizar essa tarefa 2, no Excel, sugerimos que crie a visualização com os dados da aba "Dados_Agrupados_UF", pois os dados foram sumarizados por estado. A tarefa 1 deve ser realizada com os dados da aba "Dados"