ISSN: 1415-7314
ISSN online: 2317-6717

ENGEVISTA
Página da revista: http://www.uff.br/engevista/seer/

Análise de reclamações sobre produtos e serviços no programa de proteção e defesa do consumidor utilizando mineração de dados

Analysis of claims on products and services in the consumer protection and defense program using data mining

Bruno Rabbi
Diane Bosser Klug Rabbi
Virgínia Siqueira Gonçalves
Elias Rocha Gonçalves Júnior

Resumo: O programa de proteção e defesa do consumidor (PROCON) é um órgão público destinado à proteção e à defesa dos direitos e interesses dos consumidores, exercendo as funções de acompanhamento e fiscalização das relações de consumo. Este trabalho identifica informações úteis através da mineração de dados, possibilitando um melhor entendimento das reclamações dos consumidores sobre produtos e serviços, evidenciando principalmente, as reclamações sobre as empresas. Tem como objetivo estabelecer um padrão de dados viável de ser implantado que sirva como modelo a ser utilizado nas bases de dados do PROCON brasileiro. Para tal, foi utilizado o processo de KDD (*Knowledge Discovery in Database*) como metodologia para viabilizar a mineração de dados através do software WEKA (Waikato Environment For Knowledge Analysis). Com o mapeamento obtido através da aplicação de arvore de decisão, com o algoritmo j48, foram reveladas informações relevantes tal como o tempo de permanência de reclamações mais problemáticas. Os conceitos utilizados neste artigo poderão ser aplicados no estudo de outros órgãos regulamentadores, bem como ser útil no contexto de outros órgãos.
Palavras-chave: KDD; PROCON; Mineração de Dados; Produtos e serviços.

Abstract: The Consumer Protection and Defense Program (PROCON) is a public organ dedicated to the protection and defense of consumer rights and interests, exercising the functions of monitoring and supervising consumer relations. This work identifies useful information through the mining of data, enabling a better understanding of consumer complaints about products and services, highlighting mainly the complaints about companies. It aims to establish a viable data standard to be deployed that serves as a model to be used in Brazilian PROCON databases. For this purpose, the KDD (Knowledge Discovery in Database) process was used as a methodology to enable data mining through Weka (Waikato Environment For Knowledge Analysis) software. With the mapping obtained through the application of tree and j48, relevant information was revealed, such as the dwell time of the most problematic complaints. The concepts used in this article can be applied in the study of other regulatory bodies, as well as being useful in the context of other organs.
Keywords: KDD, PROCON, Data Mining, Products and services.

ISSN: 1415-7314
ISSN online: 2317-6717

1. Introdução

Gradativamente a sociedade apresenta-se mais consumista e necessitada de respostas rápidas estas por sua vez, são frutos de processos tecnológicos que são capazes de transformar dados em informação e simplificar aquilo que antes era burocrático. Logo, quando há aumento de consumo, possivelmente há também um aumento de insatisfações e uma diminuição do tempo útil do consumidor.

Segundo Albrecht e Zemke (2002), a percepção dos consumidores em relação à qualidade dos serviços, é o efeito da comparação entre as expectativas antes da prestação dos serviços, e o resultado concretizado dos mesmos.

Com base nas informações publicadas pela Prefeitura de Assis-SP, para facilitar a relação entre fornecedor e consumidor, o PROCON - Programa de Proteção e Defesa do Consumidor funciona como um órgão auxiliar do Poder Judiciário, tentando solucionar previamente os conflitos entre o consumidor e a empresa que vende um produto ou presta um serviço, e quando não há acordo, encaminha o caso para o Juizado Especial Cível com jurisdição sobre o local. O PROCON pode ser estadual ou municipal, e segundo o artigo 105 da Lei 8.078/90 (Código de Defesa do Consumidor), é parte integrante do Sistema Nacional de Defesa do Consumidor.

A Mineração de Dados aplicada a base de dados do PROCON tem como foco a descoberta de conhecimentos não induzidos sobre as reclamações não há direcionamento sobre os resultados obtidos mostrando tendências.

As ferramentas de análise de dados tradicionais, como BI (Business Intelligence) e OLAP (Online analytical processing), não são capazes de traçar perfis e relacionamentos não induzidos entre o tempo e o tipo de reclamação, havendo forte necessidade de implementar "inteligência computacional" às bases de dados registrada pelo PROCON.

A análise de dados tem sido realizada desde seus primórdios por meios estatísticos, nesse processo, os dados têm sido sistematicamente coletados e armazenados eletronicamente. Somente esse método matemático não proporciona a descoberta de informações desconhecidas em uma base de dados. Segundo Bothorel, Serrurier e Hurter (2011), a Mineração de Dados visa extrair o máximo de conhecimento a partir de bancos de dados com grande volume de informação. Neste contexto, nota-se a necessidade de explorar a base de dados relativa ao registro de reclamações, aplicando sobre esta o processo de DCBD (Descoberta de Conhecimento em Base de Dados) ou KDD (Knowledge Discovery in Databases).

Com grandes volumes de dados armazenados, de forma automática ou semi-automática podem-se utilizar vários modelos na mineração de dados, baseados em sistemas computacionais inteligentes para descoberta de conhecimentos. Tal processo foi definido por Fayyad, Piatetsky-Shapiro e Smyth (1996), como sendo: "... o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, embutidos nos dados". A Mineração de

ISSN: 1415-7314
ISSN online: 2317-6717

Dados apresenta-se como parte do processo de KDD.

Extrair conhecimentos que possam estar "ocultos", contextualizá-los e oferecer informações analíticas às empresas para minimizar as reclamações de serviços e produtos a fim de aprimorá-los de forma inteligente e mais objetiva, identificar possíveis padrões de reclamações e propor um modelo viável, aproveitando-se como base para aplicação das ferramentas de Mineração de Dados.

Baseado nas revisões bibliográficas e no estudo de caso, o propósito deste estudo é apresentar a técnica de Mineração de Dados para extração de conhecimento sobre a base de dados do PROCON.

O presente artigo está estruturado em quatro partes. A primeira contextualiza como funciona o PROCON, o processo de KDD e trata de uma breve revisão da literatura. A segunda apresenta a metodologia, descrevendo as etapas do KDD e os algoritmos utilizados, na terceira são demonstrados os resultados provenientes da Mineração de Dados aplicada, na quarta parte discorre sobre as conclusões apuradas através dos resultados.

2. Metodologia

A metodologia empregada no estudo considerou uma pesquisa bibliográfica a respeito do processo de descoberta do conhecimento em base de dados e as etapas que compõem tal processo, bem como, um estudo de caso aplicado na prática, com a finalidade de aprimorar o funcionamento da ferramenta de mineração de dados WEKA (*Waikato Environment For Knowledge Analysis*), com foco em seus algoritmos de busca, baseados em inteligência artificial, por meio de uma análise crítica dos dados registrados na base de dados do PROCON, seus atributos, domínios e abrangência, bem como a aplicação deste banco de dados na ferramenta após etapas de limpeza e transformação de dados.

2.1. Arquitetura proposta

Nesta etapa, será apresentada a estrutura conceitual, para que os dados provenientes da base de dados do PROCON possam ser utilizados no processo de KDD. A estrutura mostrada na Figura 1 foi concebida para dar suporte operacional ao modelo, a fim de que os dados provenientes da etapa de Mineração de Dados fossem passíveis de visualização.

ISSN: 1415-7314
ISSN online: 2317-6717

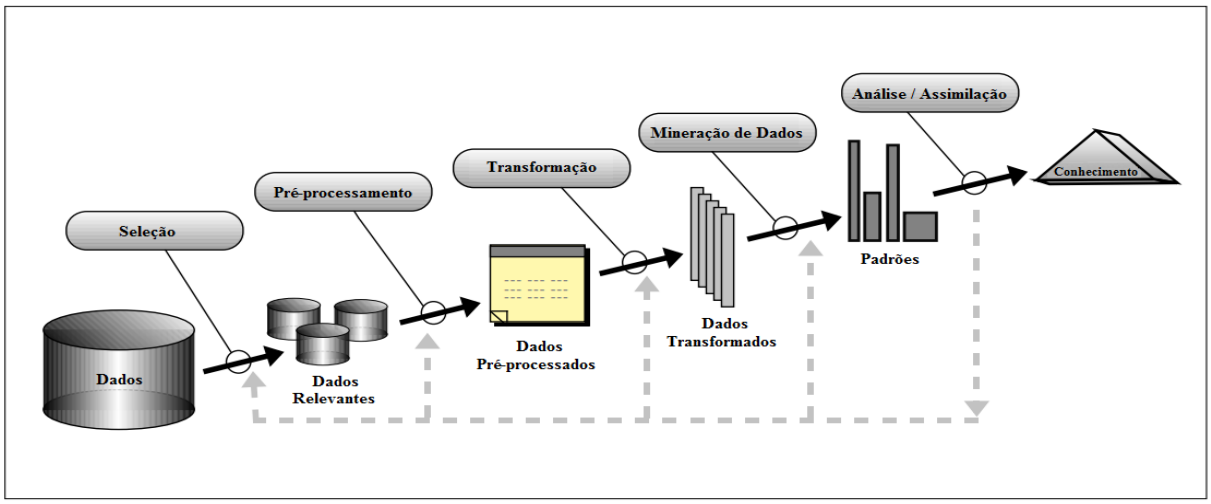

Figura 1: O ciclo do processo de KDD. Fonte: Adaptado de Fayyad et al. (1996).

O Algoritmo de apoio a decisão J48 se originou da necessidade de melhorar o algoritmo já existente C 4.5, foi criado inicialmente em linguagem de programação C, migrando depois para a linguagem Java (WITTEN et al., 2005). O algoritmo J48 criou regras e árvores de decisões que foram aplicadas neste estudo, cujo objetivo da árvore de decisão é agrupar as instâncias de forma recursiva, minimizando a entropia (ou seja, a variabilidade das classes) dentro de cada grupo. Para isso, o algoritmo utiliza os valores de atributos em cada nó não folha e deposita as instâncias nos nós folha. Dessa forma, cada nó interno corresponde a uma decisão de classificação, e a leitura e interpretação da árvore são naturais.

A versão da árvore de decisão original exige que todos os valores de atributos sejam de domínio finito e relativamente pequeno (para que o desempenho computacional seja aceitável). No entanto, o algoritmo usado pelo WEKA é o J48, uma versão do tradicional algoritmo de C4.5 que consegue, através da discretização de valores numéricos utilizando heurísticas, processar atributos de qualquer tipo na entrada. No entanto, a árvore não é capaz de realizar regressão e desempenha somente o papel de classificador, isto é, o atributo de classe deve ser discreto (RAMISCH; ENGEL, 2009). Diversos parâmetros permitem ajustar o comportamento da árvore, a maioria diz respeito às heurísticas usadas para podar uma árvore.

Um dos parâmetros é o número mínimo de nós em uma folha, que define um critério de parada para o algoritmo. Isso significa que, a partir do momento que uma folha atinge certo número de instâncias, mesmo que ainda exista ruído, ela não será subdividida. Valores altos para esse parâmetro, chamado M no WEKA, geram árvores mais genéricas enquanto valores baixos (próximos de 1) geram árvores com alta precisão.

Após construir a árvore de decisão, um passo de "poda da árvore" pode ser executado para reduzir o tamanho da árvore de decisão. Árvores de decisão que sejam grandes demais são susceptíveis a um fenômeno conhecido como *overfitting*. A poda ajuda a retirar as ramificações

ISSN: 1415-7314
ISSN online: 2317-6717

da árvore inicial de uma forma que melhore a capacidade de generalização da árvore de decisão (TAN et al., 2009).

Segundo Ingargiola (2011), após definição da árvore de decisão, é possível gerar as regras e estas são escritas considerando o caminho da base até os ramos da árvore. Os procedimentos, árvores de decisão e regras de associação, ocasionalmente são utilizados em conjunto.

Para Soares e Ochi (2004), a técnica de clusterização classifica um conjunto de elementos em subconjuntos de elementos ou classes, observando para isso características conforme critério apropriado. Esses subconjuntos são chamados de clusters, de forma que os objetos pertencentes a cluster possuam similares entre si e, ao mesmo tempo, os objetos pertencentes a clusters diferentes apresentem alta dissimilaridade.

2.2. Base de dados

A base utilizada neste estudo foi adquirida do site “dados.gov.br”, do governo federal, onde são encontradas diversas informações sobre os órgãos públicos, com base na Lei nº 12.527/2011, de acesso à informação criada em 2011.

O modelo de trabalho, exposto na Figura 1, representa a fonte de dados para o processo de Extração Transformação. O arquivo obtido do PROCON fornece os dados relativos ao registro dos clientes referente às empresas reclamadas, dando origem aos lançamentos das reclamações do PROCON.

Após a realização do download da base de dados, foi necessário a realização do tratamento do mesma, removendo erros de formatação e digitação existentes no arquivo original que encontrava-se em arquivo CSV e considerando-se apenas os seguintes atributos: Sexo, Faixa, Tempo, Fantasia, Assunto, Como, Avaliação, Nota.

Após a disponibilização inicial dos dados, foi iniciada a etapa de seleção, sendo possível então a identificação de possíveis problemas. Também foi possível formular hipóteses com base na compreensão dos dados.

2.3. Etapa de pré-processamento

Nesta etapa foram realizadas: a limpeza de dados, remoção de "ruídos", escolha de estratégias para manipular campos de dados ausentes e a formatação de dados, de maneira a adequá-los à ferramenta de mineração. Outras informações foram retiradas por não fazer parte do contexto abordado.

Alguns atributos precisaram ser manipulados para suprir a ausência de conteúdo. Os atributos “FAIXA” e ”NOTA” foram alterados em seu tipo de dados, passando de numérico

ISSN: 1415-7314
ISSN online: 2317-6717

para alfanumérico, dada a restrição de campos numéricos para processamento do algoritmo APRIORI.

Outras conversões foram realizadas dinamicamente através de funções utilizadas nos comandos do Excel, como a função replace. Para um melhor entendimento, apresenta-se no Quadro 1, o conjunto de atributos utilizados e a descrição de seu conteúdo.

Quadro 1: Principais atributos constantes dos conjuntos de dados minerados

Atributo	Descrição
Sexo	Identificação do sexo do reclamante
Faixa	Faixa etária de idade do reclamante
Tempo	Tempo que durou a ligação da reclamação
Fantasia	Nome da empresa fornecedora de serviço ou produto
Assunto	Descrição da reclamação feita
Como	Local de aquisição do bem ou serviço
Avaliação	Se a reclamação foi resolvida ou não resolvida pela empresa
Nota	Avaliação de 1 a 5 – qualidade do atendimento

2.4. Etapa de pré-processamento

A etapa de transformação é a etapa que antecede a Mineração de Dados. Foram aplicadas operações que melhoraram a visualização dos dados. Também foram transformados valores discretos em contínuos, a fim de facilitar a sumarização pelos algoritmos utilizados. Alguns atributos necessitaram ter o tipo de dados alterado para facilitar o uso de algoritmos específicos como o Apriori.

Atributos e complementos que estavam ausentes também foram adicionados a partir de valores de outros atributos correlacionados, sendo estabelecido conteúdo apropriado. Logo na sequência foram apresentadas algumas situações na qual necessitou-se realizar transformações nos dados antes da etapa de mineração.

O atributo “Faixa” descreve a faixa etária de idade do reclamante. Este campo é do tipo alfanumérico, baseado no conteúdo do atributo “Faixa”, de tipo numérico, tendo os valores adaptados conforme Quadro 2.

Quadro 2: Conteúdo transformado do atributo “faixa”

ISSN: 1415-7314
ISSN online: 2317-6717

FAIXA - numérica:	FAIXA - alfa numérica:
<= 20	até 20 anos
21 a 30	entre 21 e 30 anos
31 a 40	entre 31 a 40 anos
41 a 50	entre 41 a 50 anos
51 a 60	entre 51 a 60 anos
61 a 70	entre 61 a 70 anos
>=71	+ de 71 anos

O atributo "tempo" descreve o tempo de atendimento. Este campo é do tipo alfanumérico, foi formulado de acordo com o conteúdo do atributo "TEMPO", de tipo numérico, tendo os valores adaptados conforme Quadro 3.

Quadro 3: Conteúdo transformado do atributo "tempo"

TEMPO - numérica:	TEMPO - alfa numérica:
1 a 2 minutos	>=1 ; <=2
3 a 4 minutos	>=3 ; <=4
5 a 6 minutos	>=5 ; <=6
7 a 8 minutos	>=7 ; <=8
9 a 10 minutos	>=9 ; <=10

O atributo "ASSUNTO" descreve o assunto do atendimento esse campo foi codificado devido ao texto extenso de cada assunto para ajudar no momento de exposição dos resultados conforme Quadro 4.

Quadro 4: Conteúdo transformado do atributo "assunto"

Código alfa numérico	ASSUNTO
a	Serviço de pagamento online/ via celular
b	Acessórios periféricos
c	Aéreo
d	Aparelho celular
e	Aparelho de som, vídeo e imagem (Câmera, filmadora, dvd, home theater, etc)
f	Aparelho de telefone fixo / interfone
g	Ar condicionado e aquecedor
h	Atendimento Bancário
i	Banco de Dados e Cadastros de Consumidores (SPC, Serasa, SCPC etc)
j	Cartão de Crédito / Cartão de Débito / Cartão de Loja
k	Móveis e Colchões
i	Conta corrente / Salário / Poupança /Conta Aposentadoria
m	Crédito Consignado (Empréstimo descontado em folha de pagamento)
n	Demais Empréstimos e Financiamentos (exceto imóveis e veículos)
o	Demais Seguros (exceto habitacional)
p	Eletroportáteis (batedeira, liquidificador, umidificador, secador, etc)

ISSN: 1415-7314
ISSN online: 2317-6717

q	Esporte/Lazer
r	Financiamento de Imóveis
s	Financiamentos de Veículos / Leasing
t	Fogão, microondas, forno elétrico, depurador de ar e coifa
u	Internet Fixa
v	Internet Móvel
w	Pacote de Serviços (Combo)
x	Lavadora de roupa, louça e secadora
y	Microcomputador e laptops
z	Livros e papelaria (material didático, jornais, revistas, artigos de escritório, etc.)
aa	Telefonia Móvel Pré-paga
ab	Vestuário e Artigos de Uso Pessoal (roupa, calçados, joias, bijuterias, malas, bolsas, etc)
ac	Utilidades
ad	TV por Assinatura
ae	Televisão
af	Telefonia Móvel Pós-paga
ag	Tablet
ah	Refrigerador / freezer
ai	Programas de Fidelidade / Benefícios (pontos, milhagem etc)
aj	Produtos para crianças (carrinho, brinquedos, cadeira, vestuário, mamadeira, etc)
al	Plano de Saúde (convênio, autogestão, seguro saúde)

3. Resultados e Discussão

3.1. Etapa de clusterização

Os dados de entrada da tarefa de classificação são um conjunto de registros. Os registros também são conhecidos como instâncias e são caracterizados por uma dupla (x,y), onde x é o conjunto de atributos e y o atributo especial, designado como rótulo da classe. Também conhecido como atributo alvo ou de categorização, este atributo deve ser um atributo discreto (TAN et al., 2009).

Para Salvador et al. (2009), a clusterização, uma das técnicas da mineração de dados, busca agrupar os dados de tal maneira que seja capaz de potencializar a similaridade dos objetos de um mesmo grupo e/ou diferença entre grupos distintos.

O arquivo de dados resultante inclui 3627 casos. Os testes realizados neste trabalho foram baseados na aplicação do algoritmo de agrupamento implementado pelo WEKA e baseado no *SimpleK-Means* em um resultado de experimento de reclamações do PROCON para encontrar o cluster que apresenta a maiores reclamações por empresa, assim não foram exploradas todas as informações apresentadas no relatório do algoritmo, mas sim encontrar um cluster que se apresenta maior concentração em relação aos outros.

Este exemplo ilustra o uso de *SimpleK-means* clusters com WEKA. O conjunto de

ISSN: 1415-7314
ISSN online: 2317-6717

dados da amostra utilizada para este exemplo é baseado nos "dados do PROCON" disponíveis em formato separado por vírgulas. Este documento assume que o pré-processamento de dados apropriados foi realizado. Neste caso, uma versão do conjunto inicial de dados foi criada, em que o campo de ID foi removido. A clusterização está representada na Figura 2.

Cluster#				
0	1	2	3	4
(1599.0)	(758.0)	(1894.0)	(574.0)	(322.0)
M	M	M	F	M
entre 21 a 30 anos	entre 31 a 40 anos	entre 21 a 30 anos	entre 31 a 40 anos	entre 21 a 30 anos
>=9 <=10	>=5 <=6	>=9 <=10	>=9 <=10	>=7 <=8
Tim	Vivo - Telefônica	Vivo - Telefônica	Oi Fixo	SKY
ad	af	d	ag	ad
Internet	Loja física	Internet	Telefone	Telefone
Res	nr	Res	nr	nr
OTIMO	PESSIMO	OTIMO	PESSIMO	PESSIMO
SE	SE	SE	SE	NE

Figura 2: Cluster de evidência de tempo, faixa etária e empresa. Fonte: Elaborado pelos autores.

3.2. Etapa das árvores de decisão

Segundo Fayyad et al. (1996), árvore de decisão pode ser definida modelo preditivo que pode ser visualizado na forma de uma árvore, daí seu nome. Cada ramo da árvore é uma questão de classificação e cada folha é uma partição do conjunto de dados com sua classificação.

Com o algoritmo J48, foram criadas árvores de decisão relativas ao tempo de atendimento. A seguir são mostrados os resultados provenientes desta descoberta de conhecimento. Na Tabela 1, é apresentado a taxa de confiança, extraídos do conjunto “Procon”, juntamente com as Árvores geradas mostrado na Figura 3, a partir de um conjunto de 3780 (três mil setecentos e oitenta) registros processados, relativo ao ano de 2014.

Tabela 1: Classificação das instâncias relacionadas às árvores de decisão da Figura 3

Instâncias classificadas Corretamente	Instâncias classificadas Incorretamente	Atributo Classe
68,14%	31,86%	Tempo

ISSN: 1415-7314
ISSN online: 2317-6717

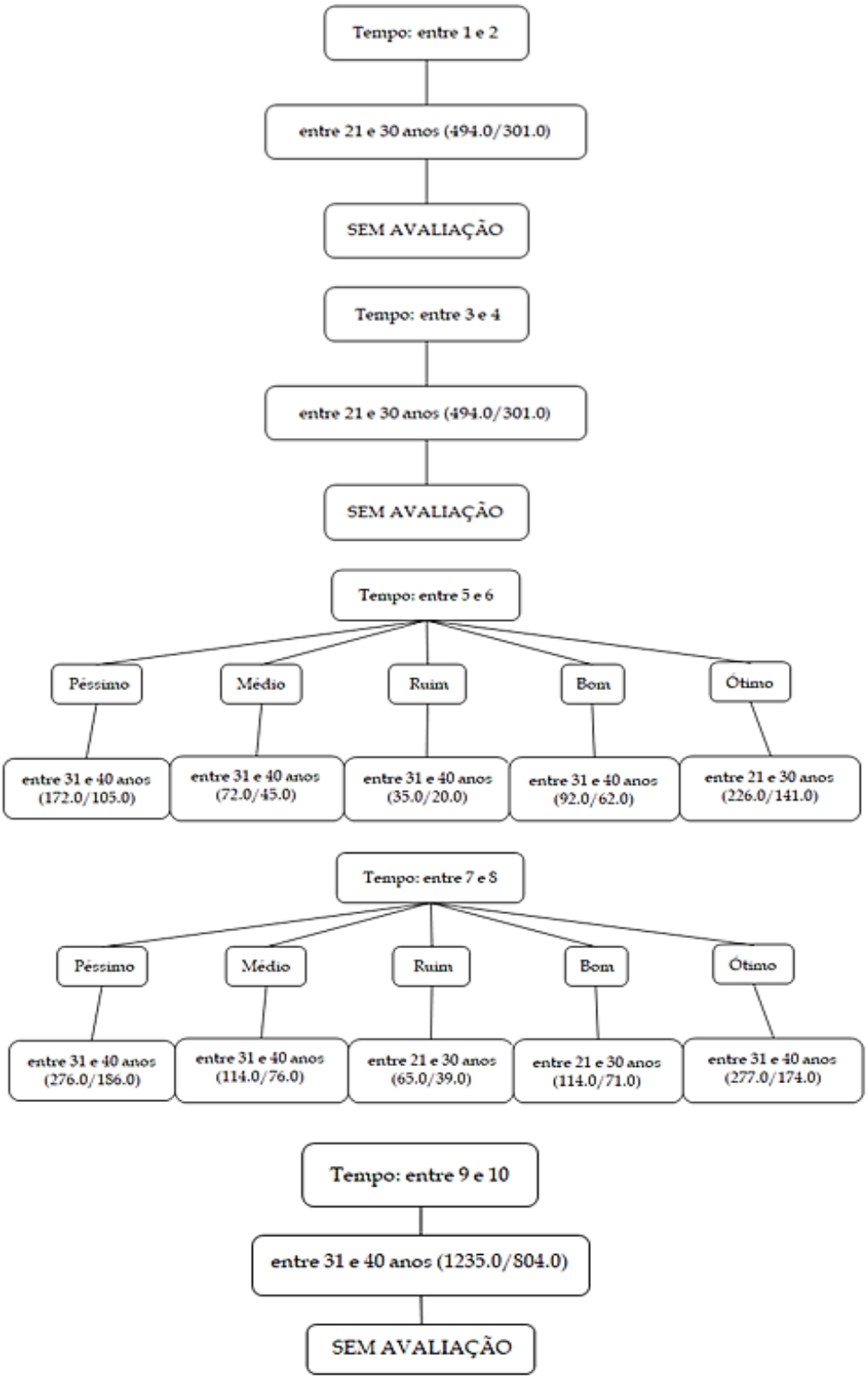

Figura 3: Árvores de decisão geradas a partir do algoritmo J48. Fonte: Elaborado pelos autores.

4. Considerações finais

A Mineração de Dados é uma das principais ferramentas a ser utilizada para exploração das grandes bases de dados existentes nas empresas privadas também se aplica às corporações públicas. As informações obtidas através do algoritmo J48 possibilitaram a identificação de

ISSN: 1415-7314
ISSN online: 2317-6717

padrões não encontrados com as ferramentas tradicionais de análise de dados, como: relatórios gerenciais e consultas por ferramentas de BI (*Business Intelligence*).

Não foram estendidos os testes para uma maior quantidade de Clusters, pois, à medida que se aumenta a quantidade de Clusters, diminui a quantidade de erros. Assim deve-se observar que chegaria um ponto onde a quantidade de clusters seria a mesma da quantidade de registros na base de dados, assim o algoritmo colocaria cada registro em um único cluster, desta forma chega-se a erro zero, torna-se assim irrelevante as informações obtidas.

Os testes foram encerrados com 05 (cinco) clusters, pois com essa quantidade foram encontrados clusters com representações bem superiores aos outros, pode-se assim enquadrar esses Clusters como os de maiores representatividades. Sugere-se para estudos posteriores a avaliação de outros algoritmos (como exemplo, o EM – *Expectation Maximization*).

Os resultados encontrados e apresentados trazem um conhecimento inovador não induzido e útil, possibilitando a tomada de decisões com alto índice de acerto. Assim este estudo, pretende induzir o uso das técnicas de Mineração de Dados no PROCON e contribuir para uma fase de maior eficácia na análise de informações.

5. Referências

ALBRECHT, K.; ZEMKE, R. 2002. *Serviço ao Cliente*. Rio de Janeiro: campus.

BOTHOREL, G.; SERRURIER, M.; HURTER, C. 2011. *Utilisation d'outils de Visual Data Mining pour l'exploration d'un ensemble de règles d'association*. Sophia Antipolis, France: Ihm.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. 1996. From Data Mining to Knowledge Discovery in Databases. *AI magazine*. p 37-54.

INGARGIOLA, G. 2016. *Building Classification Models: ID3 and C4.5*. Disponível em: http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html. Acesso em: 13 de julho de

Informações sobre Procon publicadas pela prefeitura de Assis-SP - O que é o Procon? Disponível em: <http:// http://www.assis.sp.gov.br/CIDADAO?id=2> Acesso em: 07 de Agosto de 2016.

RAMISCH, C; ENGEL, P. 2009. *Algoritmos de aprendizado para avaliação de carros*. Porto Alegre: UFRGS.

SALVADOR, H. G.; CUNHA, A. M.; CORRÊA, C. S. 2009. Vedalogic: um método de Verificação de Dados Climatológicos Apoiado em Modelos Minerados. *Revista brasileira de Meteorologia*, São Paulo, v. 24, n. 4.

SOARES, S. R. F.; OCHI, L.S. 2004. *Um algoritmo evolutivo com reconexão de caminhos para o problema de clusterização automática*. Disponível em: < http://www.ic.uff.br/satoru/conteudo/artigos/claio2004-stenio.pdf>. Acesso em: 01 de Agosto de 2016.

TAN, P.; STEINBACH, M.; KUMAR, V. 2009. *Introdução ao Data Mining: Mineração de Dados*. Rio de Janeiro, Editora Ciência Moderna.

ISSN: 1415-7314
ISSN online: 2317-6717

Agradecimentos

Os autores gostariam de agradecer à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo fomento a esta pesquisa.