UNIVERSIDADE FEDERAL DO PARANÁ

TALITA DE SOUZA RAMPÃO

MINERAÇÃO DE DADOS EM BASES JURÍDICAS: UM ESTUDO DE CASO

CURITIBA

2016

TALITA DE SOUZA RAMPÃO

MINERAÇÃO DE DADOS EM BASES JURÍDICAS: UM ESTUDO DE CASO

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção de grau de Bacharel no Curso de Gestão da Informação, Departamento de Ciência e Gestão da Informação do Setor de Ciências Sociais Aplicadas, da Universidade Federal do Paraná.

Orientadora: Prof.ª Dr.ª Denise Fukumi Tsunoda

CURITIBA

2016

TERMO DE APROVAÇÃO

TALITA DE SOUZA RAMPÃO

MINERAÇÃO DE DADOS EM BASES JURÍDICAS: UM ESTUDO DE CASO

Trabalho apresentado como requisito parcial à obtenção do grau de bacharel em Gestão da Informação no curso de graduação em Gestão da informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, pela seguinte banca examinadora:

_____

Prof.ª Dr. ª Denise FukumiTsunoda

Orientadora - Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

_____

Prof. Dr. José Simão de Paula Pinto

Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

_____

Prof. Dr. Cícero Aparecido Bezerra

Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

_____

Aurélio Câncio Peluso

Advogado – OAB 32521PR

Curitiba, 05 de dezembro de 2016

AGRADECIMENTOS

Primeiramente à Deus que iluminou o meu caminho durante esta caminhada.

Aos meus pais, Juarez e Celi, que com muito carinho e apoio não mediram esforços para que eu chegasse até esta etapa de minha vida.

A minha irmã, Francieli, pelo amor, incentivo e apoio incondicional.

Ao meu namorado, Marcelo, que de forma especial e carinhosa me deu força e coragem, me apoiando e auxiliando para que esse dia chegasse.

A minha orientadora, profª. Drª. Denise Tsunoda, pela paciência e incentivo que tornaram possível a conclusão desse trabalho.

Ao escritório de advocacia, pela confiança e apoio constante, pois além de permitirem a utilização das bases de dados, foram atenciosos e auxiliaram com o possível.

Aos colegas e amigos que conheci ao longo dessa jornada, pelo companheirismo e compreensão ao compartilharmos madrugadas, finais de semana e feriados. Cada esforço valeu a pena e foi essencial para que chegássemos até aqui.

“A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo”.

Albert Einstein

RESUMO

Estudo de caso sobre mineração de dados aplicada a uma base de dados jurídica contendo processos cíveis de direito do consumidor com enfoque em: tarifa, tarifa e dano moral, revisional, indenizatória e outras. Objetiva a aplicação de técnicas de mineração de dados na área jurídica para verificar a existência de padrões de decisões judiciais de acordo com o Estado em que tramita o processo. Constitui-se de um estudo de caso com pesquisa descritiva, finalidade aplicada e abordagem quantitativa. Realiza a aplicação das tarefas de classificação e associação por meio dos métodos Apriori, PART, Decision Table, J48 (C4.5) e REPTree. Demonstra que é possível prever padrões de decisões judiciais de acordo com o órgão julgador, tipo de ação e região que tramita o processo. Propõem a análise e continuidade do estudo para verificar a aplicação de técnicas de mineração em outras bases de dados jurídicas, a fim de validar a proposta e comparar as variações nos resultados obtidos.

Palavras-chave: Direito. Gestão da Informação. Descoberta de Conhecimento em Bases de Dados. Mineração de Dados. Tomada de Decisão.

ABSTRACT

Case study on data mining applied to a legal database containing civil cases of consumer law focusing on: tariff, tariff and moral damages, revisional, indemnification and others. It aims to apply data mining techniques in the legal area to verify the existence of patterns or tendencies of judicial decisions according to the State in which the process is being processed. It is a case study with descriptive research, applied purpose and quantitative approach. It performs the application of classification and association tasks through the Apriori, PART, Decision Table, J48 (C4.5) and REPTree methods. It shows that it is possible to predict trends in judicial decisions according to the adjudicating body, type of action and region that processes the process. They propose the analysis and continuity of the study to verify the application of mining techniques in other legal databases, in order to validate the proposal and compare the variations in the results obtained.

Keywords*:* Law. Information management. Knowledge Discovery in Database. Data Mining. Decision-Making

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A era da informação trouxe mudanças no paradigma da sociedade, facilitando o acesso, uso e compartilhamento instantâneo das informações com o auxílio das Tecnologias da Informação e Comunicação (TIC). Contudo, trouxe consigo o excesso informacional, no qual Braga (s/d) afirma que há informação demais e tempo de menos, tornando cada vez mais complexo o processo de tomada de decisão. Um levantamento quantitativo realizado pelo autor demonstra esse crescimento do volume de informações: atualmente existem mais de três bilhões de páginas disponíveis na *internet*; estão em circulação mais de 100 mil revistas científicas no planeta; mais de 1.000 novos títulos de livros são editados por dia em todo o mundo.

De acordo com Sidney (2010), essa grande quantidade de dados torna a análise humana onerosa e métodos tradicionais de recuperação de dados, mesmo que sejam sofisticados, não são eficazes para descoberta de conhecimentos “ocultos” em massas de dados como *big data*, por exemplo. Nesse contexto, a descoberta de conhecimento em bases de dados ou *Knowledge Discovery in Databases* (KDD) surge como alternativa para auxiliar a descoberta automática de conhecimento por meio do processo completo de conversão de dados brutos em informações úteis. (TAN; STEINBACH; KUMAR, 2009, p. 4).

O KDD possui como propósito realizar a descoberta de informações relevantes a partir de análise de padrões de grandes conjuntos de dados, de modo a apoiar decisões estratégias. Para isso, conta com as fases de seleção, pré-processamento, transformação, mineração dos dados e interpretação de resultados. Todas as fases são importantes, no entanto, a etapa de mineração de dados recebe maior destaque na literatura, considerando que passou a ser vista como um diferencial competitivo, auxiliando os tomadores de decisão a realizarem escolhas estratégicas.

Castro e Ferrari (2016, p. 17) destacam que são típicas aplicações da mineração de dados para análise e predição de crédito, detecção de fraudes, predição do mercado financeiro, relacionamento com clientes, predição de falência corporativa e muitas outras. De acordo com os autores, exemplos de segmentos de aplicação incluem setor financeiro; planejamento estratégico empresarial; planejamento do setor portuário; setores de energia; educação; logística; planejamento das cadeias de produção, distribuição e suprimentos; meio ambiente; e *internet.* Os autores afirmam

que aplicações típicas incluem a identificação ou segmentação de clientes, parceiros, colaboradores; detecção de fraudes e anomalias em sistemas e processos; ações estratégicas de *marketing*, CRM (*Customer Relationship Management*) e recursos humanos; jogos e atividades educacionais; gestão do conhecimento; análise de padrões de consumo; compreensão de bases de dados industriais, biológicas, empresariais e acadêmicas; predição de retorno sobre investimento, despesas, receitas, investimentos, etc.; e mineração de dados da *web*.

Tendo em vista a variedade de aplicações da mineração de dados, a presente pesquisa visa a sua aplicação na análise de dados referentes aos processos jurídicos que discutem tarifas bancárias, revisão de cláusulas contratuais, indenização por danos morais e outras, buscando identificar padrões de decisões jurídicas por Estado. Estrada (2015) destaca que a utilização de algoritmos para prever resultados já é utilizada em várias áreas de impacto social, sendo que as previsões orientadas por dados podem fornecer informações adicionais para apoiar a análise dos advogados.

1.1 PROBLEMATIZAÇÃO

Os avanços da tecnologia, tanto de hardware quanto de comunicação – têm produzido um problema de “superabundância” de dados, pois a capacidade de coletar e armazenar dados tem superado a habilidade de analisar e extrair conhecimentos deles. Nesse contexto, é necessária a aplicação de técnicas e ferramentas que transformem, de maneira inteligente e automática, os dados disponíveis em informações úteis, que representem conhecimento para a tomada de decisão estratégica nos negócios e até no dia a dia. (CASTRO; FERRARI, 2016, p. 3)

O aumento dos fluxos informacionais tem gerado para as organizações uma mudança de paradigma referente ao gerenciamento das informações. Trouxe consigo o “excesso informacional”, no qual encontrar a informação certa no momento oportuno tem se tornado fonte de vantagem competitiva. Castro e Ferrari (2006, p. 4) afirmam que é nesse contexto de superabundância de dados que surgiu a mineração de dados, como um processo sistemático, interativo e iterativo, de preparação e extração de conhecimentos a partir de grandes bases de dados.

Na área jurídica, o avanço tecnológico proporcionou a tramitação dos processos em meio eletrônico, otimizando as atividades dos profissionais da área. No entanto, devido ao grande volume de processos, torna-se complexo extrair padrões

de decisões, considerando que as mesmas variam de acordo com o Estado e a Comarca em que estão tramitando. De acordo com Coelho (2006, apud Delgado, s/d) “à maioria das pessoas será, hoje, familiar a notícia de dois processos idênticos decididos de modo opostos”. Ainda nesse aspecto, Silva (2011) complementa que:

> Existe um número muito grande de processos "repetidos", isto é, processos em que uma das partes é a mesma e que versam sobre uma mesma questão jurídica. Esses processos se arrastam durante anos pelo Judiciário até obter uma decisão final, que, em tese, deveria ser a mesma para todos aqueles que estão em uma mesma situação. Afinal de contas, o direito deve ser idêntico para as pessoas que estão na mesma situação de fato e de direito, caso contrário, o direito seria uma loteria. Não é preciso meditar muito para se concluir que casos tais devem ser objeto de um único processo de conhecimento. Não é razoável que existem milhares (ou milhões) de processos de conhecimento para se decidir uma mesma questão jurídica. É preciso que questões "repetidas" (na realidade, a questão é uma só) sejam objeto de um único processo de conhecimento, que deve produzir efeitos para todas as pessoas. (SILVA, 2011, s/n)

Como consequência da falta de uniformidade das decisões judiciais, muitos escritórios de advocacia enfrentam dificuldades em identificar os critérios adotados em cada Comarca. Além disso, com a advocacia em massa, torna-se um fator de vantagem competitiva tomar decisões com base em informações fundamentadas.

Esta pesquisa realiza a aplicação de técnicas de mineração de dados sobre uma base jurídica cedida por uma organização atuante no ramo, de modo a identificar se existem padrões, conforme o Estado em que tramita o processo. A base de dados analisada é constituída por processos cíveis no direito do consumidor em contratos de financiamento, contendo diferentes tipos de processos: tarifa, tarifa e dano moral, revisional, indenizatória e outras. Para facilitar a compreensão de alguns termos jurídicos utilizados, foi elaborado um glossário de termos (Apêndice A).

A pesquisa tem por foco a seguinte pergunta: é possível prever padrões de decisões judiciais a partir da aplicação de técnicas de mineração de dados?

1.2 OBJETIVOS

Para responder à questão levantada na pesquisa foram definidos os objetivos a serem alcançados, sendo estes desmembrados em objetivo geral e específicos.

1.2.1 Objetivo Geral

O objetivo geral consiste na aplicação de técnicas de mineração de dados em uma base de dados jurídica, de modo a identificar padrões de decisões judiciais.

1.2.2 Objetivos Específicos

São definidos os objetivos específicos para que se possa alcançar o objetivo geral previamente estabelecido, sendo eles:

- pesquisar e definir o(s) método(s) de mineração de dados que será(ão) utilizado(s) na base de dados jurídica;
- classificar os Estados com base nos padrões de decisões identificadas;
- identificar o formato da base de dados que viabilize a aplicação de métodos de mineração de dados.

1.3 JUSTIFICATIVA

Com o grande volume de processos tramitando nos tribunais brasileiros, torna-se complexo extrair padrões entre as decisões proferidas devido à falta da uniformização processual. Torna-se comum encontrar processos com pedidos e alegações semelhantes, mas com julgamentos divergentes, de acordo com o entendimento do juízo em que está tramitando a ação.

Foi realizado um levantamento em 10 de março de 2016 na base principal da *Web of Science*, de modo a verificar as pesquisas existentes na área considerando o acervo bibliométrico. Para a pesquisa foi utilizado o parâmetro "*Data mining"*, pesquisando pelo título e considerando todos os anos e índices. Como resultado foram obtidos 9.684 registros. Combinando "*Data mining*" com "*Law*" foram obtidos apenas 11 registros. O resultado indica que existe pouca pesquisa desenvolvida na área, destacando a relevância da contribuição do desenvolvimento do estudo.

A Figura 1 demonstra as dez principais áreas de desenvolvimento de pesquisas em mineração de dados.

Figura 1 - Áreas de pesquisa *Data mining*

Campo: Áreas de pesquisa	Contagem do registro	% de 9684	Gráfico de barras
COMPUTER SCIENCE	6087	62.856 %	
ENGINEERING	2777	28.676 %	
OPERATIONS RESEARCH MANAGEMENT SCIENCE	628	6.485 %	
BUSINESS ECONOMICS	476	4.915 %	
AUTOMATION CONTROL SYSTEMS	451	4.657 %	
TELECOMMUNICATIONS	441	4.554 %	
MATHEMATICS	425	4.389 %	
BIOCHEMISTRY MOLECULAR BIOLOGY	264	2.726 %	
MEDICAL INFORMATICS	240	2.478 %	
PHARMACOLOGY PHARMACY	232	2.396 %	

FONTE: *Web of Science* (2016)

A Figura 2 demonstra as áreas de pesquisa correspondentes aos 11 resultados, sendo que em ambas se destaca a Ciência da Computação (*Computer Science*).

Figura 2 - Áreas de pesquisa - *Data mining and Law*

Campo: Áreas de pesquisa	Contagem do registro	% de 11	Gráfico de barras
COMPUTER SCIENCE	7	63.636 %	
GOVERNMENT LAW	3	27.273 %	
MATHEMATICAL COMPUTATIONAL BIOLOGY	3	27.273 %	

FONTE: *Web of Science* (2016)

Outro levantamento foi realizado na base de dados *SciELO (Scientific Electronic Library Online)*, pesquisando por publicações realizadas no Brasil. Para a pesquisa foi utilizado como parâmetro "Mineração de Dados*",* considerando todos os anos e índices. Como resultados foram obtidos 98 registros. Combinando "Mineração de Dados" com "Direito" não foi obtido nenhum resultado.

Além da falta de estudos desenvolvidos na área, outra motivação para o desenvolvimento da pesquisa foi em decorrência do período de estágio supervisionado desenvolvido em uma organização atuante no setor. Ao vivenciar o dilema do profissional com a falta de uniformização dos processos surgiu o interesse em pesquisar métodos ou ferramentas que pudessem auxiliar na identificação de padrões para otimizar a tomada de decisão.

Com a identificação de padrões referentes aos julgamentos os advogados podem passar estimativas mais fiéis aos seus clientes quanto aos resultados dos processos. Surden (2014) destaca que os advogados usam uma mistura de formação jurídica, resolução de problemas, análise, experiência, raciocínio analógico, senso comum, intuição e outras “habilidades cognitivas de ordem superior” para fazer avaliações sofisticadas e assim conseguirem possíveis resultados. Estrada (2015) complementa que a utilização de algoritmos permite fornecer informações adicionais para apoiar a análise do advogado e destaca que o uso de algoritmos para prever resultados já é muito usado em várias áreas de impacto social, como a Economia, por exemplo.

Estrada (2015) destaca que já é possível com a utilização de algoritmos inteligentes a criação de leis, encontrar violações de contratos comerciais e trabalhistas, dentre outros, assim como fraudes eleitorais, ou seja, também poderia ser muito útil para o legislador, criando leis mais eficientes conforme as necessidades da sociedade tendo por base os dados que filtrar. No Brasil, alguns softwares da área jurídica já estão sendo desenvolvidos com o intuito de auxiliar os profissionais da área na tomada de decisão. Um exemplo é o Jurimetria®, um sistema que consiste no levantamento de documentos jurídicos para aferição de tendência jurisprudencial.

Em um levantamento realizado em 2015, Rover (2015) aponta os cinco principais softwares utilizados por empresas e escritórios de advocacia: CPJ-Preâmbulo®, Espaider®, RR Jurídico®, e-Xyon® e CP-PRO®, destacando as suas funcionalidades. É possível identificar que os sistemas não apresentam funcionalidades de análise de dados, sendo uma ferramenta que apresenta grande potencial para a área jurídica. Assim, com a falta de desenvolvimento de ferramentas na área, esse estudo aborda a utilização da base de dados do cliente, a qual já possui as informações referentes aos andamentos processuais, para realizar uma análise crítica com base nos padrões identificados com o processo de mineração.

A seguir é apresentada a delimitação da pesquisa, de modo a identificar o escopo do estudo.

1.4 DELIMITAÇÃO DA PESQUISA

A pesquisa abrange uma base de dados cedida por uma organização atuante no setor jurídico. Nela são encontradas apenas ações cíveis, cujo processo originário

verse sobre direito do consumidor com enfoque em: tarifa, tarifa e dano moral, revisional, indenizatória e outras.

As reclamações de tarifas discutem a legitimidade da cobrança das tarifas administrativas para concessão e cobrança dos créditos objetos de contratos bancários, comumente identificadas como Taxa de Abertura de Crédito (TAC) e Taxa de Emissão de Carnê ou boleto (TEC), assim como outras, correlatas, bem como a possibilidade do pagamento parcelado do Imposto sobre Operações Financeiras (IOF). Os processos de tarifa e dano moral incluem o mesmo tipo de reclamação, porém cumulada com indenização por danos morais. Nesses casos a parte autora requer a devolução dos valores indevidamente cobrados e indenização por danos morais. De acordo com a ministra Maria Isabel Gallotti (2012), a "discriminação dos encargos contratuais em nada onera o consumidor, ao contrário". No entendimento do Superior Tribunal de Justiça (STJ), a fixação de tarifas administrativas em contrato de financiamento é prática legal, desde que elas sejam pactuadas em contrato e em consonância com a regulamentação do Banco Central.

As ações revisionais de acordo com Raddatz (2014) objetivam revisar contratos de financiamento ou empréstimos de instituições financeiras autorizadas, tanto para uso pessoal quanto para compra de móveis, veículos, equipamentos (industriais, agrícolas), com ou sem alienação fiduciária. Incluem entre as principais reclamações a cobrança indevida de capitalização de juros, juros abusivos, comissão de permanência, entre outros encargos.

A ação indenizatória, também chamada de ressarcitória ou reparatória, de acordo com o CC, art. 159 “visa a restabelecer uma situação existente antes do ato ilícito ocorrer, seja ele por negligência ou imprudência de outrem, para ressarcimento do dano causado”. Inclui entre as principais reclamações a inexistência de débito/contrato (fraude) e inscrição indevida nos serviços de proteção financeira; obrigação de fazer para baixa do gravame; vício no veículo, entre outros.

As ações classificadas como outras correspondem aos demais tipos de reclamações associadas a financiamento/*leasing* bancário que não se adequam as categorias anteriormente apresentadas, possuindo menor recorrência. São exemplos: obrigação de fazer, consignatória, reintegração de posse, entre outras.

1.5 ESTRUTURA DO DOCUMENTO

O documento está dividido em cinco seções principais. A primeira corresponde à base introdutória sobre o estudo realizado, identificando a problemática que o originou, bem como a justificativa para a sua realização, delimitando para isso o escopo da pesquisa.

A segunda seção corresponde à revisão da literatura pertinente relacionada ao tema abordado, apresentando uma contextualização sobre a gestão da informação e o papel fundamental da recuperação da informação para a tomada de decisão, destacando o processo de descoberta de conhecimento em bases de dados (KDD) para a sua otimização. Nesse contexto é abordado o passo da mineração de dados, apresentando brevemente os principais grupos, bem como seus algoritmos.

A terceira seção aborda a metodologia utilizada para o desenvolvimento do estudo. Identifica o ambiente da pesquisa, a sua caracterização, bem como os processos de análise documental, mineração de dados e validação da proposta realizados para atingir os objetivos propostos na primeira seção.

A quarta seção apresenta os resultados obtidos, realizando uma análise e descrição estatística da base de dados e apresentando os resultados da mineração de dados.

A quinta seção apresenta as considerações finais, destacando as dificuldades e contribuições do estudo, os resultados alcançados e a possibilidade da sua continuidade.

2 REVISÃO DE LITERATURA

A seguir é apresentada a fundamentação teórica proposta para nortear a pesquisa de acordo com a problemática investigada e os objetivos traçados. Para isso, foram abordados como temas: gestão da informação, recuperação da informação, tomada de decisão, KDD (*Knowledge Discovery in Databases*), pré-processamento, mineração de dados (*data mining)* e pós-processamento.

2.1 GESTÃO DA INFORMAÇÃO

Para Davenport (1998, 173), a Gestão da Informação (GI) é vista como um conjunto estruturado de atividades que incluem o modo como as empresas obtêm, distribuem e usam a informação e o conhecimento. Tarapanoff (2006, p. 22) complementa que o principal objetivo da gestão da informação é identificar e potencializar recursos informacionais de uma organização ou empresa e sua capacidade de informação, ensinando-a a aprender e adaptar-se a mudanças ambientais.

Em seu livro “Ecologia da informação: por que só a tecnologia não basta para o sucesso na era da informação” Davenport (1998, p. 18) realiza uma abordagem ecológica para o gerenciamento da informação. Para isso, o autor realiza a distinção entre dado, informação e conhecimento.

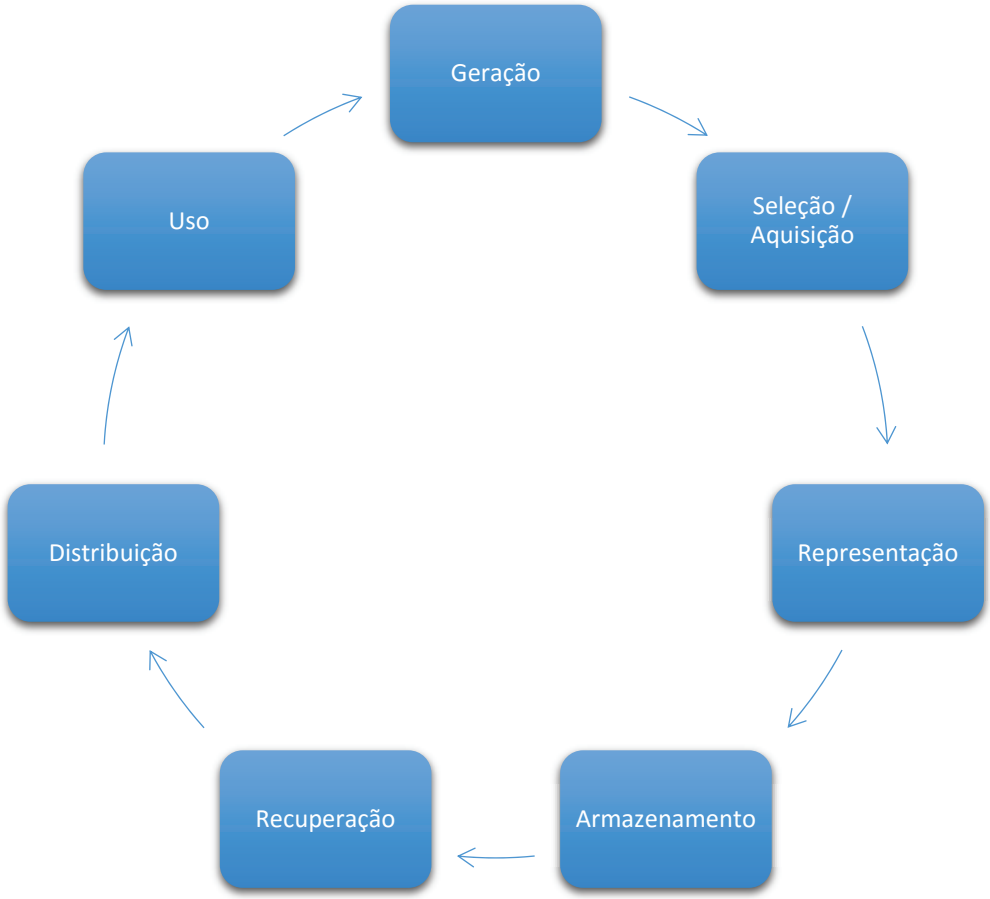
Quadro 1 - Diferença entre dados, informação e conhecimento

Dado	**Informação**	**Conhecimento**
Simples observações sobre o estado do mundo – Facilmente estruturado – Facilmente obtido por máquina – Frequentemente quantificado – Facilmente transferível	Dados dotados de relevância e propósito – Requer unidade de análise – Exige consenso em relação ao significado – Exige necessariamente a mediação humana	Informação valiosa da mente humana – Inclui reflexão, síntese, contexto – De difícil estruturação – De difícil captura em máquinas – Frequentemente tácito – De difícil transferência

FONTE: Davenport (1998, p. 18).

Tarapanoff (2006, p. 22) destaca que, de forma simples, pode-se definir a gestão da informação como a aplicação do ciclo da informação (processo da Ciência da Informação) às organizações, conforme Figura 3.

Figura 3 - Ciclo informacional



FONTE: Ponjuan Dante (1998, p. 47, apud Tarapanoff, p. 22).

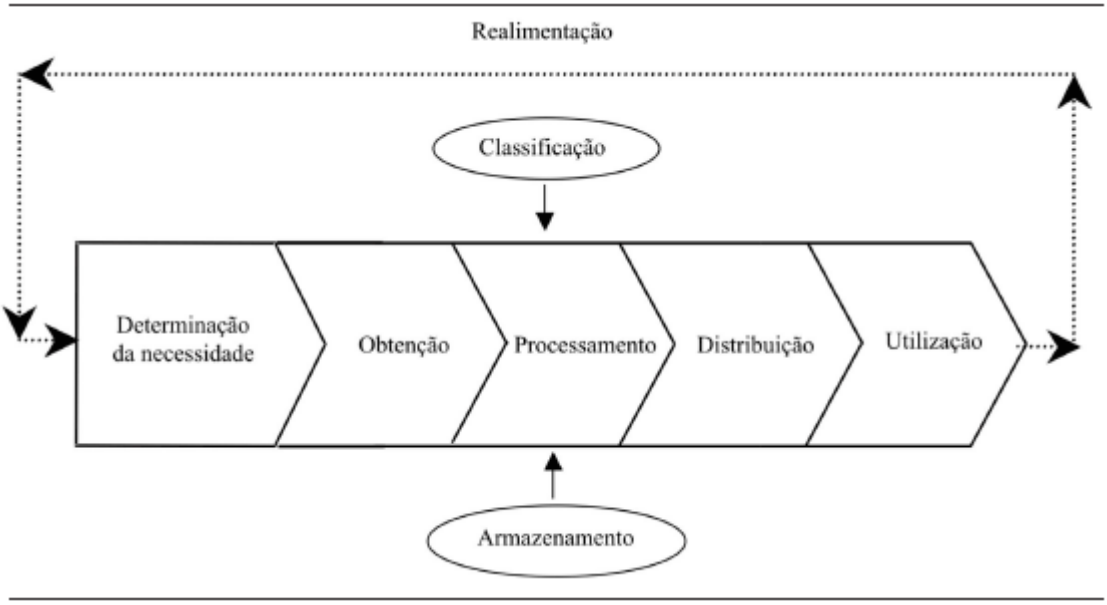
O início do ciclo informacional ocorre com a "identificação de uma necessidade informacional, um problema a ser resolvido, uma área ou assunto a ser analisado. É um processo que se inicia com a busca da solução a um problema, da necessidade de obter informações sobre algo, e passa pela identificação de quem gera o tipo de informação necessária, as fontes e o acesso, a seleção e aquisição, registro, representação, recuperação, análise e disseminação da informação, que, quando usada, aumenta o conhecimento individual e coletivo. (TARAPANOFF, 2006, p. 23).

O processo de gestão de informações inclui várias fases ou etapas, dependendo da abordagem com a qual se está trabalhando. Davenport (1998, p. 175)

apresenta um processo genérico de GI baseado em quatro passos: determinação das exigências, obtenção, distribuição e utilização.

Conforme Moraes e Filho (2006) as etapas relacionadas ao processo de gestão da informação podem ser sintetizadas em: determinação da necessidade de informação; obtenção; processamento; distribuição e apresentação; e utilização, conforme demonstra a Figura 4.

Figura 4 - Processo de gerenciamento da informação



FONTE: Moraes e Filho (2006)

Moraes e Filho (2006) definem os processos conforme detalhado na sequência.

A determinação da necessidade de informação envolve compreender as fontes e os tipos de informações necessárias para um bom desempenho do negócio, bem como suas características, fluxos e necessidades.

A obtenção inclui as atividades relacionadas à coleta dos dados. Davenport (1998) destaca que o processo mais eficaz é aquele que incorpora um sistema de aquisição contínua. O autor desmembra esse passo em várias atividades: exploração do ambiente informacional; classificação da informação em uma estrutura pertinente; e formatação e estruturação das informações.

O processamento compreende atividades de classificação (define o melhor modo de acessar as informações necessárias) e de armazenamento (seleciona o melhor lugar e os recursos para o arquivamento) das informações obtidas.

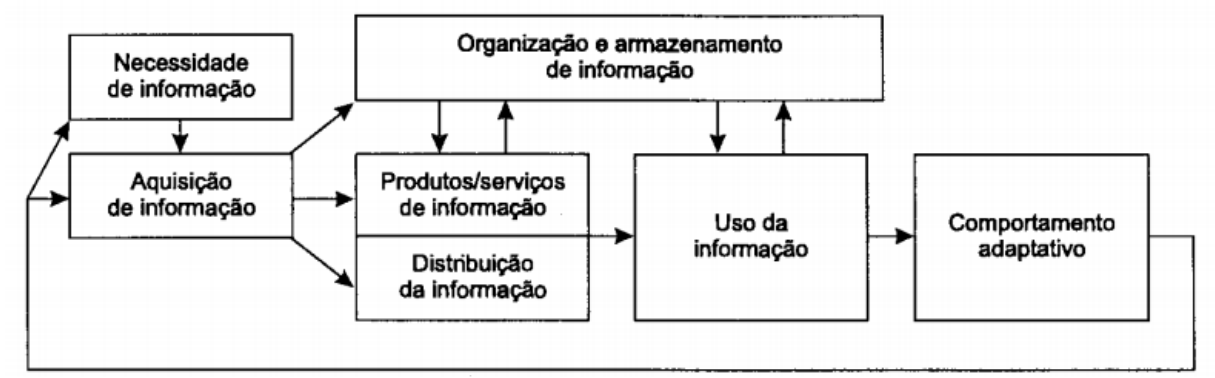
A distribuição e apresentação envolve escolher, entre diferentes metodologias, qual pode ser mais adequada para se apresentar a informação, disponibilizando-a aos usuários por diferentes formas e fontes e estilos.

Por fim, após a apresentação da informação, segue-se a etapa da sua utilização pelas pessoas da empresa, que as incorporarão às etapas de elaboração, execução e avaliação da estratégia empresarial, auxiliando, assim, o processo de gestão estratégica.

Ainda de acordo com os autores, após a última etapa, em que a informação foi utilizada e auxiliou na formulação da estratégia, uma nova demanda torna necessária a busca de informação, impulsionando o reinício do processo de gerenciamento da informação, já que esse processo, para ser estratégico, deve ser contínuo.

Em outra abordagem, Choo (2003, p 404) analisa a gestão da informação como um conjunto de seis processos correlatos: identificação das necessidades de informação; aquisição da informação; organização e armazenamento da informação; desenvolvimento de produtos e serviços de informação; distribuição da informação; e uso da informação. Assim, o processo é representado por um ciclo, conforme demonstra a Figura 5.

Figura 5 - Ciclo de Gestão da Informação



FONTE: Choo (2003, p. 404)

A necessidade de informação para Choo (2003, p. 405) nasce de problemas, incertezas e ambiguidades encontradas em situações e experiências específicas, de modo que não se deve preocupar-se com o significado da informação, mas sim com as condições, padrões e regras de uso, que a tornam significativas para determinados indivíduos em determinadas situações.

A aquisição da informação, segundo Choo (2003, p. 407), possui como princípio importante a variedade indispensável: as fontes para monitorar o ambiente devem ser suficientemente numerosas e variadas para refletir todo o espectro de fenômenos externos. Ainda de acordo com o autor, uma maneira eficaz de administrar a variedade de informações é envolver o maior número possível de pessoas na coleta de informações. O autor ressalta que a seleção e o uso das fontes de informação têm de ser planejados e continuamente monitorados e avaliados, como qualquer outro recurso vital para a organização.

A organização e armazenamento da informação, de acordo Choo (2003, p. 409), considera que parte da informação que é adquirida ou criada é fisicamente organizada e armazenada em arquivos, bancos de dados computadorizados e outros sistemas de informação, de modo a facilitar sua partilha e sua recuperação. A informação armazenada representa um componente importante e frequentemente consultado da memória da organização.

O desenvolvimento de produtos e serviços de informação, segundo Choo (2003, p. 412), consiste em garantir que as necessidades de informação dos membros da organização sejam atendidas com uma mistura de equilibrada de produtos e serviços. Para o autor, para darem resultados, os produtos e serviços de informação precisam abranger não apenas a área do problema, mas também as circunstâncias específicas que afetam a resolução de cada problema ou cada tipo de problema.

A distribuição da informação é definida por Choo (2003, p. 414) como o processo pelo qual as informações se disseminam pela organização, de maneira que "a informação correta atinja a pessoa certa no momento, lugar e formato adequados". Para o autor, uma ampla distribuição da informação pode acarretar muitas consequências positivas: o aprendizado organizacional torna-se mais amplo e mais frequente; a recuperação da informação torna-se mais provável; e novas informações podem ser criadas pela junção de itens esparsos. O autor afirma que o objetivo da distribuição da informação é promover e facilitar a partilha de informações, que é fundamental para a criação de significado, a construção de conhecimento e a tomada de decisões.

O uso da informação é identificado por Choo (2003, p. 415) como um processo social dinâmico de pesquisa e construção que resulta na criação de significado, na construção de conhecimento e na seleção de padrões de ação. Para o autor, a informação organizacional contém múltiplos significados, de modo que cada

representação é resultado de interpretações cognitivas e emocionais de indivíduos ou grupos. Por fim, Choo destaca que:

> O resultado do uso eficiente da informação é o comportamento adaptativo: a seleção e execução de ações dirigidas para objetivos, mas que também reagem às condições do ambiente. As reações da organização interagem com as ações de outras organizações, gerando novos sinais e mensagens aos quais se devem atentar e, dessa forma, mantendo novos ciclos de uso da informação. (CHOO, 2003, p. 404)

Considerando as abordagens apresentadas pelos autores, a seguir é analisado o enfoque na recuperação da informação, tendo em vista a sua influência para a tomada de decisão de forma estratégica.

2.2 RECUPERAÇÃO DA INFORMAÇÃO

A Recuperação da Informação (RI) trata da recuperação, armazenamento, organização e acesso a itens de informação, como documentos, páginas Web, catálogos *online*, registros estruturados e semiestruturados, objetos multimídia, etc. A representação e a organização dos itens devem fornecer aos usuários facilidade de acesso às informações de seu interesse (BAEZA-YATES; RIBEIRO-NETO, 2013).

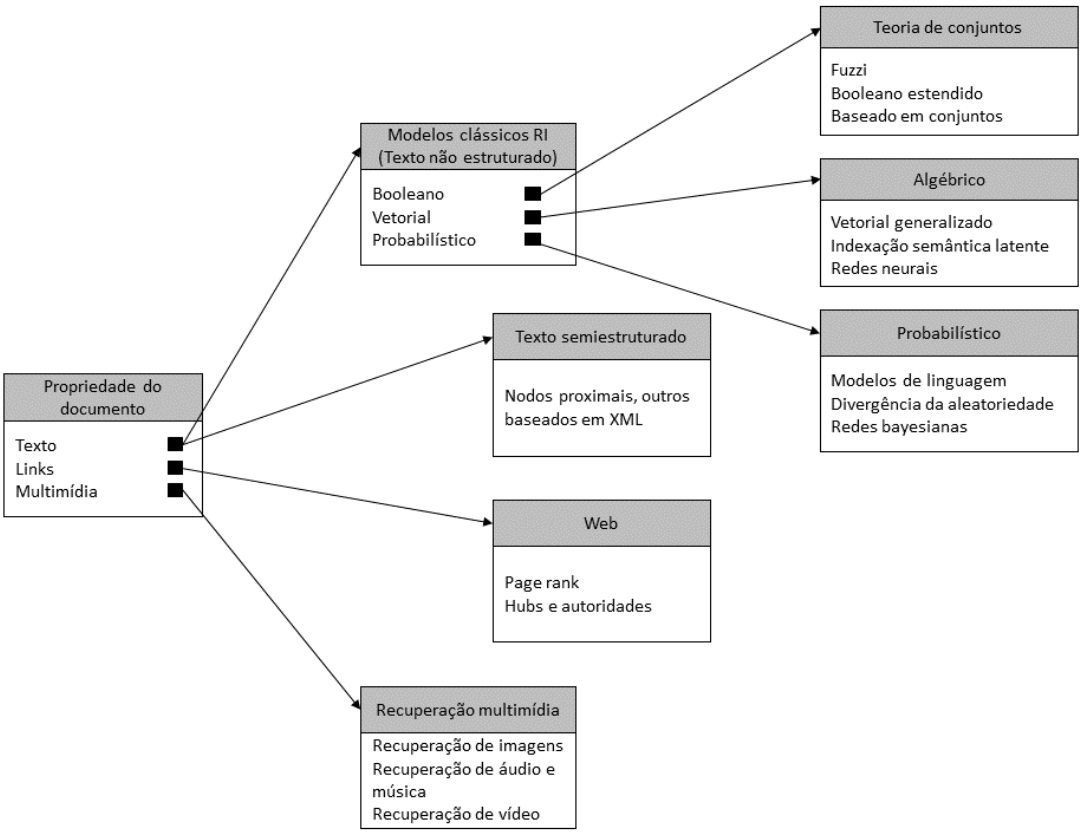
De acordo com Baeza-Yates e Ribeiro-Neto (2013) a RI pode ser estudada sob dois pontos de vistas distintos e complementares em termos de pesquisa: um centrado no computador e o outro centrado no usuário. Os autores afirmam que na visão centrada no computador, a RI consiste principalmente na construção de índices eficientes, no processamento de consultas com alto desempenho e no desenvolvimento de algoritmos de ranqueamento, a fim de melhorar os resultados. Já na visão centrada no usuário, os autores afirmam que a RI consiste principalmente em estudar o comportamento o usuário, entender suas principais necessidades e determinar como esse entendimento afeta a organização e operação do sistema na recuperação.

Ainda de acordo com os autores, os modelos de RI são fundamentalmente baseados em texto, isto é, eles usam o texto dos documentos para ranqueá-los em relação à consulta. Na *Web,* contudo, os autores destacam que também é necessário utilizar a informação sobre a estrutura de links para alcançar um bom rastreamento.

Porém, os autores explicam que os objetos multimídia não são da mesma forma que o texto, de modo que imagens são codificadas como *bitmaps* de *pixels*, vídeos são codificados como fluxos (*streams)* temporais de imagens e objetos de áudios são codificados como fluxos discretizados de som. Decorrentes dessas peculiaridades em suas formas de representação, os autores afirmam que os objetos multimídia são ranqueados de maneira diferente, ou então são recuperados sem ranqueamento. Dadas essas características, os autores distinguem três categorias de modelos de RI: baseadas em texto, as baseadas em links e as baseadas em objetos de multimídia.

A Figura 6 demonstra a taxonomia de modelos de RI elaborada pelos autores, bem como os modelos de recuperação de objetos de multimídia. Quanto aos modelos baseados em texto, os autores distinguem entre modelos para texto não estruturado e modelo que levam em conta a estruturação do texto.

Figura 6 - Uma taxonomia de modelos de RI



FONTE: Baeza-Yates e Ribeiro-Neto (2013).

Na primeira categoria do modelo, o texto é modelado simplesmente como uma sequência de palavras. Na segunda categoria, componentes estruturais do texto (como título, seções, subseções e parágrafos) são parte do modelo, que geralmente

é chamado de semiestruturado, porque ele também inclui texto não estruturado. Quanto ao texto não estruturado, os três modelos clássicos são chamados de Booleano, vetorial e probabilístico. No modelo Booleano, documentos e consultas são representados como conjuntos de termos de indexação. Assim, os autores esclarecem que o modelo é da teoria de conjuntos. No modelo vetorial, documentos e consultas são representados come vetores em um espaço com *t* dimensões. Por isso o modelo é considerado algébrico. No modelo probabilístico, o arcabouço para modelar as representações dos documentos e consultas é baseado na teoria das probabilidades. Dessa forma, como o nome indica, o modelo é probabilístico (BAEZA-YATES; RIBEIRO-NETO, 2013).

Os autores destacam que, ao longo dos anos, outros modelos de RI baseados nos modelos clássicos foram propostos. Quanto aos modelos alternativos baseados na teoria dos conjuntos, os autores distinguem entre o *fuzzy,* o booleano estendido e o baseado em conjuntos. Quanto aos modelos algébricos alternativos, os autores distinguem entre o modelo vetorial generalizado, a indexação semântica latente e o modelo de redes neurais. Quanto aos modelos probabilísticos alternativos, os autores distinguem o BM25, o de redes bayesianas, a divergência da aleatoriedade e os modelos de linguagem.

Ainda de acordo com os autores, em relação aos modelos para a recuperação de textos semiestruturados (isto é, modelos que lidam com a estrutura fornecida pelo texto), os autores consideram técnicas de indexação como os nodos proximais e os métodos de indexação baseados em XML.

Na *Web*, devido ao grande número de documentos (ou páginas *Web*), o ranqueamento baseado em texto por si só não é suficiente. Também é necessário considerar os *links* entre páginas Web como parte integrante do modelo. Isso leva aos modelos de recuperação baseados em *links*, particularmente o *PageRank* e o Hubs & Autoridades (BAEZA-YATES; RIBEIRO-NETO, 2013).

Os autores afirmam que um conjunto diferente de estratégias de recuperação é empregado para dados multimídia. Para recuperar imagens de interesse do usuário, são necessários vários passos intermediários que não são requeridos na busca em coleções textuais. Os autores exemplificam que, em vez de escrever uma consulta, o usuário pode especificar sua necessidade de informação apontando para uma dada imagem. Essa imagem consulta é comparada pelos autores às imagens da coleção para recuperar imagens relacionadas. Assim, os autores explicam que os métodos

para a recuperação multimídia são muito distintos dos métodos de RI para texto, pois, por exemplo, muitos deles não incluem nenhuma forma de ranqueamento. Por essa razão, os autores afirmam que eles são representados separadamente pelos autores na taxonomia.

Ainda de acordo com os autores, a forma mais simples de recuperação multimídia é a recuperação de imagens, porque as imagens são estáticas. Os autores explicam que no caso de áudio e vídeo, a representação dos objetos multimídia precisa incluir também uma dimensão temporal, o que torna os arquivos muito maiores e o problema mais difícil.

Considerando que o propósito principal de um modelo de RI é produzir um conjunto de resultados que provavelmente seja relevante para o usuário, implementações modernas de sistemas de RI incluem características de vários modelos de RI, e não de apenas um. Por exemplo, funções de ranqueamento na Web combinam características dos modelos clássicos de RI com características de modelos baseados em links para melhorar a recuperação (BAEZA-YATES; RIBEIRO-NETO, 2013).

O processo de recuperação da informação auxilia os tomadores de decisões a realizarem escolhas estratégicas com base em informações fundamentadas. A seguir será abordado sobre esse processo.

2.3 TOMADA DE DECISÃO

A informação auxilia no processo decisório, pois quando devidamente estruturada é de crucial importância para a empresa, associa os diversos subsistemas e capacita a empresa a impetrar seus objetivos. A informação é um instrumento de valor que, se trabalhada de forma eficaz, torna-se responsável por agregar ainda mais valor à organização. Esse valor deve ser medido pela liderança, a partir de uma análise de obtenção de resultado ocorrido através da informação transmitida (OLIVEIRA, 1992, apud VIEIRA, 2011).

A decisão é o processo de análise e escolha entre as alternativas disponíveis de cursos de ação que a pessoa deverá seguir. Toda decisão envolve seis elementos: o tomador de decisão: é a pessoa que faz uma escolha ou opção entre várias alternativas futuras de ação; os objetivos: são o que o tomador de decisão pretende alcançar com suas ações, as preferências: são os critérios que o tomador de decisão

usa para fazer sua escolha; a estratégia: é o curso de ação que o tomador de decisão escolhe para atingir seus objetivos dependendo dos recursos que pode dispor; a situação: são os aspectos do ambiente que envolve o tomador de decisão, alguns deles fora do seu controle, conhecimento ou compreensão e que afetam sua escolha; o resultado: é a consequência ou resultado de uma estratégia (CHIAVENATO, 2003, p. 348).

Chiavenato (2003, p. 349) destaca ainda que o processo decisório exige sete etapas: percepção da situação que envolve algum problema; análise e definições do problema; definição dos objetivos; procura de alternativas de solução ou de cursos de ação; escolha da alternativa mais adequada ao alcance dos objetivos; avaliação e comparação das alternativas; e implementação da alternativa escolhida.

A tomada de decisão é estudada sob duas perspectivas: a do processo e a do problema. A perspectiva do processo concentra-se na escolha dentre as possíveis alternativas de solução daquela que produza melhor eficiência. Dentro dessa perspectiva, o objetivo é selecionar a melhor alternativa de decisão. Focaliza o processo decisório como uma sequência de três etapas simples: definição do problema; quais as alternativas possíveis de solução do problema; e qual é a melhor alternativa de solução (escolha). Já a perspectiva do problema está orientada para a resolução de problemas. Sua ênfase está na solução final do problema. Essa perspectiva é criticada pelo fato de não indicar alternativas e pela sua deficiência quando as situações demandam vários modelos de implementação. Na perspectiva do problema, o tomador de decisão aplica métodos quantitativos para tornar o processo decisório o mais racional possível concentrando-se na definição e no equacionamento do problema a ser resolvido. Preocupa-se com a eficácia da decisão (CHIAVENATO, 2003, p. 442).

Maximiano (2000, p. 142) classifica as principais formas de decisões em razão dos diferentes problemas e situações que variam em termos de natureza, urgência, impacto sobre a organização e outros fatores, conforme demonstra o Quadro 2.

Quadro 2 - Descrição dos principais tipos de decisões

Tipo de decisão	Descrição
Programadas	Aplicam-se a problemas repetitivos
Não-programadas	Aplicam-se a problemas que não são familiares
Estratégicas	Escolhem objetivos para a organização
Administrativas	Colocam decisões estratégicas em prática
Operacionais	Definem meios e recursos
Individuais	São tomadas unilateralmente
Coletivas	São tomadas em grupo
Satisfatórias	Qualquer alterativa serve
Maximizadas	Procuram o melhor resultado possível
Otimizadas	Equilibram vantagens e desvantagens de diversas alternativas

FONTE: Maximiano (2000, p. 142)

Devido ao ambiente turbulento em que as empresas atualmente se encontram inseridas, o uso de informações imprecisas na tomada de decisões pode ser muito arriscado, uma vez que isso pode prejudicá-las quanto a sua produtividade, competitividade e até mesmo determinar a sua permanência ou não no mercado (CAETANO, 2000).

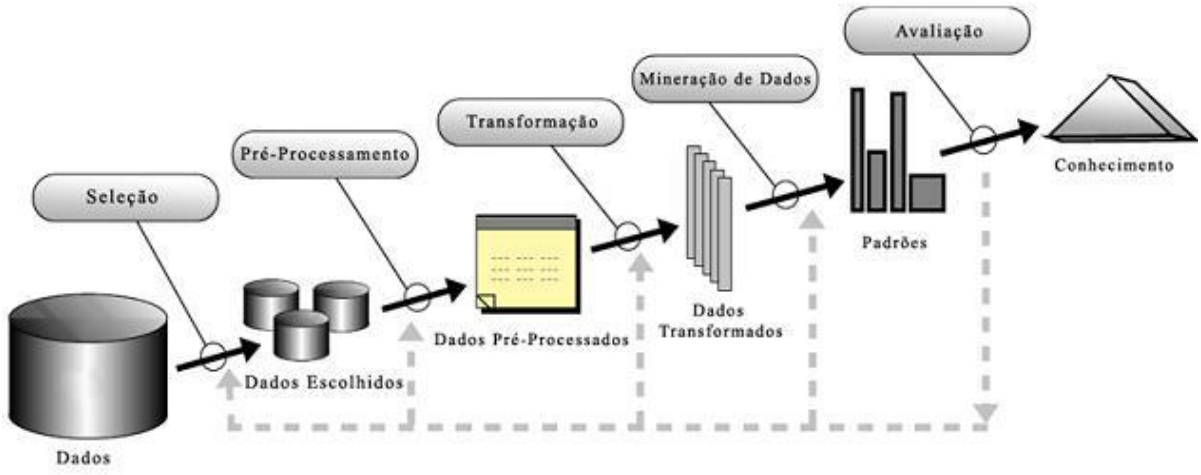
Caetano (2000) afirma que para apoiar a tomada de decisão ferramentas estão sendo utilizadas, de modo a auxiliar na análise dos problemas, bem como transformar informações complexas em informações relevantes, auxiliando na avaliação de resultados. A próxima subseção apresenta essa abordagem.

2.4 KDD (KNOWLEDGE DISCOVERY IN DATABASE)

O KDD (*Knowledge Discovery In Database* – ou Descoberta de Conhecimento em Bases de Dados, em português) consiste no processo de descoberta de padrões pela análise de grandes conjuntos de dados, tendo como principal etapa o processo de mineração, consistindo na execução prática de análise e de algoritmos específicos que, sob limitações de eficiência computacionais aceitáveis, produz uma relação particular de padrões a partir de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Fayyad, Piatetsky-Shapiro e Smyth (1996, p.41) afirmam que o processo de KDD é interativo e iterativo, envolvendo vários passos com muitas decisões tomadas pelo usuário. Os autores consideram o processo de KDD dividido em nove etapas, ilustradas na Figura 7: seleção, criação de dados alvo; pré-processamento dos dados; transformação; escolha da tarefa de mineração dos dados; escolha do algoritmo de mineração dos dados; mineração de dados; interpretação/avaliação dos resultados; e utilização dos padrões descobertos.

Figura 7 - Processos de KDD



FONTE: Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 41)

Fayyad, Piatetsky-Shapiro e Smyth (1996) explicam cada um dos passos conforme detalhamento apresentado na sequência:

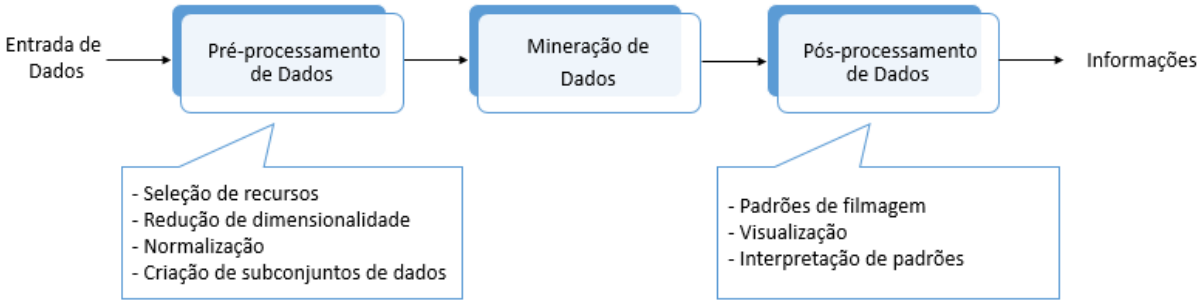
1. o primeiro passo consiste no conhecimento do domínio da aplicação: inclui o conhecimento relevante e as metas do processo KDD para a aplicação;
2. o segundo passo consiste na criação de um banco de dados alvo: inclui selecionar um conjunto de dados ou dar ênfase para um subconjunto de variáveis ou exemplo de dados nos quais o "descobrimento" será realizado;
3. o terceiro passo consiste na limpeza de dados e pré-processamento: inclui operações básicas como remover ruídos, coleta de informação necessária para modelagem, decidir estratégias para manusear (tratar) campos perdidos etc;

4. o quarto passo consiste na redução de dados e projeção: inclui encontrar formas práticas para se representar dados, dependendo da meta do processo e o uso de redução dimensionável e métodos de transformação para reduzir o número efetivo de variáveis que deve ser levado em consideração ou encontrar representações invariantes para os dados;
5. o quinto passo consiste na escolha da tarefa de mineração de dados inclui a decisão do propósito do modelo derivado do algoritmo de mineração de dados (Ex. classificação, regressão, regras de associação e agrupamento);
6. o sexto passo consiste em encontrar o algoritmo de mineração de dados: inclui selecionar métodos para serem usados para procurar por modelos nos dados, como decidir quais modelos e parâmetros podem ser apropriados e determinar um método de mineração de dados particular como modelo global do processo KDD (Ex. o usuário pode estar mais preocupado em entender o modelo do que nas suas capacidades);
7. o sétimo passo consiste na interpretação: inclui a interpretação do modelo descoberto e possível retorno a algum passo anterior como também uma possível visualização do modelo extraído, removendo modelos redundantes ou irrelevantes e traduzindo os úteis em termos compreendidos pelos usuários;
8. o oitavo passo consiste na utilização do descobrimento obtido: inclui incorporar este conhecimento no desempenho do sistema, tomando ações baseadas no conhecimento, ou simplesmente documentando e reportando para grupos interessados;
9. por fim, o nono passo consiste em agir sobre o conhecimento descoberto: inclui usar o conhecimento diretamente, incorporando-o em outro sistema de novas ações, ou simplesmente documentá-lo e denunciá-lo às partes interessadas. Este processo também inclui a verificação e resolução de possíveis conflitos com o conhecimento anteriormente extraído.

Calil et al. (2008) complementam que as fases do KDD podem ser agrupadas em três grandes grupos: pré-processamento, mineração de dados e pós-

processamento. De acordo com os autores, o pré-processamento inclui todas as etapas que consideram a preparação da base de dados, cujos dados serão fornecidos como entrada para o (s) algoritmo (s) de Mineração de Dados. Ainda de acordo com os autores, o pós-processamento contempla a depuração e/ou síntese dos padrões descobertos. Tan, Steinbach e Kumar (2009, p. 4) complementam que o processo consiste em uma série de passos de transformação, desde o pré-processamento dos dados até o pós-processamento, resultado da mineração de dados, conforme demonstra a Figura 8.

Figura 8 - O processo de descoberta de conhecimento em bancos de dados



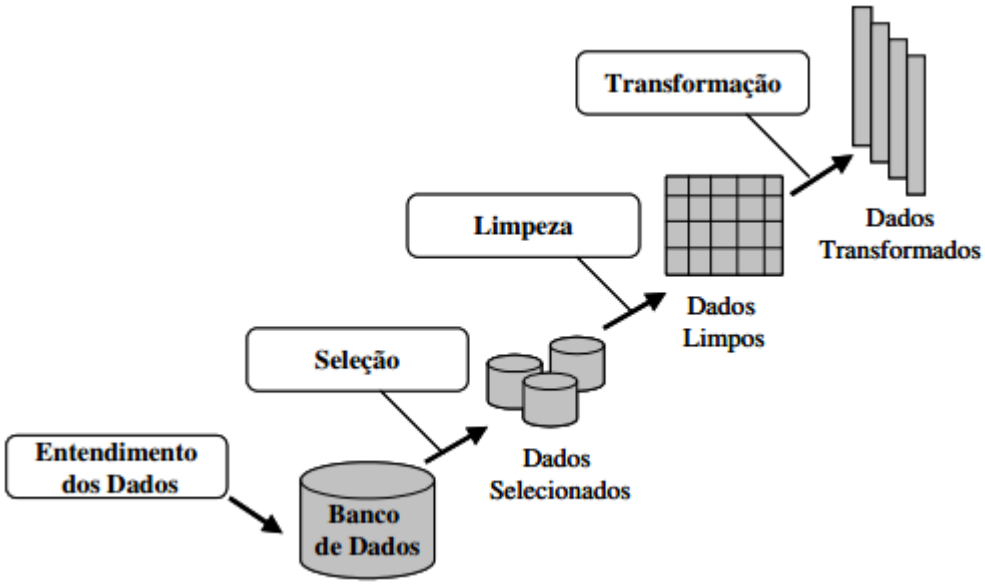
FONTE: Tan, Steinbach e Kumar (2009, p. 4)

A seguir são abordados esses três grandes grupos.

2.5 PRÉ-PROCESSAMENTO

De acordo com Neves (2003), a fase de pré-processamento é composta pelas subfases: entendimento, seleção, limpeza e transformação de dados, conforme demonstra a Figura 9.

Figura 9 - Subfases de pré-processamento



FONTE: Neves (2003)

A Autora define os processos conforme explicado na sequência.

O entendimento dos dados consiste em analisar os dados fornecidos pelos especialistas, entendendo do que se tratam as tabelas envolvidas, o significado, relevância, formato, tamanho e tipo de dado dos atributos; identificando os atributos chaves; realizando levantamentos estatísticos e verificando a qualidade dos dados.

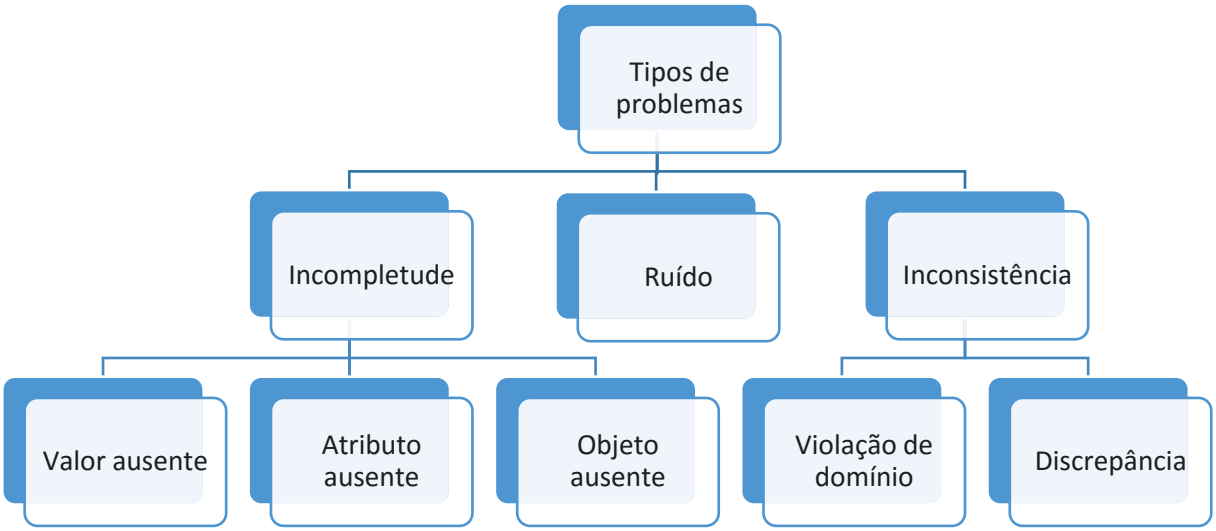
A seleção de dados envolve a escolha da (s) tabela (s), atributos e instâncias da (s) mesma (s) em relação aos objetivos do usuário, considerando-se ainda que, na necessidade de se manipular informações de várias tabelas cabe a integração das mesmas, de modo a obter-se um conjunto único de instâncias sobre o qual será dada a continuidade do pré-processamento e/ ou do processo KDD.

A limpeza de dados refere-se a garantia da qualidade dos dados que pode ser obtida através de algumas operações tais como: padronização de dados, tratamento de valores ausentes, eliminação de dados errôneos e de duplicatas.

A transformação de dados corresponde a operações que tornem a apresentação dos dados apropriada a técnica de mineração de dados a ser utilizada. Assim encontram-se descritas operações do tipo normalização de dados, conversões de valores simbólicos para valores numéricos, discretização e composição de atributos.

Castro e Ferrari (2016, p. 27) destacam que existem basicamente três tipos de problemas com dados (Figura 10).

Figura 10 - Principais problemas com os dados



FONTE: Castro e Ferrari (2016, p. 27)

A incompletude de uma base de dados pode ocorrer de várias formas, por exemplo, podem faltar valores de um dado atributo; pode faltar um atributo de interesse; ou pode faltar um objeto de interesse. Entretanto, nem sempre a ausência de um atributo ou objeto é percebida, a não ser quando um especialista no domínio do problema analisa a base e percebe a falta (CASTRO; FERRARI, 2006, p. 26).

Um dado inconsistente ocorre quando diferentes e conflitantes versões do mesmo dado aparecem em locais variados. Na área de mineração de dados, um dado inconsistente é aquele cujo valor está fora do domínio do atributo ou apresenta grande discrepância em relação aos outros dados. Exemplos comuns de inconsistência ocorrem quando se consideram diferentes estados de medidas ou notação, como é o caso de pesos dados em quilos (kg) ou em libras (£), e distância dadas em metros ou quilômetros (CASTRO; FERRARI, 2006, p. 26).

O ruído possui diversos significados, dependendo do contexto. Por exemplo, em vídeo, um ruído é aquele chuvisco na imagem e, em rádio, é aquela interferência no sinal de áudio. Entretanto, a noção de ruído em mineração de dados está mais próxima do conceito de ruído em estatística (variações inexplicáveis em uma amostra) e processamento de sinais (variações indesejadas e normalmente inexplicáveis em um sinal). Um dado ruidoso é aquele que apresenta alguma variação em relação ao seu valor sem ruído e, portanto, ruídos na base de dados podem levar a

inconsistências. Dependendo do nível de ruído, nem sempre é possível saber se ele está ou não presente em um dado (CASTRO; FERRARI, 2016, p. 27).

Ainda de acordo com os autores, de forma simplificada, dados são valores quantitativos ou qualitativos associados a alguns atributos. Com relação a estrutura, os autores afirmam que eles podem ser: estruturados, semiestruturados ou não estruturados.

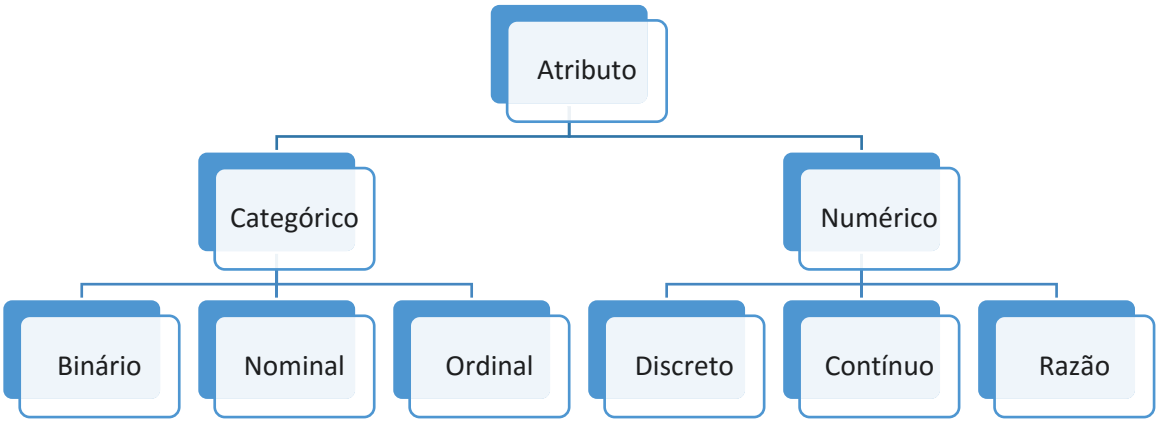
De acordo com os autores, os dados estruturados dependem da criação de um modelo de dados, ou seja, a descrição dos objetos juntamente com as suas propriedades e relações. Os autores afirmam que uma base de dados é estruturada quando os dados residem em campos fixos em um arquivo – por exemplo, uma tabela, uma planilha ou um banco de dados. Uma das vantagens dos dados estruturados apresentada pelos autores é a facilidade de armazenagem, acesso e análise.

Os dados semiestruturados são definidos pelos autores como tipos de dados que não possuem a estrutura completa de um modelo de dados, mas também não é totalmente desestruturado. Os autores explicam que nos dados semiestruturados geralmente são usados marcadores (por exemplo, *tags)* para identificar certos elementos dos dados, mas a estrutura não é rígida.

Por fim, o dado não estruturado é definido pelos autores como aquele que não possui um modelo de dados, que não está organizado de uma maneira predefinida ou que não reside em locais definidos. Os autores complementam que geralmente se refere a textos livres, imagens, vídeos, sons, páginas web, arquivos PDF, entre outros. Os autores acrescentam que costumam ser de difícil estruturação, acesso e análise.

Castro e Ferrari (2016, p. 30) definem ainda os tipos de atributos (Figura 11). De acordo com os autores, o valor de um atributo de um dado objeto é uma medida de quantidade daquele atributo, podendo ser numérica ou categórica. Os atributos numéricos podem assumir quaisquer valores numéricos – por exemplo, valores discretos (inteiros) ou contínuos (reais) ao passo que as quantidades categóricas assumem valores correspondentes a símbolos distintos.

Figura 11 - Tipos de dados



FONTE: Castro e Ferrari (2016, p. 30)

De acordo com os autores, o atributo binário é aquele que pode assumir apenas dois valores possíveis – por exemplo, “0” ou “1”.

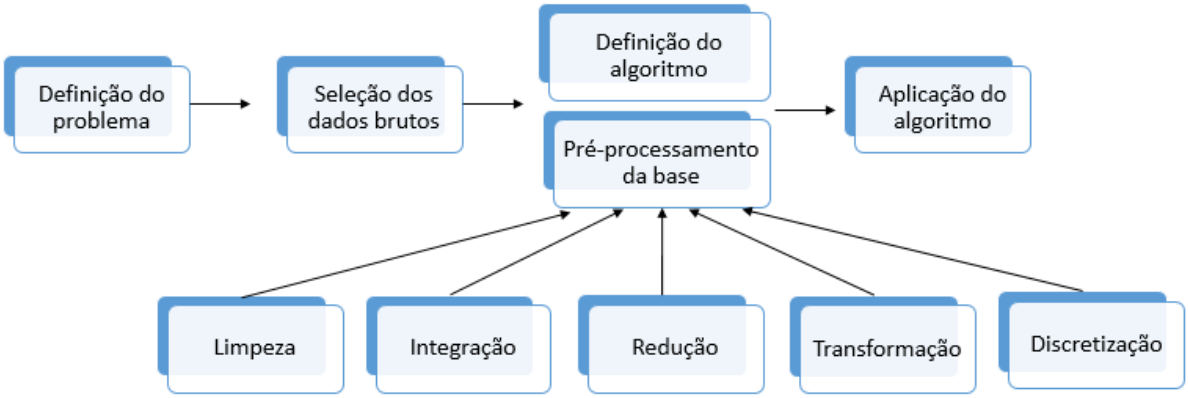
O atributo nominal é definido pelos autores como aquele cujos valores possuem símbolos ou rótulos distintos - por exemplo: o atributo “estado civil” pode assumir, por exemplo, os valores “solteiro”, “casado”, “separado”, “divorciado” e “viúvo”.

O atributo ordinal é definido pelos autores como aquele que permite ordenar suas categorias, embora não necessariamente haja uma noção explícita de distância entre as categorias - por exemplo: o atributo “nível educacional” pode assumir valores “primário”, “secundário”, “graduação”, “especialização”, “mestrado” e “doutorado”.

O atributo razão é definido pelos autores como aquele para o qual o método de medida define o ponto zero - por exemplo: a distância entre dois objetos possui naturalmente o zero quando ambos são iguais.

Os autores apresentam ainda uma visão abrangente do processo de preparação da base de dados para análise (Figura 12).

Figura 12 - Etapas do processo de preparação da base de dados



FONTE: Castro e Ferrari (2016, p. 35)

Para Castro e Ferrari (2016, p. 35) as principais atividades de pré-processamento são: limpeza, integração, redução, transformação e discretização. Os autores definem as atividades da seguinte forma:

A limpeza ocorre para imputação de valores ausentes, remoção de ruídos e correção de inconsistências. A integração é realizada para unir dados de múltiplas fontes em um único local, como um armazém de dados (*data warehouse*). A redução é realizada para reduzir a dimensão da base de dados, por exemplo, agrupando ou eliminando atributos redundantes, ou para reduzir a quantidade de objetos da base, sumarizando os dados. A transformação é realizada, segundo os autores, para padronizar e deixar os dados em um formato passível de aplicação das diferentes técnicas de mineração. Por fim, a discretização é utilizada para permitir que métodos que trabalham apenas com atributos nominais possam ser empregados a um conjunto maior de problemas. Também faz com que a quantidade de valores para um dado atributo (contínuo) seja reduzida.

2.6 MINERAÇÃO DE DADOS

A mineração de dados (ou *Data mining*, em inglês) é um dos principais passos no processo de KDD, tendo sido utilizada para melhorar sistemas de recuperação de informações. A mineração de dados corresponde parte da descoberta de conhecimento em bases de dados (KDD). Surgiu a partir da necessidade de desenvolver ferramentas mais eficientes e escaláveis que pudessem lidar com diversos tipos de dados. Assim, os trabalhos que culminaram na área de mineração

de dados constituíram-se sobre a metodologia e algoritmos que pesquisadores já haviam utilizado anteriormente, como a (1) amostragem, estimativa e teste de hipótese a partir de estatísticas e (2) algoritmos de busca, técnicas de modelagem e teorias de aprendizagem da inteligência artificial, reconhecimento de padrões e aprendizagem de máquina. A mineração de dados também foi rápida em adotar ideias de outras áreas, incluindo a otimização, computação evolutiva, teoria da informação, processamento de sinais, visualização e recuperação de informações (TAN; STEINBACH; KUMAR, 2009, p. 7).

O objetivo da mineração de dados é a extração de conhecimento implícito por meio da descoberta de padrões e regras significativas, a partir de grande quantidade de dados armazenados, de forma automática ou semiautomática, utilizando modelos computacionais construídos para descobrir novos fatos e relacionamentos entre dados, de forma repetida e interativa (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

São típicas aplicações da mineração de dados para análise e predição de crédito, detecção de fraudes, predição do mercado financeiro, relacionamento com clientes, predição de falência corporativa e muitas outras. Exemplos de segmentos de aplicação incluem setor financeiro; planejamento estratégico empresarial; planejamento do setor portuário; setores de energia (petróleo, gás, energia elétrica, biocombustíveis, etc.); educação; logística; planejamento das cadeias de produção, distribuição e suprimentos; meio ambiente; e internet (portais, redes sociais, comércio eletrônico, etc.). Aplicações típicas incluem identificação ou segmentação de clientes, parceiros, colaboradores; detecção de fraudes e anomalias em sistemas e processos; ações estratégicas de *marketing*, CRM (*Customer Relationship Management* ou Gestão de Relacionamento com o Cliente, em português*)* e RH (Recursos Humanos); jogos e atividades educacionais; gestão do conhecimento; análise de padrões de consumo; compreensão de bases de dados industriais, biológicas, empresariais e acadêmicas; predição de retorno sobre investimento, despesas, receitas, investimentos etc.; e mineração de dados da *web* (CASTRO; FERRARI, 2016, p. 17).

Tan, Steinbach e Kumar (2009, p. 8) demonstram um relacionamento da mineração e dados com as outras áreas, como os sistemas de bancos de dados que são necessários para fornecer um eficiente suporte de armazenamento, indexação e processamento de consultas; técnicas de computação de alto desempenho que são importantes para lidar com o tamanho volumoso de alguns conjuntos de dados; e as

técnicas de distribuição, que podem auxiliar abordar a questão do tamanho e são essenciais quando os dados não podem ser juntados em um único local. Castro e Ferrari (2016, p. 7) complementam que a mineração de dados envolve conhecimento de áreas como banco de dados, estatística, aprendizagem de máquina, computação de lato desempenho, reconhecimento de padrões, computação natural, visualização de dados, recuperação da informação, processamento de imagens e de sinais, conforme demonstra a Figura 13.

Figura 13 - Multidisciplinaridade da mineração de dados

FONTE: Castro e Ferrari (2016, p. 7).

Tan, Steinbach e Kumar (2009, p. 26) também abordam em seu livro os diferentes tipos de dados. Para os autores, um conjunto de dados muitas vezes pode ser visto como uma coleção de objetos de dados. Segundo os autores, eles são descritos por um número de atributos que capturam as características básicas de um objeto, como a massa de um objeto físico ou o tempo no qual um evento tenha ocorrido.

Um atributo é definido pelos autores como uma propriedade ou característica de um objeto de pode variar, seja de um objeto para outro ou de tempo para outro. Os

autores afirmam que uma forma de especificar o tipo de um atributo é identificar as propriedades de números que correspondem às propriedades relacionadas do atributo. Os autores destacam que as seguintes propriedades (operações) de números são geralmente usadas para descrever atributos: distinção (= e ≠); ordenação (<; ≤; > e ≥); adição (+ e -); e multiplicação (* e /).

Os autores complementam que dadas essas propriedades é possível definir quatro tipos de atributos: nominal, ordinal, intervalar e proporcional. As definições dos tipos de atributos e as informações sobre as operações estatísticas que são válidas para cada tipo são identificadas pelos autores no Quadro 3.

Quadro 3 - Descrição, exemplos e operações dos diferentes tipos de atributos

Tipo de atributo		**Descrição**	**Exemplos**	**Operações**
Categorizados (Qualitativos)	Nominal	Os valores de um atributo nominal são apenas nomes diferentes; ou seja, valores nominais fornecem apenas informação suficiente para distinguir um objeto do outro (=, ≠)	Códigos postais, números de ID de funcionários, cor dos olhos, sexo	Modo, entropia, correlação de contingência, teste x^2
	Ordinal	Os valores de um atributo ordinal fornecem informações suficiente para ordenar objetos. (>.<)	Dureza de minerais {boa, melhor, melhor de todas}, notas, números de ruas	Medianas, porcentagens, testes de execução, testes de assinatura
Numéricos (Quantitativo)	Intervalar	Para atributos intervalares, as diferenças entre os valores são significativas, ou seja, existe uma unidade de medida (+,-)	Datas de calendário, temperatura em Celsius ou Fahrenheit	Média, desvio padrão, correlação de Pearson, testes T e F.
	Proporcional	Para variáveis proporcionais, tanto as diferenças quanto as proporções são significativas (*, /)	Temperatura em Kelvin, quantidades monetárias, contadores, idade, massa, comprimento, corrente elétrica	Média geométrica, média harmônica, variação percentual

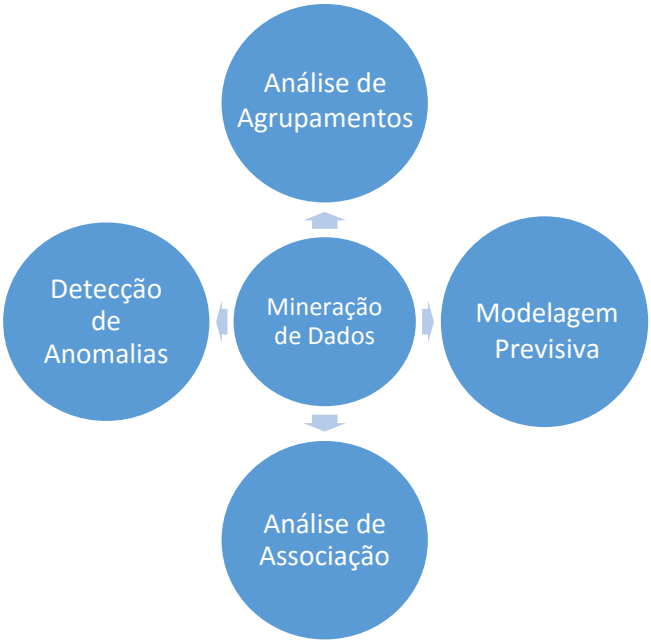
FONTE: Tan, Steinbach e Kumar (2009, p.32)

De acordo com os autores, os atributos nominais e ordinais são chamados coletivamente de atributos categorizados ou qualitativos. Assim, os autores afirmam que não possuem a maioria das propriedades dos números. Os atributos intervalar e proporcional, segundo os autores, são chamados coletivamente de atributos quantitativos ou numéricos. Os autores complementam que os atributos quantitativos são representados por números e possuem a maioria das propriedades do mesmo.

As tarefas de mineração de dados são geralmente divididas em duas categorias principais: preditivas e descritivas. As tarefas preditivas apresentam como objetivo prever o valor de um determinado atributo baseado nos valores de outros atributos. O atributo a ser previsto é comumente conhecido como a variável dependente ou alvo, enquanto que os atributos para fazer a previsão são conhecidos como as variáveis independentes ou explicativas. Já as tarefas descritivas apresentam como objetivo derivas padrões (correlações, tendências, grupos, trajetórias e anomalias) que resumam os relacionamentos subjacentes nos dados. As tarefas descritivas da mineração de dados são muitas vezes exploratórias em sua natureza e frequentemente requerem técnicas de pós-processamento para validar e explicar os resultados (TAN; STEINBACH; KUMAR, 2009, p. 8)

A Figura 14 ilustra quatro das tarefas centrais da mineração de dados de acordo com a abordagem de Tan, Steinbach e Kumar (2009, p. 9).

Figura 14 - Tarefas de mineração de dados comumente encontradas na literatura



FONTE: Adaptado de Tan, Steinbach e Kumar (2009, p. 9)

A modelagem de previsão se refere à atividade de construir um modelo para a variável alvo (também conhecida por meta ou objetivo) como uma função das atividades explicativas. Há dois tipos de tarefas de modelagem de previsão: classificação, a qual é usada para variáveis-alvo discretas, e regressão, que é usada para variáveis-alvo contínuas. A modelagem de previsão pode ser usada para identificar clientes que responderão a uma campanha de vendas, prever perturbações no ecossistema da Terra ou julgar se um paciente possui uma determinada doença baseado nos resultados de exames médicos (TAN; STEINBACH; KUMAR, 2009, p. 9).

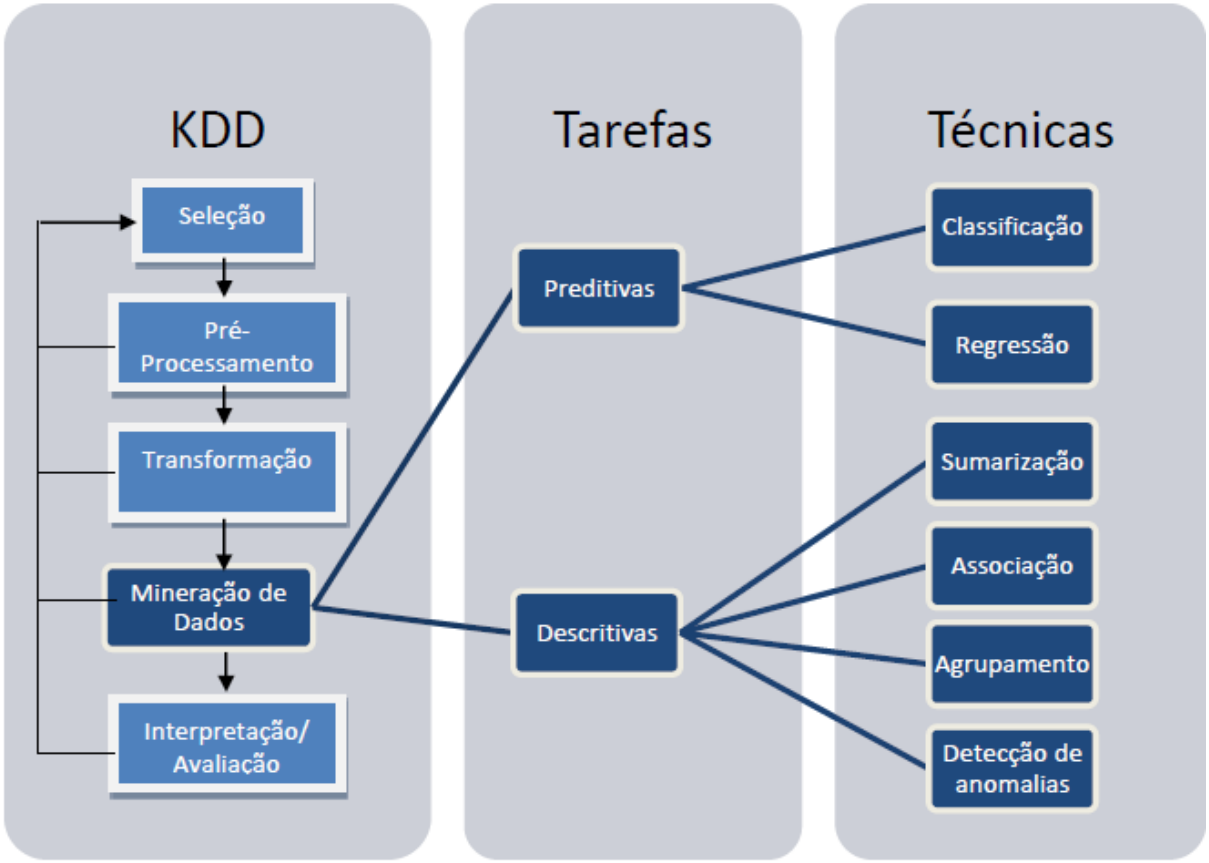
A análise de associação é usada para descobrir padrões que descrevam características altamente associadas dentro dos dados. Os padrões descobertos são normalmente representados na forma de regras de implicação ou subconjuntos de características. Devido ao tamanho exponencial de seu espaço de busca, o objetivo da análise de associação é extrair os padrões mais interessantes de forma mais eficiente. Aplicações úteis de análise de associação incluem a descoberta de genes que possuam funcionalidades associada, a identificação de páginas Web que sejam acessadas juntas ou a compreensão dos relacionamentos entre diferentes elementos do sistema climático da Terra (TAN; STEINBACH; KUMAR, 2009, p. 11).

A análise de grupo procura encontrar grupos de observações intimamente relacionadas de modo que observações que pertençam ao mesmo grupo sejam mais semelhantes entre si do que com as que pertençam a outros grupos. O a grupamento tem sido usado para juntas conjuntos de clientes relacionados, descobrir áreas do oceano que possuam um impacto significativo sobre o clima da Terra e compactar dados (TAN; STEINBACH; KUMAR, 2009, p. 12).

Por fim, a detecção de anomalias é a tarefa de identificar observações cujas características sejam significativamente diferentes do resto dos dados. Tais observações são conhecidas como anomalias ou fatores estranhos. O objetivo de um algoritmo de detecção de anomalias é descobrir as anomalias verdadeiras e evitar rotular erroneamente objetos normais como anômalos. Em outras palavras, um bom detector de anomalias deve ter uma alta taxa de detecção e uma baixa taxa de alarme falso. As aplicações da detecção de anomalias incluem a detecção de fraudes, intromissões na rede, padrões incomuns de doenças e perturbações no meio ambiente (TAN; STEINBACH; KUMAR, 2009, p. 13).

Complementando essa abordagem, Nogueira (2015) destaca a mineração de dados como um importante passo dentro do processo de descoberta de conhecimento, mostrando suas tarefas e as técnicas mais comuns por meio das quais as tarefas podem ser implementadas, conforme demonstra a Figura 15

Figura 15 - Técnicas de mineração de dados



FONTE: Nogueira (2015)

Considerando as abordagens anteriormente apresentadas, na próxima seção são analisdas as principais tarefas de mineração de dados. Destaca-se que é realizada apenas uma abordagem introdutória às tarefas, de modo que para obter o conteúdo completo sobre os algoritmos, bem como exemplos de aplicação e exercícios resolvidos, recomenda-se acessar o livro: “Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações”, de Castro e Ferrari, tendo em vista que o mesmo é voltado para estudantes e profissionais das ciências exatas, humanas e sociais aplicadas.

2.6.1 Análise de Grupos

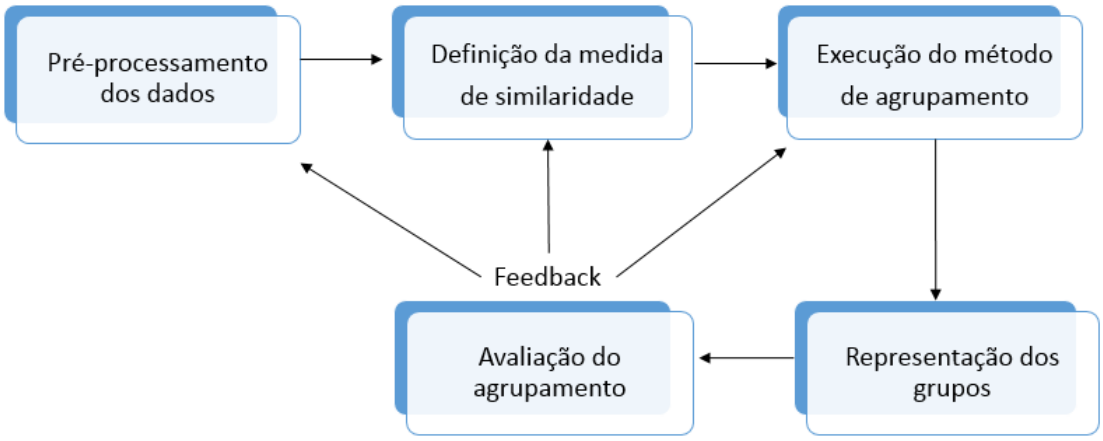
A análise de grupos, também conhecida como agrupamento de dados, é um termo genérico usado para designar um amplo espectro de métodos numéricos de análise de dados multivariados com o objetivo de descobrir grupos homogêneos de objetos. O agrupamento de objetos em diferentes grupos pode simplesmente representar uma forma conveniente de organizar grandes bases de dados de maneira que elas sejam mais facilmente compreendidas ou pesquisadas e, também, para realizar tarefas muito mais sofisticadas, como tomada de decisão em processos críticos (CASTRO; FERRARI, 2016, p. 87).

A análise de grupos é definida por Castro e Ferrari (2016, p. 87) como a organização de um conjunto de objetos (normalmente representados por vetores de características, ou seja, pontos em um espaço multidimensional) em grupos baseada na similaridade entre eles. Ainda de acordo com os autores, dito de outra forma, agrupar objetos é o processo de particionar um conjunto de dados em subconjuntos (grupos) de forma que os objetos em cada grupo (idealmente) compartilhem características comuns, em geral proximidade em relação a alguma medida de similaridade ou distância.

O agrupamento de dados é definido pelos autores como uma técnica comum em análise de dados que é utilizada em diversas áreas, incluindo aprendizagem de máquina, mineração de dados, reconhecimento de padrões, análise de imagens e bioinformática.

Castro e Ferrari (2016, p. 96) dividem o processo de agrupamento de dados e cinco etapas principais, conforme demonstra a Figura 16.

Figura 16 - Processo de agrupamento de dados



FONTE: Castro e Ferrari (2016, p. 96)

De acordo com os autores, a etapa de pré-processamento de dados, que consiste na preparação da base para o agrupamento, pode envolver todas as etapas típicas de pré-processamento de dados, como limpeza, integração, redução, transformação e discretização.

A etapa de definição da medida de similaridades (proximidade) ou dissimilaridade (distância) entre objetos é utilizada, segundo os autores, durante o agrupamento propriamente dito.

Na etapa de execução os autores afirmam que os métodos de agrupamento podem ser divididos em hierárquicos ou particionais. Os autores destacam que os métodos hierárquicos criam uma decomposição hierárquica dos dados, enquanto os métodos particionais, dado um conjunto *n* objetos, um método particional constrói *k* partições de dados, sendo que cada partição representa um *cluster* ($k \leq n$).

A representação dos grupos é definida pelos autores como o processo de extrair uma representação simples e compacta dos grupos obtidos a partir do agrupamento da base. Os autores destacam que as formas típicas de representação dos grupos são: protótipos, estruturas em grafos, estruturas em árvore e rotulação.

A etapa de avaliação do agrupamento depende, de acordo com os autores, do contexto e dos objetivos da análise. Os autores afirmam que a saída do algoritmo de agrupamento pode ser avaliada com relação à qualidade do agrupamento, o que pode ser feito por uma medida de avalição externa – isto é, os grupos encontrados são comparados com uma estrutura de agrupamento conhecida a priori ou uma avaliação

interna, ou seja, tenta-se determinar se a estrutura encontrada pelo algoritmo é apropriada aos dados.

Tan, Steinbach e Kumar (2009, p. 582) fornecem alguns exemplos específicos com base em duas categorias: agrupamento para compreensão e agrupamento por utilidade.

No agrupamento para compreensão, segundo os autores, classes ou grupos conceitualmente significativos de objetos que compartilham características comuns, desempenham um papel importante em como as pessoas analisam e descrevem o mundo. De acordo com os autores, são alguns exemplos de aplicação:

- biologia: biólogos aplicam o agrupamento para analisar as grandes quantidades de informações genéticas que agora estão disponíveis. Por exemplo, o agrupamento tem sido utilizado par encontrar grupos de genes que tenham funções semelhantes;
- recuperação de informações: o agrupamento pode ser usado para agrupar os resultados de pesquisa em um número pequeno de grupos, cada um dos quais captura um determinado aspecto da consulta. Cada categoria (grupo) pode ser dividida em subcategorias (subgrupos), produzindo uma estrutura hierárquica que auxilie mais a exploração de um usuário nos resultados da consulta;
- clima: a análise de grupos tem sido aplicada para encontrar padrões na pressão atmosférica de regiões polares e áreas do oceano que tenham um impacto significativo sobre o clima da terra;
- psicologia e medicina: a análise de agrupamentos pode ser usada para identificar diferentes subcategorias de uma doença ou condição. Por exemplo, o agrupamento tem sido usado para identificar diferentes tipos de depressão. Também pode ser usado para detectar padrões na distribuição espacial ou temporal de uma doença;
- negócios: o agrupamento pode ser usado para segmentar clientes em um número menor de grupos para análise adicional e atividades de marketing.

No agrupamento por utilidade, os autores destacam que a análise de grupos é o estudo de técnicas para encontrar os protótipos de grupos mais representativos. De acordo com os autores, são alguns exemplos de aplicação:

- resumo: muitas técnicas de análise de dados têm uma complexidade de tempo ou espaço, não sendo práticas para conjuntos grandes de dados. Nesse sentido, em vez de aplicar o algoritmo ao conjunto de dados inteiro, ele pode ser aplicado a um conjunto de dados reduzido consistindo de apenas protótipos de grupos;
- compactação: protótipos de grupos também podem ser usados para a compactação de dados;
- encontrando eficientemente os vizinhos mais próximos: se os objetos estiverem relativamente próximos do protótipo de seus grupos, então é possível usar os protótipos para reduzir o número de cálculos de distância que são necessários para encontrar o vizinho mais próximo de um objeto.

2.6.1.1 Algoritmos de agrupamento

Entre os principais algoritmos de agrupamento presentes na literatura, Castro e Ferrari (2016) destacam: algoritmo k-médias; algoritmo k-medoides; algoritmo *fuzzy* k-médias; árvore geradora mínima; DBSCAN; *single-linkage*; e *complete-linkage.*

O algoritmo k-médias ou *k-means*, de acordo com Castro e Ferrari (2016, p. 116) toma como entrada o parâmetro *k*, correspondente ao número de grupos desejados, e particiona o conjunto de *n* objetos em *k* grupos, de forma que a similaridade intragrupo seja alta e a similaridade intergrupo seja baixa. Os autores afirmam que a similaridade intragrupo é avaliada considerando o valor médio dos objetos em um grupo, que pode ser visto como o seu centro de gravidade ou o centroide.

No particionamento realizado pelo k-médias, os autores explicam que cada objeto pertence ao grupo do centroide mais próximo a ele. Os autores complementam que o algoritmo padrão do k-médias opera por meio de uma técnica de refinamento iterativo da seguinte forma: os *k* centroides iniciais dos grupos são determinados aleatoriamente ou selecionando-se de modo aleatório alguns dos objetos da própria base de dado. Feito isso, os autores afirmam que deve ser calculada a distância entre os objetos da base e cada um dos centroides, atribui-se cada objeto ao centroide mais próximo. Em seguida, os autores explicam que os novos centróides calculados tomando-se a média dos objetos pertencentes a cada centroide, o que

pode promover um reposicionamento dos centroides e uma nova alocação de objeto a grupos. Por fim, os autores destacam que o algoritmo converge quando não há mais alterações nos centroides e mudanças nas alocações de objetos aos grupos.

Para definir o algoritmo k-medoides, Castro e Ferrari (2016, p. 119) definem antes um “medoide” como o objeto com a menor dissimilaridade média a todos os outros objetos, ou seja, é o objeto mais centralmente localizado no grupo. Assim, os autores destacam que o algoritmo k-medoides é um método de agrupamento relacionado ao k-médias, mas que usa um objeto da base como protótipo em lugar de um centroide. Os autores destacam que ambos são algoritmos particionais que visam minimizar o erro quadrático entre os objetos de um grupo e seu protótipo (no caso do k-médias, o centroide do grupo, e do k-medoides, o medoide do grupo). Para os autores, uma diferença importante entre esses métodos é que o k-medoides escolhe objetos da própria base como os centros do grupo, ao passo que o k-médias calcula o centro do grupo a partir dos objetos neles contidos. Além disso, os autores afirmam que o algoritmo k-medoides é mais robusto a ruído e a valores discrepantes do que o k-médias, pois o centro do grupo será necessariamente um objeto da base e, portanto, ruídos e valores discrepantes podem não influenciar tão fortemente a definição do centro.

O método *fuzzy* k-médias é definido por Castro e Ferrari (2016, p. 122) como uma extensão do algoritmo k-médias na qual cada objeto possui um grau de pertinência em relação aos grupos da base. De acordo com os autores, no algoritmo *fuzzy* k-médias um objeto pode pertencer a mais de um grupo, porém, com variados graus de pertinência.

A árvore geradora mínima (*Minimal Spanning Tree* – MST), de acordo com Castro e Ferrari (2016, p. 127), baseia-se em teoria dos grafos, onde deve atender aos requisitos: uma árvore é uma árvore geradora se ela é um subgrafo que contém todos os nós do grafo; uma árvore geradora mínima de um grafo é uma árvore com peso mínimo, onde o peço de uma árvore é definido como a soma dos pesos de suas arestas; e um caminho *minimax* entre um parte de nós é aquele que minimiza o custo (peso máximo do caminho) sobre todos os caminhos. De acordo com os autores, a MST sempre percorre os caminhos *minimax*, forçando a conexão entre dois nós próximos antes de sair em busca de outro nó. Os autores complementam que, em linhas gerais, o método opera da seguinte forma: “construa uma árvore geradora

mínima dos dados de entrada no qual os nós correspondem às coordenadas dos objetos e as arestas à distância (similaridade) entre eles. Em seguida, defina um critério de inconsistência para as arestas, de modo que arestas inconsistentes sejam removidas da árvore; as subárvores (subgrafos) resultantes corresponderão aos grupos de objetos da base". Os autores destacam que, esse método requer que a base de dados esteja representada de forma numérica.

O algoritmo DBSCAN (*Density Based Spatial Clustering of Applications with Noise),* de acordo com Castro e Ferrari (2016, p. 134) foi desenvolvido para encontrar agrupamentos de diferentes formatos e ruídos nas bases de dados, baseando-se na densidade de objetos no espaço. Os autores explicam que a noção de densidade está relacionada à quantidade de objetos dentro de um raio de vizinhança. Ainda de acordo com os autores, o número de grupos é definido automaticamente pelo algoritmo, sendo que cada grupo possui pelo menos um objeto de núcleo. O objeto de núcleo é definido pelos autores como um objeto com uma quantidade mínima de objetos em seu raio de vizinhança, e o grupo é formado agregando os objetos da vizinhança do objeto do núcleo. Assim, os autores destacam que, partindo dos novos objetos adicionais, seus vizinhos também serão agregados ao grupo até que não haja mais objetos na vizinhança. Por fim, os autores afirmam que esse processo é iterativo e repetido até que todos os objetos sejam visitados, sendo que os objetos que não foram adicionados em grupo nenhum são definidos como ruído pelo algoritmo.

O *single-linkage* é definido por Castro e Ferrari (2016, p. 137) como um método aglomerativo hierárquico no qual novos grupos são criados unindo os grupos mais semelhantes. Os autores afirmam que o agrupamento inicial é formado apenas por *singletons (*grupos formados por apenas um objeto), e a cada interação do método um novo grupo é formado por meio da união dos dois grupos mais similares da iteração anterior. Os autores destacam que nesse método a distância (proximidade) entre o novo grupo e os demais é determinada como a menor distância entre os elementos do novo grupo e os grupos remanescentes.

O algoritmo *complete-linkage*, de acordo com Castro e Ferrari (2016, p. 139), opera de maneira similar ao *single-linkage*, mas a distância do novo grupo aos demais é calculada como a distância máxima entre os elementos do novo grupo aos grupos restantes.

2.6.2 Estimação

A estimação corresponde à tarefa de predizer um valor contínuo de uma variável, a qual apresenta aprendizagem do tipo supervisionada e, portanto, requer pares entrada-saída desejada para a construção do estimador e possui muitas características e processos em comum com a classificação. A preparação da base de dados, a separação dos dados em treinamento e teste, a definição dos critérios de parada do algoritmo e o treinamento e teste são feitos de forma equivalente à classificação. Uma diferença importante entre essas tarefas, entretanto, encontra-se na avaliação da saída. No caso dos classificadores, essa avaliação é baseada em algumas medidas de acurácia do classificador, ou seja, a quantidade de objetos classificados corretamente. No caso dos estimadores, a quantidade costuma ser medida calculando-se uma distância ou um erro entre a saída do estimador e a saída desejada (CASTRO; FERRARI, 2016, p. 200).

Castro e Ferrari (2016, p. 200) afirmam ainda que a tarefa de classificação pode ser vista como um caso particular da estimação, no qual a saída é discreta (ou discretizada). Assim, de acordo com os autores, praticamente todos os algoritmos de estimação podem ser usados para classificação, mas a recíproca não é verdadeira.

2.6.2.1 Algoritmos de estimação

Entre os principais métodos de estimação na literatura, Castro e Ferrari (2016) destacam: regressão linear; regressão polinomial; rede neural do tipo *Adaline*; rede neural do tipo *Multi-Layer Perceptron* (MLP); e redes neurais do tipo Função de Base Radial (RBF).

A regressão linear é a relação entre uma ou mais variáveis de resposta (também chamadas de variáveis de saída, dependentes, preditas ou explicadas) e os preditores (também chamados de variáveis de controle, independentes, explanatórias ou regressores.

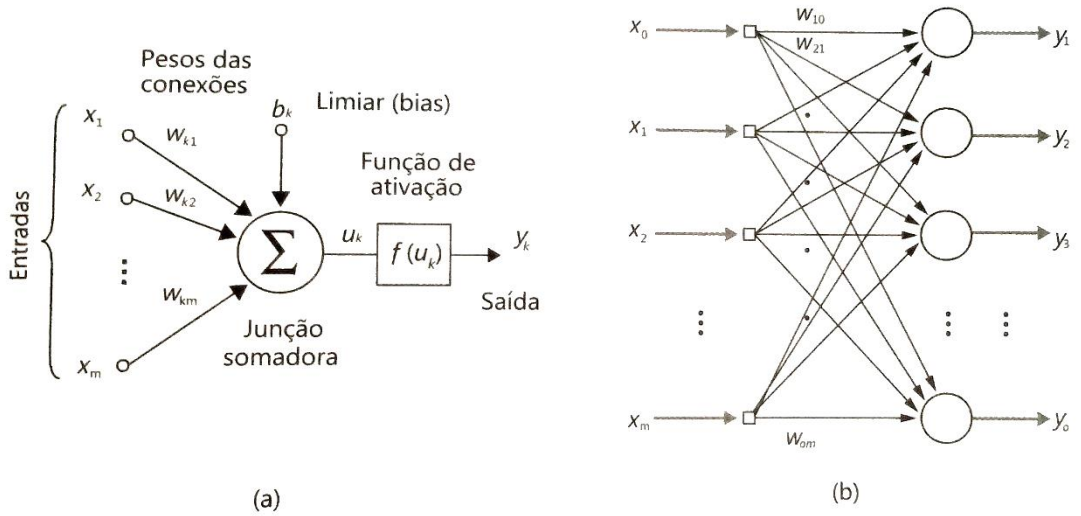
Em linhas gerais, regressão corresponde ao problema de estimar uma função a partir de pares entrada-saída e, se há mais de uma variável de resposta, a regressão linear é chamada de multivalorada. Se a forma do relacionamento funcional entre as variáveis dependentes e independentes é conhecida, mas podem existir parâmetros cujos valores são desconhecidos e podem ser estimados a partir do conjunto de

treinamento, então a regressão é dita paramétrica. Por outro lado, se não há conhecimento prévio sobre a forma da função que está sendo estimada, então a regressão é dita não paramétrica (CASTRO; FERRARI, 2016, p. 205).

A regressão polinomial é definida por Castro e Ferrari (2016, p. 208) como um modelo de regressão no qual a relação entre as variáveis independentes e a variável dependente pode ser não linear e tem a forma de um polinômio de grau n.

Rede neural do tipo *Adaline* (*Adaptive Linear Element*) é similar ao *Perceptron*, ou seja, uma arquitetura mais simples de rede neural capaz de classificar padrões linearmente separáveis. O *Perceptron* consiste em uma rede neural com uma única cama de pesos, ou seja, um conjunto de neurônios de entrada e um conjunto de neurônios de saída, com pesos sinápticos e bias ajustáveis. Se os padrões de entrada forem linearmente separáveis, o algoritmo de treinamento do *Perceptron* possui convergência garantida, ou seja, é capaz de encontrar um conjunto de pesos que classifica corretamente os dados. Os pesos dos neurônios que compõem o *Perceptron* serão tais que as superfícies de decisão produzidas pela rede neural estarão apropriadamente posicionadas no espaço. Entretanto, a rede *Adaline* se diferencia por apresentar neurônios usando função de ativação linear em vez de função de sinal, conforme demonstra a Figura 17 (CASTRO; FERRARI, 2016, p. 209).

Figura 17 - Neurônio (a) e a rede neural do tipo *Perceptron* e *Adaline* (b)



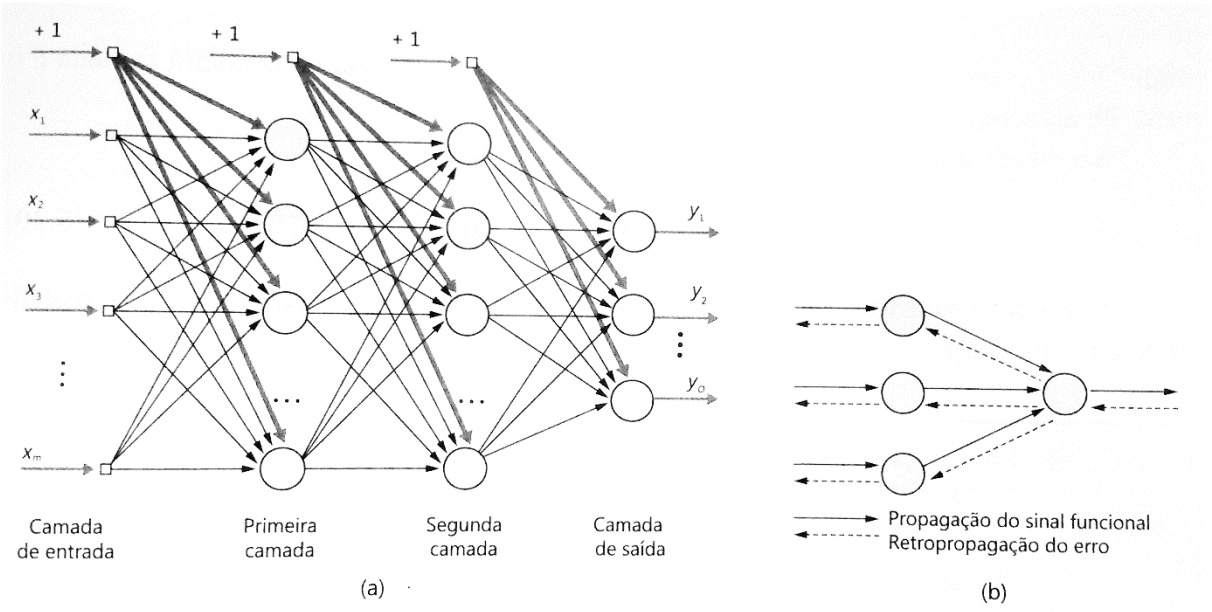
FONTE: Castro e Ferrari (2016, p. 209)

Rede neural do tipo Multi-Layer *Perceptron* (MLP) ou *Perceptron* de múltiplas camadas é uma rede do tipo *Perceptron* com pelo menos uma camada intermediária.

Trata-se de uma generalização do *Perceptron* simples e da rede *Adaline*. O treinamento da rede MPL foi feito originalmente utilizando-se um algoritmo denominado retropropagação do erro, conhecido como backpropagation. Esse algoritmo consiste basicamente em dois passos: propagação positiva do sinal funcional, durante a qual todos os pesos da rede são mantidos fixos; e retropropagação do erro, durante a qual os pesos da rede são ajustados com base no erro. (CASTRO; FERRARI, 2016, p. 215). Ainda de acordo com os autores, uma rede MPL típica possui três características principais, conforme demonstra a Figura 18.

- os neurônios das camadas intermediárias (e de saída) possuem uma função de ativação não linear e do tipo sigmoidal;
- a rede possui uma ou mais camadas intermediárias; e
- a rede possui altos graus de conectividade.

Figura 18 - Rede neural de múltiplas camadas (a), sentido de propagação do sinal de entrada e retropropagação do erro (b)



FONTE: Castro e Ferrari (2016, p. 215)

Nas redes neurais do tipo Função de Base Radial (RBF, do inglês *Radial Basis Function*), o projeto das redes com múltiplas camadas a propagação positiva do sinal é visto como um problema de aproximação de função em um espaço multidimensional. A camada de entrada é responsável por conectar a rede a seu ambiente, a camada intermediária aplica uma transformação não linear do espaço de

entrada para o espaço intermediário e a camada de saída é linear (CASTRO; FERRARI, 2016, p. 222).

2.6.3 Classificação

A classificação como tarefa de organizar objetos em uma entre diversas categorias pré-definidas é definida por Tan, Steinbach e Kumar (2009, p. 171) como um problema universal que engloba muitas aplicações diferentes. De acordo com os autores, os exemplos incluem a detecção de mensagens de *spam* em e-mails baseada no cabeçalho e conteúdo da mensagem, a categorização de células como malignas ou benignas baseada nos resultados da varredura MRI (*Magnetic Resonance Imaging* ou ressonância magnética, em português) e a classificação de galáxias baseada nos seus formatos.

De acordo com os autores, classificação é definida como: “a tarefa de aprender uma função alvo *f* que mapeie cada conjunto de atributos x para um dos rótulos de classes *y* pré-determinados”, conforme demonstra a Figura 19.

Figura 19 - Classificação como a tarefa de mapear um conjunto de atributos x no seu rótulo de classe y

FONTE: Tan, Steinbach e Kumar (2009, p. 172).

Camilo e Silva (2009) complementam que a classificação, visa identificar a qual classe um determinado registro pertence. De acordo com os autores, nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de “aprender” como classificar um novo registro (aprendizado supervisionado). Por exemplo, categorizamos cada registro de um conjunto de dados contendo as informações sobre os colaboradores de uma empresa: Perfil Técnico, Perfil Negocial e Perfil Gerencial. O modelo analisa os registros e então é capaz de dizer em qual categoria um novo colaborador se encaixa. A tarefa de classificação pode ser usada, por exemplo, para:

- determinar quando uma transação de cartão de crédito pode ser uma fraude;
- identificar em uma escola, qual a turma mais indicada para um determinado aluno;
- diagnosticar onde uma determinada doença pode estar presente;
- identificar quando uma pessoa pode ser uma ameaça para a segurança.

De acordo com Castro e Ferrari (2016, p. 165), classificar um objeto significa atribuir a ele um rótulo, chamado *classe*, de acordo com a categoria à que ele pertence. De acordo com os autores, para que isso seja possível, um algoritmo de classificação é usado na construção de um modelo de classificação, também chamado *classificador*, o qual é construído com base em um conjunto de treinamento com dados rotulados. Os autores destacam que há uma grande variedade de algoritmos de classificação na literatura, sendo possível separá-los de acordo com a sua estrutura em: baseados em conhecimento; baseados em árvore; conexionistas; baseados em distância; baseados em função; e probabilísticos.

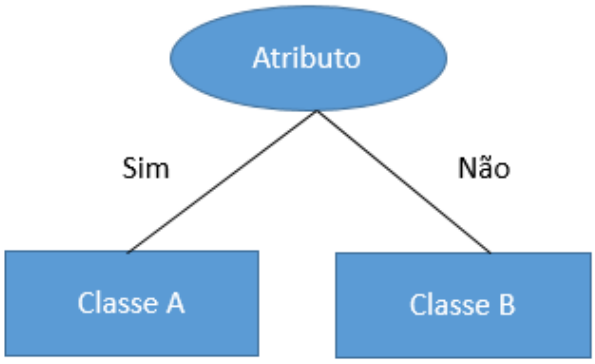
De acordo com os autores, o modelo baseado em conhecimento opera por meio de um conjunto de regras usadas para atribuir determinada classe a um objeto caso ele satisfaça condições predefinidas (Figura 20).

Figura 20 - Modelo baseado em conhecimento

FONTE: Castro e Ferrari (2016, p. 165)

Nos modelos baseados em árvores, os autores afirmam que o nó raiz e os nós intermediários das árvores representam testes sobre um atributo, os ramos representam os resultados desses testes e os nós folhas, os rótulos de classe (Figura 21).

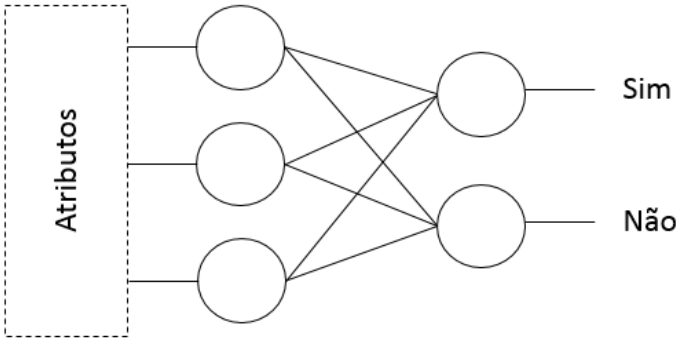
Figura 21 - Modelo baseado em árvores



FONTE: Adaptado de Castro e Ferrari (2016, p. 166)

Os modelos conexionistas são definidos pelos autores como aqueles baseados em redes de unidades (nós) interconectadas. Os autores afirmam que os sistemas conexionistas são um tipo de grafo e, embora haja diferentes sistemas conexionistas, os mais comuns são as Redes Neurais Artificiais (Figura 22).

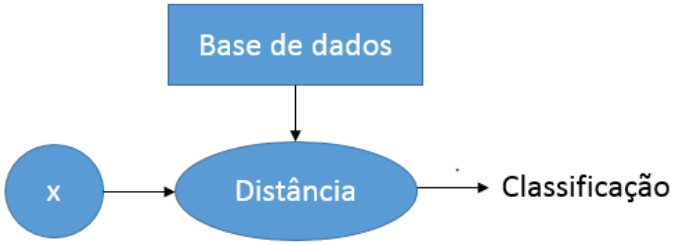
Figura 22 - Modelo conexionista



FONTE: Castro e Ferrari (2016, p. 166)

Nos modelos baseados em distância os autores afirmam que o processo de classificação se dá calculando a distância entre o objeto cuja classe se deseja conhecer e um ou mais objetos rotulados. Os autores afirmam que a classe do objeto desconhecido passa a ser a mesma daqueles objetos que estão a uma menor distância dele (Figura 23).

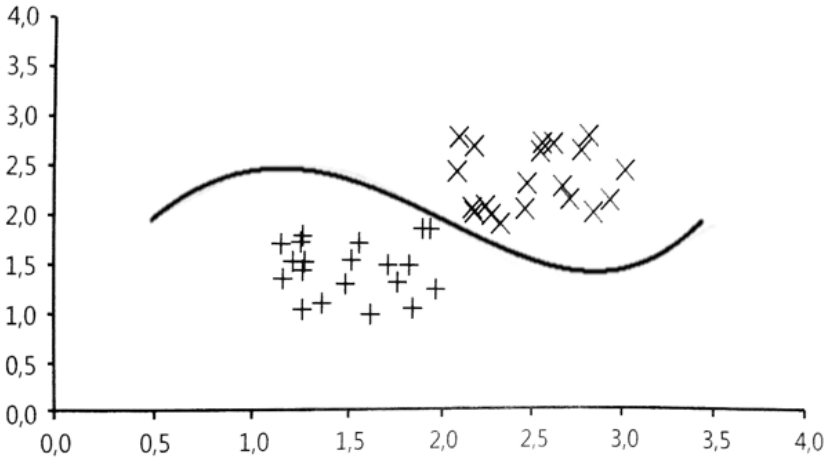
Figura 23 - Modelo baseado em distância



FONTE: Castro e Ferrari (2016, 166)

Os modelos baseados em função são, segundo os autores, paramétricos baseados em funções predefinidas e cujos parâmetros são ajustados durante o processo de treinamento. Os autores afirmam que após o treinamento, um novo objeto de classe desconhecida é apresentado à função, cujo valor é calculado e que representa, se alguma forma, a classe desse objeto (Figura 24).

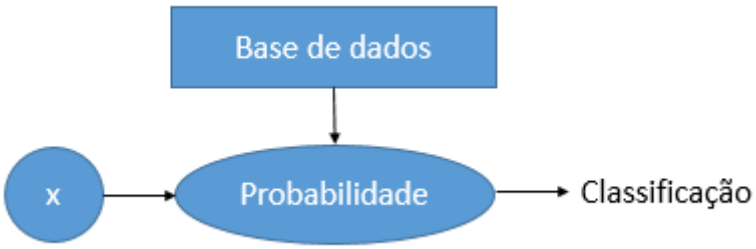
Figura 24 - Modelo baseado em função



FONTE: Castro e Ferrari (2016, p. 167)

Por fim, os autores afirmam que o modelo probabilístico permite atribuir uma probabilidade de um objeto pertencer a uma ou mais classes possíveis, conforme demonstra a Figura 25.

Figura 25 - Modelo probabilístico



FONTE: Castro e Ferrari (2016, p. 167)

Com base nas categorias apresentadas por Castro e Ferrari (2016), serão abordados os principais métodos de classificação presentes na literatura.

2.6.3.1 Algoritmos de classificação

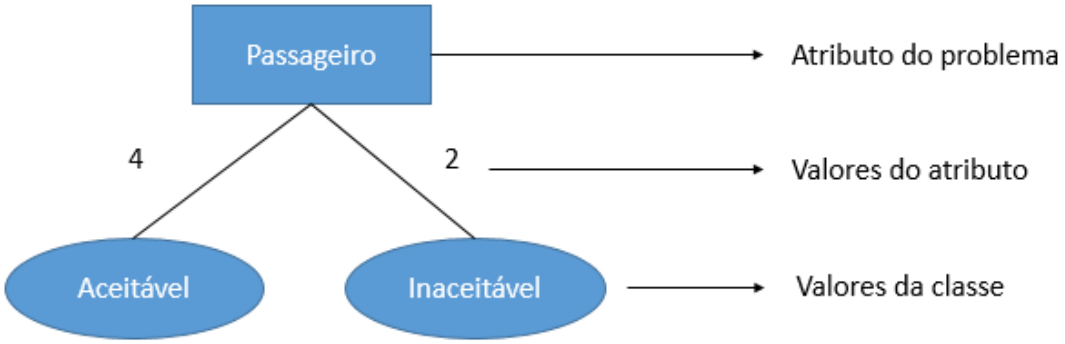
Entre os principais métodos de classificação na literatura, Castro e Ferrari (2016) destacam: classificador k-NN; árvores de decisão; regras de classificação; classificador one-rule (1R); e classificador *naive* Bayes.

O classificador k-NN (k-*nearest neighbors*) corresponde ao método dos k-vizinhos mais próximos. Ele opera da seguinte maneira: dado um objeto $\mathbf{x_0}$ cuja classe se deseja inferir, encontra-se os k objetos $\mathbf{x_i}$, $i = 1, \ldots, k$ da base que estejam mais próximos a $\mathbf{x_0}$ e, depois, se classifica o objeto $\mathbf{x_0}$ como pertencente à classe da maioria dos k vizinhos mais próximos. (CASTRO; FERRARI, 2016, p. 167).

Uma árvore de decisão (*decision tree*) é uma estrutura em forma de árvore na qual cada nó interno corresponde a um teste de um atributo, cada ramo representa um resultado do teste e os nós folhas representam classes ou distribuições de classes. O nó mais elevado da árvore é conhecido como nó raiz e cada caminho da raiz até um nó folha corresponde a uma regra de classificação, conforme demonstra a Figura 26 (CASTRO; FERRARI, 2016, p. 170).

De acordo com Camilo e Silva (2009), “o sucesso das árvores de decisão, deve-se ao fato de ser uma técnica extremamente simples, não necessita de parâmetros de configuração e geralmente tem um bom grau de assertividade”.

Figura 26 - Exemplo de árvore de decisão



FONTE: Castro e Ferrari (2016, p. 170)

As regras de classificação são, de acordo com Tan, Steinbach e Kumar (2009, p. 245), uma técnica para classificar registro usando um conjunto de regras “se ... então”. Castro e Ferrari (2016, p. 180) complementam que elas constituem uma alternativa popular às arvores de decisão. Ainda de acordo com os autores, o antecedente de uma regra é uma série de testes similares àqueles feitos nos nós da árvore de decisão e o consequente da regra fornece a classe ou as classes (ou a distribuição de probabilidades sobre as classes) aplicáveis aos objetos cobertos por aquela regra. Os autores explicam que é fácil ler um conjunto de regras diretamente de uma árvore de decisão: uma regra é gerada para cada nó folha da árvore; o antecedente da regra inclui uma condição para cada nó do caminho da raiz à folha; e o consequente da regra é a classe especificada pela folha.

O classificador *one-rule* (1R) é definido por Castro e Ferrari (2016, p. 193) como uma “forma fácil de encontrar regras de classificação que testam um único atributo da base de dados. Os autores afirmam que além de simples, o algoritmo IR tem baixo custo computacional e muitas vezes é capaz de descobrir boas regras que caracterizam a estrutura dos dados. Ainda de acordo com os autores, frequentemente regras simples são capazes de fornecer altos valores de acurácia, talvez porque a estrutura intrínseca de muitas bases dedados seja rudimentar e pelo fato de que um único atributo é suficiente para determinar a classe de um objeto com uma acurácia”. De acordo com os autores, a ideia geral do algoritmo é a seguinte:

- são construídas regras que testam um único atributo, ramificando-o, sendo que cada ramo corresponde a diferentes valores do atributo;

- a melhor classificação de cada ramo é aquela que usa a classe que ocorre com mais frequência nos dados de treinamento;
- assim, a taxa de erro das regras pode ser facilmente determinada por maio da contagem do número de erros que ocorre para os dados de treinamento, ou seja, do número de objetos que mão possuem a maioria nas classes.

Os classificadores bayesianos, por sua vez, são classificadores estatísticos fundamentados no teorema de Bayes e usados para predizer a probabilidade de pertinência de um objeto a determinada classe. Estudos indicam que os algoritmos simples de classificação bayesiana, conhecidos como *naive Bayes*, possuem desempenho comparável a redes neurais artificiais e árvores de decisão para alguns problemas. Eles também apresentam alta acurácia e velocidade de processamento quando aplicados a grandes bases de dados. Os classificadores *naive Bayes* assumem que o efeito do valor de um atributo em uma dada classe é independente dos valores dos outros atributos. Essa premissa denominada independência condicional da classe, tem como objetivo simplificar os cálculos e, por causa dela, o algoritmo é denominado *naive* (CASTRO; FERRARI, 2016, p. 186).

2.6.4 Regras de Associação

De acordo com Castro e Ferrari (2016, p. 235), a mineração de regras de associação é uma técnica usada na construção de relações sob a forma de regras entre itens de uma base de dados transacional. Os autores explicam que diferentemente do agrupamento, que busca relações de similaridade entre objetos, as regras de associação buscam relações entre os atributos dos objetos, ou seja, os itens que compõem a base. Assim, os autores afirmam que o objetivo é encontrar regras fortes de acordo com alguma medida do grau de interesse da regra.

Ainda de acordo com os autores, as regras de associação não são diferentes das regras de classificação, exceto pelo fato de que elas podem ser usadas para predizer qualquer atributo, não apenas a classe. Os autores afirmam que essa característica que dá a liberdade de predizer combinações de atributos também. Além disso, os autores complementam que as regras de associação não são planejadas para serem usadas em conjunto, como no caso das regras de classificação. Ainda de

acordo com os autores, diferentes regras de associação expressam diferentes regularidades da base de dados e geralmente são usadas para estimar relações distintas entre itens.

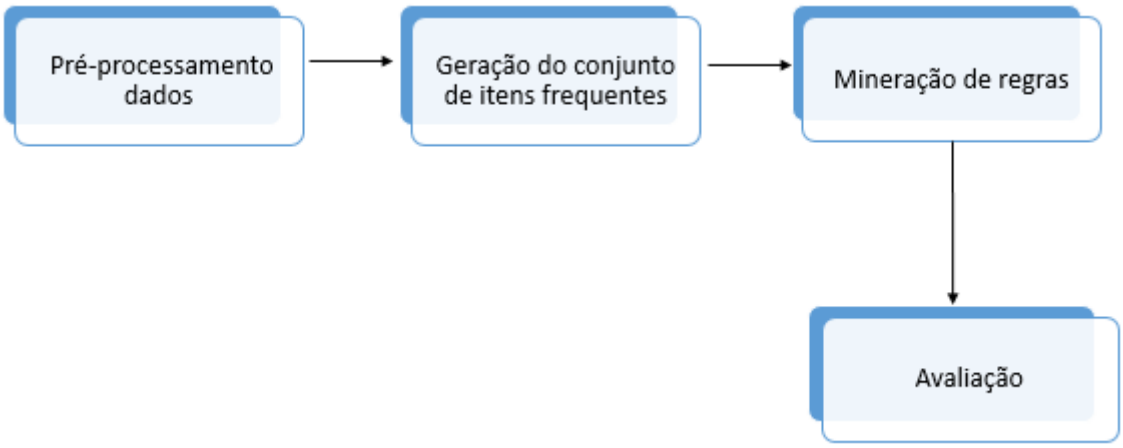
Os autores complementam que como uma grande quantidade de regras de associação pode ser derivada a partir de uma base de dados, mesmo que pequena, normalmente se objetiva a derivação de regra que suportam um grande número de transações e que possuam uma confiança razoável para as transações as quais elas são aplicáveis. Os autores afirmam que esses requisitos estão associados a dois conceitos centrais em mineração de regras de associação:

- suporte: o suporte, ou cobertura de uma regra de associação é o número de transações para as quais ela faz a predição correta. Também pode ser entendida como uma utilidade de urna dada regra;
- confiança: a confiança ou acuraria, de uma regra é o número de transações que ela prediz corretamente proporcionalmente às transações para as quais ela se aplica. Também pode ser entendida como a certeza de uma dada regra.

Os conceitos de suporte e confiança permitem definir o problema geral de mineração de regras de associação: “Dado um conjunto de transações, o problema de minerar regras de associação corresponde a encontrar todas as regras que satisfaçam um valor mínimo predefinido de suporte (chamado de *minsup)* e um valor predefinido de confiança (chamado de *minconf)*” (CASTRO; FERRARI, 2016, p. 236).

Castro e Ferrari (2016, p. 240) dividem o processo geral de minerar regras de associação em quatro passos: pré-processamento da base; geração do conjunto de itens frequentes; mineração e regras; e avaliação, conforme demonstra a Figura 27.

Figura 27 - Processo de mineração de regras de associação



FONTE: Castro e Ferrari (2016, p. 240)

De acordo com os autores, o pré-processamento da base, além de poder envolver todas as etapas típicas de pré-processamento de dados, como limpeza, integra e discretização, a preparação de uma base transacional para a mineração de regras pode exigir que essa base seja transformada em urna base binária ou baseada em frequência (e outras informações relevantes), na qual cada elemento da tabela corresponde a quantas unidades de um item há em cada transação.

Segundo os autores, a geração do conjunto de itens frequentes é realizada pela adoção de algum critério mínimo de frequência, por exemplo, itens que aparecem ao menos em determinado número de transações. Os autores afirmam que essa etapa do processo existe para que a construção das regras de associação seja feita de maneira mais parcimoniosa, uma vez que a quantidade de possíveis regras a serem mineradas de uma base cresce exponencialmente com o número de itens.

Na mineração das regras, os autores afirmam que as regras de associação propriamente ditas são geradas em uma etapa específica usando apenas os itens frequentes da base. Segundo os autores, uma forma direta de gerar as regras de associação seria fazer todas as combinações possíveis dos itens frequentes, mas essa abordagem resulta em um problema combinatório que se torna computacionalmente intratável para bases de tamanho moderado a grande.

Na avaliação os autores destacam que as regras de associação podem ser avaliadas utilizando-se deferentes medidas de interesse, dependendo do contexto. Os autores afirmam que praticamente todas elas utilizam o suporte e a confiança das regras para quantificar sua utilidade.

De acordo com Tan, Steinbach e Kumar (2009, p. 390), a análise de associações é aplicável para domínios de aplicação como bioinformática, diagnósticos médicos, mineração na Web e análise de dados científicos. Os autores destacam que na análise dos dados da ciência da Terra, por exemplo, os padrões de associação podem revelar conexões interessantes entre o oceano, a terra e os processos atmosféricos. Os autores afirmam que tais informações podem ajudam cientistas da Terra a desenvolver uma melhor compreensão de como os diferentes elementos do sistema da Terra interagem entre si.

2.6.4.1 Algoritmos de regras de associação

Entre os principais algoritmos de descoberta de regras de associação encontrados na literatura, Castro e Ferrari (2016) destacam: algoritmo Apriori e algoritmo FP-*Growth.*

O algoritmo Apriori é o método mais conhecido para a mineração de regras de associação e emprega busca em profundidade e gera conjuntos de itens candidatos de *k* elementos a partir de conjuntos de itens com *k – 1* elementos. Os itens candidatos não frequentes são eliminados, e toda a base de dados é rastreada e os conjuntos de itens frequentes são obtidos a partir dos conjuntos de itens candidatos. A estratégia adotada pelo algoritmo consiste em decompor o problema em duas sub tarefas: geração do conjunto de itens frequentes e geração das regras. A geração do conjunto de itens frequentes consiste em encontrar todos os conjuntos cujo suporte seja maior que o *minsup* especificado. Já a geração de regras corresponde ao uso dos conjuntos de itens frequentes para gerar as regras desejadas. A ideia geral é que se, por exemplo, ABCD e AB são frequentes, então é possível determinar se a regra AB → CD é válida calculando a razão *confiança = suporte(ABCD) / suporte(AB).* Se a confiança for maior ou igual a *minconf*, então a regra é válida (CASTRO; FERRARI, 2016, p. 246).

O algoritmo FP-*Growth* (*Frequente Pattern Growth),* de acordo com Castro e Ferrari (2016, p. 252) é baseado em uma estrutura em árvore de prefixos para os padrões frequentes, denominada FP-*Tree* (*Frequent Pattern Tree),* a qual armazena de forma comprimida a informação sobre os padrões frequentes. Ainda de acordo com os autores, o algoritmo FP-*Growth* extrai o conjunto completo de padrões frequentes.

Os autores destacam que a essência do algoritmo está baseada em três aspectos centrais:

- a compreensão da base de dados sem uma estrutura em árvore (FP-*Tree*) cujos nós possuem apenas itens frequentes de comprimento unitário (F_1) e organizada de modo que aquelas nós que ocorrem mais frequentemente tenham maiores chances de compartilhar nós do que os de baixa frequência;
- o uso de um algoritmo de mineração da árvore que evita a geração de uma grande quantidade de conjuntos candidatos. Esse algoritmo inicia com um padrão frequente de comprimento 1 (padrão sufixo inicial) avalia apenas o conjunto de itens frequentes que co-ocorrem com o padrão sufixo, constrói sua FP-Tree e executa uma mineração recursiva na árvore;
- o uso de um método particional para decompor a tarefa de mineração em subtarefas menores, reduzindo significativamente o espaço de busca.

2.6.5 Detecção de Anomalias

Uma anomalia é definida por Castro e Ferrari (2016, p. 269) como um valor discrepante, ou seja, um valor que se localiza significativamente distante dos valores considerados normais. Os autores destacam que é importante notar que uma anomalia não é necessariamente um erro ou um ruído, podendo caracterizar um valor ou uma classe bem definida, porém de baixa ocorrência, às vezes indesejada, ou que reside fora de agrupamentos ou classes típicas.

Na detecção de anomalias Tan, Steinbach e Kumar (2009, p. 777) destacam que o objetivo é encontrar objetos que sejam diferentes da maioria dos outros objetos. De acordo com os autores, a detecção de anomalias também é conhecida como detecção de desvios, considerando que os objetos anômalos têm atributos que se desviam significativamente dos valores de atributos esperados ou típicos ou, como mineração de exceções, considerando que as anomalias são excepcionais em algum sentido. Castro e Ferrari (2016, p. 269) complementam que a importância da detecção de anomalias deve-se ao fato de que elas normalmente correspondem a dados significativos, às vezes críticos, para a análise. Os autores exemplificam que uma fraude em uma transação de cartão de crédito, um intruso em uma rede de

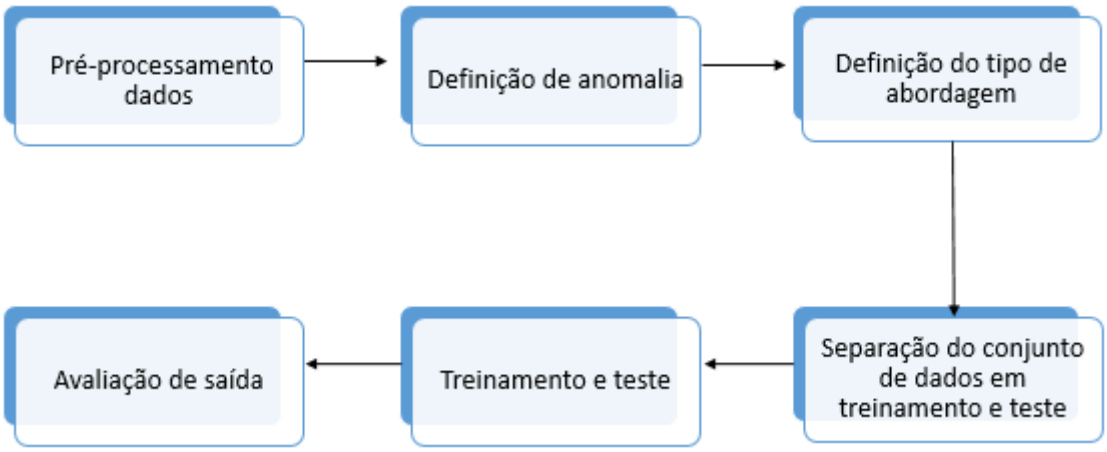
computadores ou uma folha na operação de uma turbina de avião são anomalias que causam algum tipo de prejuízo, risco ou dano ao sistema.

Castro e Ferrari (2016, p. 270) apontam que as principais aplicações de detecção de anomalias incluem:

- detecção de fraudes: em transações de cartões de crédito, em uso de telefones celulares, em medição de consumo de energia etc.;
- análise de crédito: identificação de clientes potencialmente problemáticos, inadimplentes ou fraudulentos;
- detecção de intrusão: identificação de acesso não permitido a redes de computadores e ambientes diversos;
- monitoramento de atividades: acompanhamento e identificação de negociações suspeitas em mercados financeiros, comportamentos incomuns de usuários, desempenho de atletas, operações de sistemas, etc.;
- desempenho de rede: monitoramento do desempenho de redes de comunicação para a identificação de gargalos;
- diagnóstico de faltas: em motores, geradores, redes, instrumentos, etc.;
- análise de imagens de vídeos: identificação de novas características, objetos, comportamento etc.;
- monitoramento de séries temporais: em aplicações que envolvem séries temporais, por exemplo, consumo de energia elétrica de subestações, análise de batimentos cardíacos etc.;
- análise de textos: identificação de novas histórias, riscos, situações etc.

A detecção de anomalias em bases de dados é identificada por Castro e Ferrari (2016, p. 273) como essencialmente um problema de classificação binária, no qual se deseja determinar se um ou mais objetos pertencem à classe normal ou à classe anômala. Assim, os autores afirmam que esse processo é muito similar ao fluxo convencional da tarefa de predição, contanto com dois passos adicionais: definição do que é uma anomalia e definição do tipo de abordagem, conforme demonstra a Figura 28.

Figura 28 - Fluxo do processo de classificação de anomalias



FONTE: Castro e Ferrari (2016, p. 273)

De acordo com os autores, a definição do que é uma anomalia corresponde a definição de algum contorno ou vizinhança ao redor de uma das classes (normal ou anomalia) e, a partir deste, estabelece um liminar de normalidade ou anomalia. Já a definição do tipo de abordagem envolve a definição se a abordagem será supervisionada ou não supervisionada.

2.6.5.1 Algoritmos de detecção de anomalias

Entre os principais métodos de detecção de anomalias presentes na literatura, Castro e Ferrari (2016) destacam: métodos estatísticos paramétricos e não paramétricos; e métodos algorítmicos baseados em proximidades, redes neurais artificiais e em aprendizagem de máquina.

Castro e Ferrari (2016, p. 277) os métodos estatísticos para detecção de anomalias normalmente geram um modelo probabilístico de dados e testam se determinado objeto foi gerado por tal modelo ou não. Assim, os autores afirmam que essas técnicas são essencialmente baseadas em modelo, ou seja, assumem ou estimam um modelo estatístico que captura a distribuição dos objetos da base e avalia os objetos em relação a quão bem eles se ajustam ao modelo. Os autores explicam que se a probabilidade de certo objeto ter sido gerado por esse modelo for muito baixa, então ele é rotulado como uma anomalia. Os autores destacam que os métodos estatísticos foram os primeiros a ser utilizados para detecção de anomalias, e muitos deles operam com dados unidimensionais

Os métodos paramétricos assumem que os dados gerados por uma distribuição conhecida e, na maioria das vezes, ajustam um modelo específico aos dados; portanto, a fase de treinamento envolve estimar os parâmetros do modelo para uma base de dados. Os métodos paramétricos permitem que o modelo seja avaliado rapidamente para novos objetos e são adequados a grandes bases de dados, pois o modelo cresce somente com a complexidade do próprio modelo, e não com o tamanho da base de dados. No entanto, os autores afirmam que a sua aplicação é limitada, pois eles assumem uma distribuição preespecificada dos dados (CASTRO; FERRARI, 2016, p. 277).

Os métodos não paramétricos são definidos por Castro e Ferrari (2016, p. 285) como aqueles que não assumem uma distribuição predefinida dos dados nem um modelo específico que deverá ser ajustado aos dados. Os autores afirmam que as abordagens mais populares nessa categoria são as baseadas em histogramas, geralmente usadas de forma não supervisionada.

Os métodos baseados em proximidades são, de acordo com Castro e Ferrari (2016, p. 288) normalmente simples de implementar e não arrumem nenhuma premissa sobre a distribuição dos objetos da base. Os autores afirmam que podem ser aplicados tanto de forma não supervisionada quanto supervisionada, e o princípio básico da operação desses métodos é o cálculo de alguma medida de similaridade ou distância entre pares de objetos da base. Assim, os autores explicam que os objetos são mapeados em um espaço métrico definido sobre um conjunto finito de atributos.

Os métodos baseados em redes neurais artificiais são discutidos por Castro e Ferrari (2016, p. 294) em relação a sua utilização para a detecção de anomalias. Os autores destacam que as redes neurais competitivas fazem parte do paradigma de aprendizagem não supervisionada, sendo empregadas geralmente para explorar dados não rotulados. Portanto, os autores afirmam que os seus algoritmos de treinamento utilizam apenas os dados de entrada, tomados como amostras independentes de uma distribuição de probabilidade desconhecida. Assim, os autores afirmam que as redes neurais devem desenvolver uma capacidade de criar representações internas que codificam as características dos dados de entrada, tornando-se capaz de identificar a quais classes (grupos) novos objetos pertencem.

A aprendizagem de máquina é definida por Castro e Ferrari (2016, p. 296) como uma área de pesquisa que visa desenvolver programas computacionais capazes de automaticamente melhorar sem desempenho pela experiência. Os autores destacam

que alguns dos métodos utilizados com enfoque para a detecção de anomalias são: algoritmos de classificação, algoritmos de agrupamento e regras de classificação.

Os algoritmos de classificação, como árvores de decisão e as regras de classificação, são empregados de acordo com Castro e Ferrari (2016, p. 297) como uma abordagem supervisionada. Nesse caso, os autores afirmam que se assume que a informação sobre as duas classes (normal e anomalia) para treinar o classificador e, feito o treinamento, o objeto de teste é avaliado para conferência de em qual classe ele é categorizado. Os autores afirmam que o processo de aplicação do algoritmo para detecção de anomalia é padrão: treinamento do classificador e teste dos novos objetos.

Castro e Ferrari (2016, p. 291) destacam que as técnicas de agrupamento, bem como o k-médias e o k-medoides podem ser aplicados de forma supervisionada para a detecção de anomalias, considerando critérios baseados em proximidade.

Ainda de acordo com os autores, os algoritmos de mineração de regras de associação normalmente geram regras que satisfazem algum critério de supor mínimo (*minsup*), ou seja, regras que aparecem com uma frequência mínima na base. Os autores afirmam que como esses algoritmos buscam associações entre os atributos da base, nenhuma informação de classe é necessária e, portanto, eles podem ser aplicados de maneira não supervisionada, assumindo que as anomalias ocorrem com baixa frequência na base, é possível definir um suporte mínimo tal que o algoritmo de mineração de regras de associação gere regras que desconsiderem as anomalias.

2.6.6 Escolha do Método de Mineração de Dados

A escolha do método de mineração de dados é um processo complexo, pois de acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), apesar de cada um possuir suas peculiaridades e apresentar melhor resultado com um certo tipo de dado, não existe uma classificação única para a escolha e aplicação destes métodos.

Chen, Han e Yu (1996, p. 4) destacam que diferentes esquemas de classificação podem ser usados para categorizar métodos de mineração de dados sobre os tipos de bancos de dados a serem estudados, os tipos de conhecimento a serem descobertos e os tipos de técnicas a serem utilizadas, conforme mostrado a seguir:

- que tipos de bancos de dados trabalhar: um sistema de descoberta de conhecimento pode ser classificado de acordo com os tipos de bancos de dados sobre os quais técnicas de mineração de dados são aplicadas, tais como: bancos de dados relacionais, bancos de dados de transação, orientados a objetos, dedutivos, espaciais, temporais, de multimídia, heterogêneos, ativos, de herança, banco de informação de Internet e bases textuais;
- qual o tipo de conhecimento a ser explorado: vários tipos de conhecimento podem ser descobertos por extração de dados, incluindo regras de associação, regras características, regras de classificação, regras discriminantes, agrupamento, evolução e análise de desvio;
- qual tipo de técnica a ser utilizada: a extração de dados pode ser categorizada de acordo com as técnicas de mineração de dados subordinadas. Por exemplo, extração dirigida a dados, extração dirigida a questionamento e extração de dados interativa. Pode ser categorizada, também, de acordo com a abordagem de mineração de dados subordinada, tal como: extração de dados baseada em generalização, baseada em padrões, baseada em teorias estatísticas ou matemáticas, abordagens integradas etc.

Castro e Ferrari (2016, p. 11) apresentam uma lista de considerações que podem servir como guia para uma mineração eficiente e eficaz:

- estabelecer a significância da mineração: tanto a significância estatística quanto a prática da mineração devem ser consideradas. A significância estatística está relacionada à confiabilidade dos resultados obtidos, ou seja, se a base de dados foi preparada adequadamente para a análise, se os resultados apresentados são coerentes e se os algoritmos propostos têm o desempenho desejado. Por exemplo, uma amostragem ou normalização inadequada da base pode gerar resultados que não tenham nenhuma significância estatística e que, portanto, são inúteis. A significância pratica, por sua vez, questiona sobre a aplicabilidade práticas das análises realizadas, ou seja, se essas análises podem ser usadas em algum processo de tomada de decisão;

- reconhecer que as características da base de dados influenciam todos os resultados: o processo de mineração opera, quase em sua totalidade, sobre uma base de dados pré-processada. Assim, é importante reconhecer que a quantidade de objetos da base, a dimensão (número de atributos) desses objetos, o tipo de atributos e seus domínios, a ausência de valores na base, as inter-relações entre os atributos e muitas outras características dos dados afetarão fortemente o resultado da análise, podendo, inclusive, invalidá-la;
- necessidade de reconhecer os dados: a discussão apresentada implica que análises preliminares dos dados – mais especificamente as técnicas de análise descrita, como medidas de tendência central (por variável), análise de componentes principais e muitos outros métodos (estatísticos) simples – devem ser aplicados à base com o objetivo de entende-la melhor antes de se iniciar a mineração propriamente dita;
- buscar pela parcimônia: boa parte dos algoritmos de mineração resultam em uma espécie de modelos dos dados que poderá ser utilizado posteriormente para fazer alguma inferência ou predição. É possível que a escolha de diferentes amostras dos dados, ou mesmo diferentes execuções dos algoritmos, resultem em modelos com características distintas. Nesses casos, a escolha por um ou outro modelo deve considerar, entre outros aspectos, a parcimônia da solução, ou seja, a complexidade do modelo resultante. Muitas vezes, a complexidade de criação do modelo é um aspecto crucial na escolha de uma ferramenta dentro de um conjunto de possiblidades;
- verificar os erros: todos os algoritmos de mineração podem ter seu desempenho avaliado. Nos casos dos algoritmos de agrupamento, há medidas que permitem avaliar a quantidade dos agrupamentos propostos; nas tarefas de predição, é possível avaliar que o erro de predicação; na mineração de regras de associação, avalia-se a significância das regras; e, para os algoritmos de detecção de anomalias, verifica-se o seu desempenho por meio de medidas específicas para esse tipo de problema. Em todos os casos, é preciso fazer um diagnóstico de desempenho do algoritmo, identificando os erros, o porquê de sua ocorrência, e empregar esse conhecimento para realimentar o processo de análise;

- validar os resultados: os resultados de uma análise precisam ser validados de diversas formas, por exemplo, comparando com resultados de outras técnicas, analisando a capacidade de generalização dos métodos, combinando com outras técnicas e até utilizando um especialista de domínio capaz de validar se os resultados apresentados fazem sentido e se são se boa qualidade. Assim como no caso anterior, a validação é central para realimentar o processo de análise de dados.

2.6.7 Softwares para Mineração de Dados

Castro e Ferrari (2016, p. 348) destacam os Softwares mais utilizados para dar suporte às tarefas de mineração de dados: Weka; Matlab, R, Wolfram Mathematica; RapidMiner; SAS; SSPS; Orange; Mahout; Elki; e Libsvm, conforme demonstra o Quadro 4.

Quadro 4 - Descrição dos principais Softwares de mineração de dados

(continua)

Software	Descrição	Licença
Weka	O Weka é um software de código aberto, desenvolvido em Java e mantido pela Universidade de Waikato. Possui interface gráfica que permite ao usuário realizar tarefas de pré-processamento, classificação, regressão, agrupamento e visualização dos dados, assim como planejar e executar análises ou experimentos mais complexos por meio da construção de fluxogramas que encadeiam as tarefas de mineração de dados. Há também a possibilidade de utilização do Weka por meio da integração de suas bibliotecas em um ambiente de desenvolvimento Java, como NetBeans ou Eclipse. Essa integração dá maior poder de personalização à mineração dos dados, mas exige maior conhecimento da ferramenta e da linguagem Java.	Gratuita
Matlab	O Matlab é uma linguagem de programação de alto nível aliada. Originalmente desenvolvido para atuar em operações matriciais e álgebra linear na década de 1970, o software ficou famoso por suas aplicações nas mais diversas áreas de engenharia. Atualmente, conta com pacotes para diferentes processos de mineração de dados, tais como agrupamento, classificação e estimação, com ferramentas específicas para dados financeiros e biológicos, processamento de sinais e imagens, entre outros. Em virtude da natureza matricial da linguagem, o pacote de redes neurais artificias é bastante completo, permitindo a construção, o treinamento e a execução de diferentes modelos de redes neurais.	Paga

(continuação)

R	O R é uma linguagem para computação estatística e visualização de dados. Seu ambiente é capaz de ser instalado em diferentes sistemas operacionais. Em seu site, é possível realizar o *download* de diferentes manuais que orientam desde iniciantes na linguagem até integrações com outras linguagens (C, C++, Fortran etc.). Em virtude de sua origem estatística, o R foi adotado por diferentes grupos de pesquisas que acabaram por criar pacotes de algoritmos para as mais diversas tarefas de mineração de dados. Todos os pacotes são documentados e validados pela comunidade, servindo de suporte para diferentes pesquisadores. O conjunto de pacotes mais famoso é o Bioconductor, desenvolvido para análises de dados de bioinformática, tais como análises de dados genômicos, tendo seus algoritmos e processos amplamente aceitos nas comunidades de pesquisas médicas e científicas	Gratuita
Wolfram Mathematica	O Mathematica é um software com forte embasamento matemático. Lançado em 1988 para computação técnica em diversas áreas, atualmente o programa possui ferramentas para atuar nas arcas de análise de dados e imagens, computação de grafos, sistemas dinâmicos e complexos.	Paga
RapidMiner	O RapidMiner é um software que atua em processos de mineração de dados. A ferramenta permite a construção visual, por meio de blocos e fluxogramas, de processos complexos de análise e mineração de dados, podendo conectar-se a diferentes fontes de dados, tais conto arquivos e diferentes SGBDs. Mesmo contendo diferentes algoritmos para todos os processos de mineração de dados, o software possui uma API para extensão que permite o desenvolvimento de algoritmos customizados, ampliando consideravelmente a gama de análises possíveis.	Gratuita e Paga
SAS	A empresa SAS (*Statistical Analysis Sustem*) possui uma família de softwares para gerenciamento de bases de dados, análise preditiva, mineração de dados e visualização de dados. Fundada em 1976. Atualmente a empresa figura entre as mais importantes na análise preditiva empresarial.	Paga
SPSS	O SPSS é um software de análise descritiva e preditiva da IBM que possui ferramentas para coleta de dados, realização de estatísticas, construção de modelos, análise de mídia social e atuação em base de dados de larga escala	Paga
Orange	O Orange é um software que permite a construção visual, por meio de blocos e fluxogramas de processos complexos de análise e mineração de dados, sendo baseado na linguagem de programação Python. O Software pode ser instalado nos sistemas operacionais Windows, Mac OS X e Linux, e trabalha com diferentes pacotes de extensão da linguagem Python. Além disso, possui pacotes adicionais para atuar nas áreas de bioinformática, mineração de textos e visualização de dados.	Gratuita

(conclusão)

Mahout	O Mahout é um projeto da Apache com o objetivo de desenvolver algoritmos de aprendizado de máquina escaláveis. A ferramenta conta com algoritmos para agrupamento, classificação e recomendação que podem ser aplicados em bases de dados de larga escala (*Big Data*) com processamento paralelo ou distribuído.	Gratuita
ELKI	O ELKI (*Environment for Developing KDD - Apllications Supported by Index-Structures)* é um software de mineração de dados de código aberto desenvolvido em Java e mantido pela Universidade de Munique Ludwig-Maximilians. O principal foco do ELKI é a pesquisa em algoritmos não supervisionados para tarefas de agrupamento e detecção de anomalias.	Gratuita
LIBSVM	O LIBSVM é uma biblioteca para desenvolvimento de Máquinas de Vetores de Suporte (*Support Vector Machine* - SVM) desenvolvida em Java C++, e mantida por uma comunidade de pesquisadores. As SVMs são técnicas que podem ser aplicadas para tarefas de otimização, classificação, regressão, estimação, entre outras.	Gratuita

FONTE: Elaborado pela autora (2016). Adaptado de Castro e Ferrari (2016)

2.7 PÓS PROCESSAMENTO

De acordo com Milani e Carvalho (2013), o pós-processamento tem como principal objetivo apoiar na verificação de até que ponto estes padrões contribuem na solução do problema inicialmente identificado. As autoras destacam que para operacionalizar o pós-processamento existem várias estratégias propostas, entre elas eliminar a redundância, generalizar, identificar no conjunto aqueles com maior potencial de serem interessantes etc.

Entre as estratégias têm-se a proposta por Hussain *et al.* (2000¸ apud Milani; Carvalho*,* 2013) que constitui um método que identifica, a partir de um conjunto de padrões descobertos, um subconjunto de regras que representam regras de exceção e, além disso, atribui uma medida de interesse para cada regra. O Quadro 5 mostra a estrutura geral das regras de exceção. Nesta tabela A, B e C são conjuntos não-vazios de itens de dados associados e o símbolo "¬" denota a negação lógica. Os autores destacam que uma regra de exceção é uma especialização de uma regra geral e uma regra de exceção associa a um item de dados que nega aquele identificado pela regra geral. Este método assume que regras de senso comum representam padrões conhecidos pelo usuário, tendo em vista que aquelas regras têm uma grande

cobertura, ao contrário das regras de exceção, que em geral são desconhecidas, uma vez que elas têm baixa cobertura. Sendo assim, as regras de exceção tendem a ser surpreendentes, dado o fato de representarem uma contradição em relação à regra de senso comum. É importante observar que a regra de referência auxilia na explicação da causa da regra de exceção.

Quadro 5 - Estrutura das regras de exceção

Regra	Classificação da regra
A → C	Regra geral (alta cobertura e alta confiança)
A, B ¬ C	Regra de exceção (baixa cobertura, alta confiança)
B → ¬ C	Regra de referência (baixa cobertura e/ou baixa confiança)

FONTE: Hussain *et al.* (2000 apud Milani; Carvalho, 2013)

Tan, Steinbach e Kumar (2009, p. 5) destacam que o pós-processamento deve assegurar que somente resultados válidos e úteis sejam incorporados ao sistema de apoio a decisão. Os autores exemplificam a visualização, a qual permite que os analistas explorem os dados e os resultados da mineração dos mesmos a partir de uma diversidade de pontos de vista. Ainda de acordo com os autores, as medições estatísticas ou métodos de teste de hipóteses também podem ser aplicados durante o pós-processamento para eliminar os resultados não legítimos da mineração de dados.

Após a apresentação dos principais conceitos relacionais ao tema, a próxima seção aborda a metodologia proposta para o desenvolvimento do estudo e alcance dos resultados.

3 METODOLOGIA

Esta seção apresenta a caracterização da pesquisa, o seu ambiente de desenvolvimento, bem como a descrição da base de dados objeto do estudo.

3.1 AMBIENTE DA PESQUISA

A pesquisa é realizada em uma organização privada de Curitiba atuante no segmento jurídico. Por motivos de privacidade, o seu nome não é divulgado.

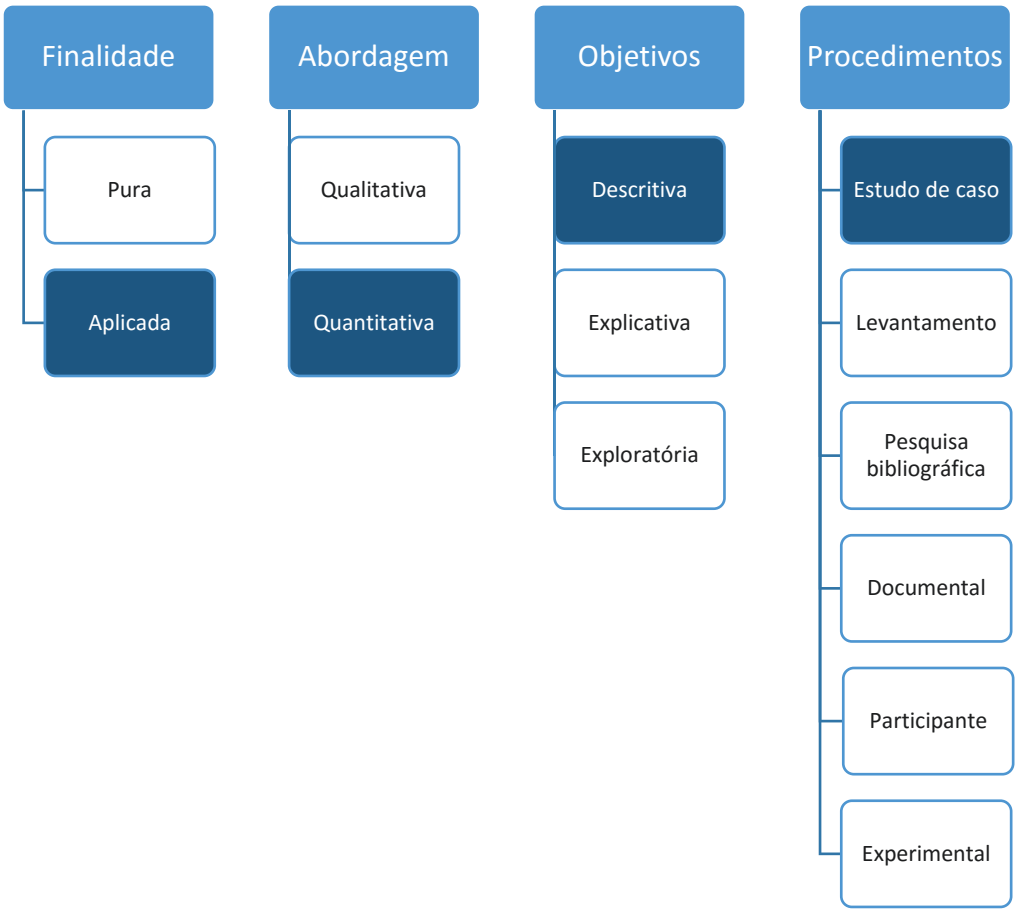
A organização presta serviços em diversas áreas do direito: contencioso; contratos civis e mercantis e apoio legal em licitações; direito administrativo; direito tributário; direito do consumidor; direto do trabalho; direito societário; planejamento sucessório; entre outros. Para a administração dos processos, a organização utiliza um software de gerenciamento específico para esse fim, permitindo realizar o acompanhamento de prazos e processos, verificação da movimentação processual, capturas automáticas pela *internet,* entre outras funcionalidades.

Para o desenvolvimento da pesquisa a organização cedeu a sua base de dados processual, considerando que nenhuma informação pessoal ou ainda que identifique alguma empresa/organização é mencionada nesta pesquisa.

3.2 CARACTERIZAÇÃO DA PESQUISA

A caracterização da pesquisa é realizada com base nos métodos e técnicas de destacadas por Gil (1991) sob o ponto de vista de sua finalidade, abordagem, objetivos e procedimentos, conforme demonstra a Figura 29.

Figura 29 - Caracterização da pesquisa



FONTE: Elaborado pela autora (2016).

Quanto à finalidade da pesquisa, Gil (1991) destaca que a mesma pode ser pura ou aplicada. Essa pesquisa pode ser caracterizada como aplicada, considerando que objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos. O autor destaca que a pesquisa aplicada tem como característica fundamental o interesse na aplicação, utilização e consequências práticas dos conhecimentos.

Quanto à forma de abordagem do problema, a pesquisa pode ser caracterizada como qualitativa ou quantitativa. Essa pesquisa pode ser classificada como quantitativa, pois considera que tudo pode ser quantificável, o que significa, de acordo com Moresi (2003) traduzir em números opiniões e informações para classificá-las e analisá-las.

Quanto aos objetivos, a pesquisa pode ser classificada de acordo com Gil (1991) como exploratória, descritiva ou explicativa. Com base nessa classificação essa pesquisa pode ser considerada descritiva, pois “têm como objetivo primordial a

descrição das características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis". (GIL, 1991). Fernandes e Gomes (2003) complementam que se trata de uma modalidade de pesquisa cujo objetivo principal é descrever, analisar ou verificar as relações entre fatos e fenômenos (variáveis), ou seja, tomar conhecimento do que, com quem, como e qual a intensidade do fenômeno em estudo.

Quanto aos procedimentos técnicas, Gil (1991) destaca que a pesquisa pode ser caracterizada como: estudo de caso; levantamento; pesquisa bibliográfica; documental; participante; ou experimental. A presente pesquisa constitui-se em um estudo de caso, considerando que descreve uma determinada situação do contexto em que está sendo desenvolvida determinada investigação. Fernandes e Gomes (2003) destacam que se trata do estudo de casos isolados, em que a análise deve ser feita com profundidade, detalhadamente e de forma exaustiva, considerando as influências internas e externas.

Yin (2005) destaca que para os estudos de caso, são especialmente importantes cinco componentes de um projeto de pesquisa: as questões de um estudo; suas proposições, se houver; sua (s) unidade (s) de análise; a lógica que une os dados as proposições; e os critérios para se interpretar as descobertas.

O presente estudo de caso é baseado em três etapas principais: análise documental que corresponde à etapa de pré-processamento; mineração de dados; e validação dos resultados, correspondendo a etapa de pós-processamento.

3.3 ANÁLISE DOCUMENTAL

Para Gil (2002) os documentos são fontes ricas e estáveis de dados, pois subsistem ao longo dos tempos e analisá-los requer do pesquisador apenas tempo, haja vista que não apresenta outros custos. Além dessas vantagens, o autor afirma que a análise documental dispensa interação com sujeitos, o que facilita a pesquisa, considerando que quando exige consulta a envolvidos a disponibilidade do respondente é um fator relevante e impactante para o processo de coleta de dados.

A base de dados jurídica analisada contém aproximadamente mil processos cíveis de direito do consumidor. Encontra-se em formato .xls, sendo os seus principais atributos descritos no Quadro 6. Essa descrição é de suma importância para

identificar o tipo de atributo, ou seja, nominal, ordinal, intervalar ou proporcional, sendo essas características decisivas para a escolha do método de mineração de dados.

A análise documental é aplicada aos atributos que contém texto livre, considerando que em alguns casos torna-se necessário consultar os autos processuais para extrair as informações necessárias para a base de dados. Essa etapa pode ser comparada ao pré-processamento, considerando que contempla a seleção, limpeza e transformação de dados.

Quadro 6 - Descrição dos atributos da base de dados jurídica com base no valor e tipo de atributo

(continua)

Atributo	Valor do atributo	Tipo de atributo	Descrição
Mês Base	Variável Ex: jan/16	Categórico - Nominal	Mês e ano atual.
Tipo de Movimentação	Encerrado	Categórico - Ordinal	Situação em que o processo se encontra
	Saldo		
	Entrada		
Data de Entrada	Variável Ex: 99/99/9999	Numérico – Intervalar	Data do recebimento do processo
Réu	Variável	Categórico - Nominal	Nome do réu
Autor	Variável	Categórico - Nominal	Nome do autor
Escritório	Variável	Categórico - Nominal	Nome do escritório
Tipo de Ação	Variável	Categórico - Nominal	Descrição livre sobre o tipo de ação
Número do Processo	Variável Ex: 0000000-00.0000.0.00.0000	Numérico - Intervalar	Número do processo, conforme padrão CNJ
Risco (Mês Anterior)	Remoto	Categórico – Ordinal	Classificação do risco da ação em relação ao mês anterior com base no julgamento, podendo ser remoto (sentença favorável ao réu), possível (sem sentença) ou provável (sentença desfavorável ao réu)
	Possível		
	Provável		
Risco (Mês Atual)	Remoto	Categórico – Ordinal	Classificação do risco da ação em relação ao mês atual com base no julgamento, podendo ser remoto (sentença favorável ao réu), possível (sem sentença) ou provável (sentença desfavorável ao réu)
	Possível		
	Provável		

(continuação)

Valor da Causa	**Variável**	**Numérico - Proporcional**	**Valor monetário referente aos pedidos do autor**
Valor do Risco	Variável	Numérico - Proporcional	Valor monetário referente ao risco que o processo apresenta em caso de julgamento desfavorável
Valor da Provisão	Variável	Numérico - Proporcional	Valor monetário que deve ser reservado para atender às despesas ou eventuais condenações.
Data de Ajuizamento	Variável Ex: 99/99/9999	Numérico – Intervalar	Data de distribuição do processo
Comarca	Variável	Categórico - Nominal	Nome da cidade em que foi distribuído o processo e
UF	Variável Ex: PR	Categórico - Nominal	Estado referente à Comarca que tramita a ação
Órgão Julgador (Número)	Variável Ex: 1	Numérico – Intervalar	Número da instância que a ação foi ajuizada
Órgão Julgador	Procon	Categórico – Ordinal	Nome da instância que ajuizou a ação
	Vara Cível		
	Juizado Especial Cível		
Contrato	Variável Ex: 00000000000	Numérico - Proporcional	Número do contrato de financiamento ou leasing
CPF/CNJP	Variável Ex: 000.000.000-00	Numérico - Proporcional	Número do CPF/CNJP da parte autora
Manutenção de Posse	Deferido	Categórico - Ordinal	Antecipação de tutela que determinou a autorização para a parte se manter na posse do veículo até o encerramento do processo
	Deferido - Condicionado		
	Deferido - Condicionado e com Multa		
	Revogado		
Serasa	Deferido	Categórico - Ordinal	Antecipação de tutela que determinou a remoção da restrição do nome da parte autora
	Deferido - Condicionado		
	Deferido - Condicionado e com Multa		
	Revogado		
Consignação	Deferido	Categórico - Ordinal	Antecipação de tutela que determinou a autorização para a parte autora realizar a consignação em juízo
	Deferido – Condicionado		
	Deferido – Condicionado e com Multa		

(continuação)

	Revogado		
Veículo	Variável	Categórico - Nominal	Nome do automóvel referente ao contrato,
Safra	Variável Ex: 99/99/9999	Numérico – Intervalar	Data da constituição do contrato de financiamento/leasing
Reclamação	Variável	Categórico - Nominal	Texto livre com o assunto abreviado da reclamação da parte autora
Assunto	Variável	Categórico - Nominal	Texto livre com a condensação das alegações e pedidos da parte autora,
Resultado	Variável	Categórico - Nominal	Texto livre com a condensação da sentença
Status Contrato	Em atraso	Categórico - Ordinal	Situação referente ao pagamento do financiamento/leasing
	Regular		
	Quitado		
Classificação da Ação	Tarifa	Categórico - Ordinal	Principais tipos de ações, apresentando discussão de tarifas, tarifas cumuladas com dano moral, indenização por danos morais, revisão de cláusulas contratuais, e outras, que correspondem a todas as demais que não se enquadram nas categorias anteriores.
	Tarifa e Dano Moral		
	Indenizatória		
	Revisional		
	Outras		
Suspenso	Variável	Categórico - Nominal	Se o processo se encontra suspenso
Valor Recuperado	Variável	Numérico - Quantitativo	Valor monetário recuperado na ação.
Motivo Encerramento	Condenação	Categórico – Ordinal	Motivo do encerramento do processo
	Acordo		
	Improcedência		
	Extinção sem mérito		
	Outros		
Valor Final	Variável	Numérico – Proporcional	Valor monetário total pago no processo para o seu encerramento
Data Encerramento	Variável	Numérico – Intervalar	Data do encerramento do processo
Fase Atual	Em grau de recurso	Categórico – Ordinal	Fase atual em que o processo se encontra

(conclusão)

	Julgado		
	Em instância Superior		
	Concluso para sentença		
Tarifa de Cadastro	Variável	Numérico - Proporcional	Valor monetário referente a cobrança de tarifa de cadastro
Registro de Contrato	Variável	Numérico - Proporcional	Valor monetário referente a cobrança de registro de contrato
Gravame	Variável	Numérico - Proporcional	Valor monetário referente a cobrança de inserção de gravame
Seguro Prestamista	Variável	Numérico - Proporcional	Valor monetário referente a cobrança de seguro prestamista
Serviço de Correspondente	Variável	Numérico - Proporcional	Valor monetário referente a cobrança de serviço de correspondente
Serviço de Terceiros	Variável	Numérico - Proporcional	Valor monetário referente a cobrança de serviço de terceiros
Dano Moral	Variável	Numérico - Proporcional	Valor monetário referente a cobrança de dano moral
Tarifa de Avaliação do Bem	Variável	Numérico - Proporcional	Valor monetário referente a cobrança de tarifa de avaliação do bem

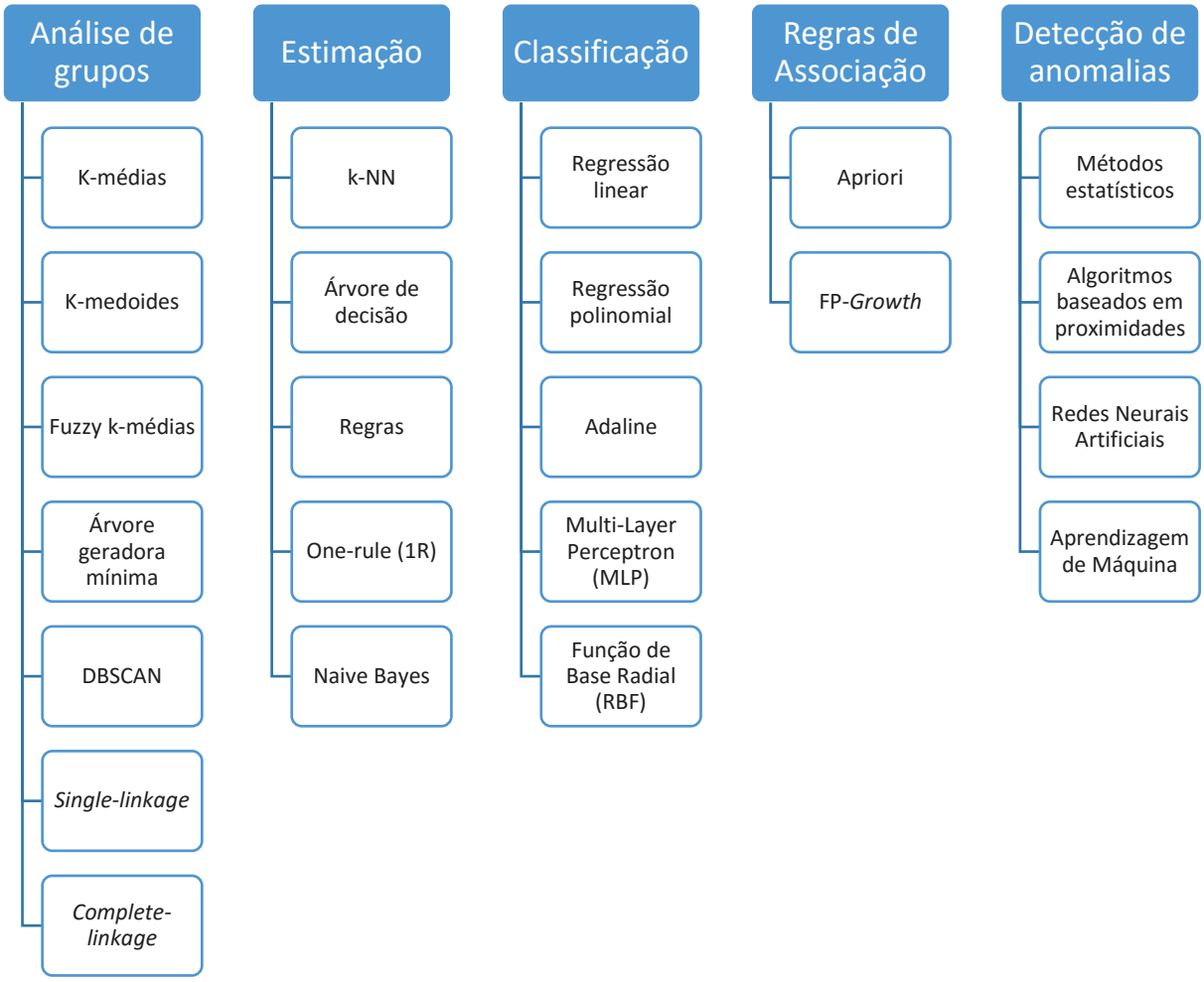
FONTE: Elaborado pela autora (2016)

Tendo concluído a análise documental, a próxima seção aborda o processo de mineração da base de dados.

3.4 MINERAÇÃO DE DADOS

Na etapa de mineração são escolhidos os algoritmos que melhor atendam aos requisitos da base de dados. A escolha é realizada com base nos tipos de atributos previamente descritos no pré-processamento, levando em consideração ainda a complexidade para análise e interpretação dos resultados. Os métodos e algoritmos definidos na revisão da literatura foram sintetizados na Figura 30, de modo a facilitar a visualização e escolha do algoritmo que melhor extraia conhecimento da base de dados jurídica.

Figura 30 - Principais algoritmos de mineração de dados



FONTE: Elaborado pela autora (2016)

Essa etapa também envolve a escolha do software de mineração de dados a ser utilizado. Como critérios, primeiramente foram eliminados os softwares pagos. Por fim, optou-se pela utilização do *Weka*, considerando a familiaridade com o sistema pelos seguintes critérios: já ter sido utilizado em disciplina acadêmica, estar disponível nos laboratórios didáticos do setor, ser livremente distribuído e apresentar compatibilidade com diversos sistemas operacionais.

3.5 VALIDAÇÃO DA PROPOSTA

A validação da proposta corresponde a etapa de pós-processamento, na qual os resultados são analisados para verificar se houve efetivamente alguma descoberta de conhecimento. Para isso, os resultados são apresentados a um profissional da área

jurídica para verificar se o conhecimento extraído contribuiu de alguma forma para a tomada de decisão ou e foram obtidos apenas resultados que já eram evidentes para a área. Castro e Ferrari (2016, p. 11) destacam que um especialista de domínio é capaz de validar se os resultados apresentados fazem sentido e se possuem “boa qualidade”.

A validação da proposta é realizada por meio de uma entrevista com um dos sócios da organização que cedeu a base de dados, visando avaliar os resultados obtidos e obter sugestões de melhorias.

Após a apresentação dos principais elementos da metodologia utilizada para o desenvolvimento do estudo, a próxima seção apresenta os resultados alcançados.

4 RESULTADOS

A seguir é apresentada a análise e descrição estatística da base de dados, a escolha do algoritmo de mineração e os resultados alcançados.

4.1 ANÁLISE DA DASE DE DADOS

A base de dados analisada contém processos jurídicos recebidos pela organização desde 2014 até o mês de janeiro de 2016, totalizando 1169 ativos em diversas fases processuais. Para análise, primeiramente definiu-se como atributo meta a coluna “Motivo Arquivamento”, a qual possui as categorias: condenação; acordo, improcedência; extinção sem mérito; e outros.

Para análise foram filtrados todos os processos da base de dados arquivados - aqueles que já foram julgados e encaminhados ao arquivo - entre o período de abril de 2014 a janeiro de 2016, totalizando 701 processos. Primeiramente foram retiradas as colunas da base de dados que não apresentariam influência sobre o atributo meta analisado. Com isso, restaram as colunas: data de entrada; réu; autor; processo; ação; comarca; UF; Órgão Julgador; risco (mês anterior); risco (atual); valor da causa; valor do risco; valor da provisão; motivo arquivamento; valor pago; e valor recuperado.

Em seguida, a coluna “órgão julgador” foi subdividida em outras duas colunas: “nome órgão julgador” e “número órgão julgador”, pois para análise importa apenas saber o tipo de juízo: Vara Cível (VC), Juizado Especial Cível (JEC) ou PROCON. Depois disso, o atributo “Valor da Causa”, “Valor do Risco” e “Valor da Provisão” foram discretizados considerando 3 (três) intervalos. Em “Valor da Causa” foram encontrados os intervalos: valor mínimo R$0,01 e valor máximo R$87.043,68. Considerando esses valores, foram criados os intervalos I1, I2 e I3 dispostos no Quadro 7.

Quadro 7 - Discretização do atributo "Valor da Causa"

87.043,68 – 0,01 = 87.043,67
87.043,67 / 3 = 29.014,55
Intervalo 1 (I1): 0,01 + 29.014,55 = 29.014,56 → **I1 ≤ 29.014,56**
Intervalo 2 (I2): 29.014,56 + 29.014,55 = 58.029,11 → **29.014,56 < I2 ≤ 58.029,11**
Intervalo 3 (I3): 58.029,11 + 29.014,55 = 87.043,66 → **I3 > 58.029,11**

FONTE: Elaborado pela autora (2016)

Ao concluir a discretização com base nos intervalos apresentados no Quadro 7, 661 valores ficaram no intervalo 1, 37 no intervalo 2 e 3 no intervalo 3.

Em “Valor do Risco” foram encontrados os intervalos: valor mínimo: R$1.000,00 e valor máximo R$35.000,00. Com base nesses valores foram criados os intervalos I1, I2 e I3 dispostos no Quadro 8.

Quadro 8 - Discretização do atributo "Valor do Risco"

35.000 – 1.000 = 34.000	
34.000 / 3 = 11.333	
Intervalo 1 (I1): 1.000 + 11.333 = 12.333	→ **I1 ≤ 12.333**
Intervalo 2 (I2): 12.333 + 11.333 = 23.666	→ **12.333 < I2 ≤ 23.666**
Intervalo 3 (I3): 23.666 + 11.333= 28.930	→ **I3 > 23.666**

FONTE: Elaborado pela autora (2016)

Ao concluir a discretização conforme os intervalos apresentados no Quadro 8, 670 valores ficaram no intervalo 1, 26 no intervalo 2 e 5 no intervalo 3.

Em “Valor da Provisão” foram encontrados os intervalos: valor mínimo R$25,80 e valor máximo R$ 46.151,69. Com base nesses valores foram criados os intervalos I1, I2 e I3 dispostos no Quadro 9.

Quadro 9 - Discretização do atributo "Valor da Provisão"

46.151,69 – 25,80 = 46.125,89	
46.125,89 / 3 = 15.375,29	
Intervalo 1 (I1): 25,80 + 15.375,29 = 15.401,09	→ **I1 ≤ 15.401,09**
Intervalo 2 (I2): 15.401,09 + 15.375,29 = 30.776,38	→ **15.401,09 < I2 ≤ 30.776,38**
Intervalo 3 (I3): 30.776,38 + 15.375,29 = 46.151,67	→ **I3 > 30.776,38**

FONTE: Elaborado pela autora (2016)

Ao concluir a discretização conforme os intervalos apresentados no Quadro 9, 678 valores ficaram no intervalo 1, 18 no intervalo 2 e 5 no intervalo 3.

Ao analisar os números de exemplos nos intervalos foi identificada uma grande discrepância entre eles, pois o intervalo 1 sempre apresenta a maior quantidade de instâncias, tornando tendenciosa a análise. Ao verificar o motivo da dessa discrepância foi identificado que 56% (Tabela 1) dos processos da base de dados tramitam no Juizado Especial Cível (JEC). Esse juizado possui ações limitadas ao teto

de 40 salários mínimos, aproximadamente R$35.200,00 (considerando o salário mínimo em vigor), o que justifica a maioria dos valores correspondentes ao intervalo 1.

Tabela 1 - Distribuição dos processos por Órgão Julgador

Órgão Julgador	Quantidade	%
JEC	393	56,06%
VC	178	25,39%
PROCON	126	17,97%
CEJUSC	2	0,29%
VARA UNICA	2	0,29%
Total Geral	**701**	**100,00%**

FONTE: Elaborado pela autora (2016)

Uma nova discretização foi realizada considerando somente os intervalos presentes nas ações tramitando no JEC, considerando que eles representam a maior categoria. Os intervalos estão dispostos na Tabela 2.

Tabela 2 - Discretização dos atributos Valor da Causa, Valor do Risco e Valor da Provisão tramitando no JEC.

Atributos	Min	Máx	Diferença	÷ 3	Intervalo 1	Intervalo 2	Intervalo 3
Valor da Causa	37,82	35.740	35.702,18	11.900,73	≤ 11.938,55	> I1 ≤ 23.839,27	> 23.839,27
Valor do Risco	15,13	28.960	28.944,87	96.48,29	≤ 9.663,42	> I1 ≤ 19.311,71	> 19.311,71
Valor da Provisão	42,00	46.147,69	46.105,69	15.368,56	≤ 15.410,56	> I1 ≤ 30.779,13	> 30.779,13

FONTE: Elaborado pela autora (2016)

A distribuição dos intervalos obtidos com a discretização estão dispostos na Tabela 3. É possível identificar que o intervalo 1 ainda detém a maior proporção de instâncias.

Tabela 3 - Distribuição dos intervalos segunda discretização

Coluna	Intervalo 1	Intervalo 2	Intervalo 3
Valor da Causa	482	98	121
Valor do Risco	536	151	14
Valor da Provisão	678	18	5

FONTE: Elaborado pela autora (2016)

Para otimizar a distribuição entre os intervalos foi realizada uma nova discretização considerando como valor máximo o correspondente ao intervalo 1: valor da causa: R$ 11.938,55; valor do risco R$ 9.663,42; valor da provação: R$15.410,56, conforme demonstra a Tabela 4.

Tabela 4 - Discretização dos atributos Valor da Causa, Valor do Risco e Valor da Provisão considerando o intervalo 1.

Atributo	Min	Máx	Diferença	÷ 3	Intervalo 1	Intervalo 2	Intervalo 3
Valor da Causa	37,82	11938,55	11900,72	3966,90	≤ 4004,73	> 4004,73 ≤ 7971,64	> 7971,64
Valor do Risco	15,13	9663,42	9648,29	3216,09	≤ 3231,23	> 3231,23 ≤ 6447,32	> 6447,32
Valor da Provisão	42	15410,56	15368,56	5122,85	≤ 5164,85	> 5164,85 ≤ 10287,71	> 10287,71

FONTE: Elaborado pela autora (2016)

A distribuição dos intervalos obtidos com a discretização encontra-se disposta na Tabela 5. É possível identificar que a distribuição ficou mais proporcional, aumentando os valores presentes nos demais intervalos.

Tabela 5 - Distribuição dos intervalos terceira discretização

Coluna	Intervalo 1	Intervalo 2	Intervalo 3
Valor_Causa	354	58	289
Valor_Risco	424	66	211
Valor_Provisão	483	87	131

FONTE: Elaborado pela autora (2016)

Ao concluir os ajustes necessários na base de dados, a próxima subseção aborda a descrição estatística da base de dados.

4.2 DESCRIÇÃO ESTATÍSTICA DA BASE

Para verificar a distribuição dos processos por Estado, foi realizada a contagem por UF. Em seguida, o resultado foi classificado em ordem decrescente, tornando mais fácil a visualização dos Estados que contém a maior quantidade de processos julgados no contexto da base de dados. O resultado encontra-se disposto na Tabela 6.

Tabela 6 - Distribuição dos processos por UF

UF	Quantidade	%
SP	148	21,11%
PR	85	12,13%
RJ	81	11,55%
DF	46	6,56%
PB	41	5,85%
RS	41	5,85%
PE	40	5,71%
SC	38	5,42%
PI	34	4,85%
MG	26	3,71%
MA	22	3,14%
GO	21	3,00%
BA	15	2,14%
ES	13	1,85%
AP	10	1,43%
SE	9	1,28%
MT	5	0,71%
RN	5	0,71%
AL	4	0,57%
MS	4	0,57%
AC	3	0,43%
PA	3	0,43%
RR	3	0,43%
AM	2	0,29%
CE	2	0,29%
Total Geral	**701**	**100,00%**

FONTE: Elaborado pela autora (2016)

A partir da análise, foi possível identificar que a distribuição dos processos encontra-se desproporcional, considerando que SP possui 147 processos, enquanto AM e CE apenas 2. Dessa forma, optou-se pela distribuição dos processos por região: sul, sudeste, centro-oeste, norte, nordeste. Os resultados foram classificados em ordem decrescente, conforme demonstra a Tabela 7.

Tabela 7 - Distribuição dos processos por Região

Região	Quantidade	%
SUDESTE	268	38,23%
NORDESTE	172	24,54%
SUL	164	23,40%
CENTRO-OESTE	76	10,84%
NORTE	21	3,00%
Total Geral	**701**	**100,00%**

FONTE: Elaborado pela autora (2016)

Após concluir a análise da distribuição dos processos por Estado, foi realizada a análise em relação ao motivo do arquivamento do processo: condenação, extinção sem mérito, improcedência, acordo, outros. A seguir é detalhada cada categoria.

Nos processos encerrados por condenação o juiz reconhece a procedência da reclamação, acolhendo os pedidos do autor, sendo, portanto, desfavorável ao réu e favorável ao autor.

Nos processos encerrados por acordo, Chagas (2012) destaca que as próprias partes se antecipam e, no curso do processo, encontram, por si mesmas, uma solução para a lide. O autor explica ainda que ao juiz, nesses casos, compete apenas homologar o negócio jurídico praticado pelos litigantes, para integrá-lo ao processo e dar-lhe eficácia equivalente ao de julgamento de mérito. Chagas (2012) exemplifica que é o que ocorre quando o autor renuncia ao direito material sobre que se funda a ação (art. 269, V – CPC), ou quando as partes fazem transação sobre o objeto do processo (art. 269, III – CPC).

Nos processos encerrados por improcedência o juiz rejeita os pedidos do autor, sendo o julgamento desfavorável ao autor e favorável ao réu. No entendimento do julgador não havia o direito postulado, ou seja, quando o autor pediu algo não lhe era devido.

Nos processos encerrados por extinção sem mérito ocorre a inexistência material do fato, impedindo resolução de mérito sobre o objeto litigioso civil, nos termos do art. 267, V, do CPC, no qual: extingue-se o processo, sem resolução de mérito “quando o juiz acolher a alegação de perempção, litispendência ou de coisa julgada”. O juiz não resolverá o mérito quando:

I. indeferir a petição inicial;
II. o processo ficar parado durante mais de 1 (um) ano por negligência das partes;

III. por não promover os atos e as diligências que lhe incumbir, o autor abandonar a causa por mais de 30 (trinta) dias;
IV. verificar a ausência de pressupostos de constituição e de desenvolvimento válido e regular do processo
V. reconhecer a existência de perempção, de litispendência ou de coisa julgada;
VI. verificar ausência de legitimidade ou de interesse processual;
VII. acolher a alegação de existência de convenção de arbitragem ou quando o juízo arbitral reconhecer sua competência;
VIII. homologar a desistência da ação;
IX. em caso de morte da parte, a ação for considerada intransmissível por disposição legal; e
X. nos demais casos prescritos neste Código. (BRASIL, 2015)

Os processos encerrados classificados como “outros” compreendem todos os demais motivos de arquivamento que não correspondem à classificação anteriormente apresentada. São comuns casos de reclamações no PROCON, obrigações de fazer, entre outros.

A classificação da distribuição dos processos por motivo de arquivamento foi realizada em ordem decrescente, conforme demonstra a Tabela 8.

Tabela 8 - Distribuição dos processos por motivo de arquivamento

Motivo Arquivamento	Quantidade	%
OUTROS	181	25,82%
CONDENACAO	170	24,25%
IMPROCEDENCIA	148	21,11%
EXTINCAO SEM MERITO	135	19,26%
ACORDO	62	8,84%
CONDENACAO	5	0,71%
Total Geral	**701**	**100,00%**

FONTE: Elaborado pela autora (2016)

Outro fator de impacto para a análise da base de dados corresponde a classificação da ação, pois o motivo do arquivamento pode estar relacionado com o tipo reclamação que deu origem a ação. Por exemplo, ações indenizatórias tem maior possibilidade de serem encerradas por condenação ou acordo. A subseção 1.4 (delimitação da pesquisa) explica o contexto de cada classificação. A Tabela 9 demonstra a distribuição dos processos por tipo de ação, ordenados em ordem decrescente.

Tabela 9 - Distribuição dos processos por tipo de ação

Ação	Quantidade	%
OUTRAS	170	24,25%
TARIFA	165	23,54%
REVISIONAL	136	19,40%
INDENIZATORIA	124	17,69%
TARIFA E DANO MORAL	106	15,12%
Total Geral	**701**	**100,00%**

FONTE: Elaborado pela autora (2016)

Ao concluir a análise e descrição estatística da base de dados, a próxima subseção aborda o processo de mineração de dados.

4.3 MINERAÇÃO DE DADOS

Com a identificação dos atributos pertinentes à análise do problema a ser investigado com a mineração é necessário preparar a base de dados removendo todos os caracteres especiais e os espaçamentos presentes nos títulos das colunas. Concluída essa etapa, a planilha é salva no formato .csv (separado por vírgulas). Contudo, ressalta-se que o arquivo deve ser aberto em um editor de texto, substituindo todos os “ponto e vírgula” por “vírgula”, pois na conversão as instâncias são separadas automaticamente por ponto e vírgula. Após esse procedimento, é realizada a conversão do arquivo .csv para o arquivo com extensão .ARFF (*Attribute-Relation File Format).* Por fim, a base de dados pode ser importada no software para mineração de dados. Para fins desse estudo optou-se pela utilização do software Weka, conforme mencionado na seção 3 (metodologia).

Ao importar a base de dados os atributos numéricos “Mês”, “Ano” e “Número Órgão Julgador” foram removidos, restando apenas os nominais. O atributo “UF” e “Comarca” também foram removidos, considerando que foi criada a classificação por região. Por fim, restaram os dez atributos dispostos no Quatro 10.

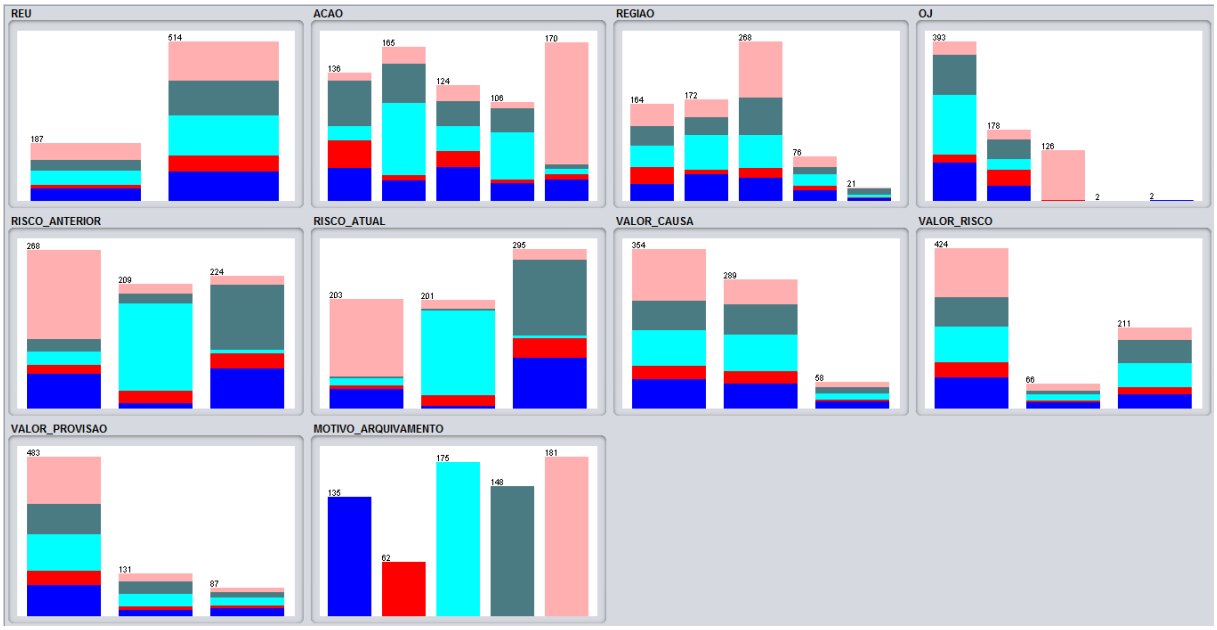
Quadro 10 - Atributos e respectivos valores da base de dados

Atributo	Valor
Reu	1; 2
Acao	Revisional; Tarifa; Indenizatoria; Tarifa e Dano Moral; Outras
Regiao	Sul; Nordeste; Sudeste; Centro-Oeste; Norte
OJ	JEC; VC; PROCON; Vara_Unica; CEJUSC
Risco_Anterior	Possivel; Provavel; Remoto
Risco_Atual	Possivel; Provavel; Remoto
Valor_Causa	I1; I2; I3
Valor_Risco	I1; I2; I3
Valor_Provisao	I1; I2; I3
Motivo_Arquivamento	Extincao_Sem_Merito; Acordo; Condenacao; Improcedencia; Outros

FONTE: Elaborado pela autora (2016)

A Figura 31 apresenta a visualização gráfica do Quadro 10. As cores são separadas de acordo com a quantidade de atributos correspondentes ao atributo meta (motivo_arquivamento).

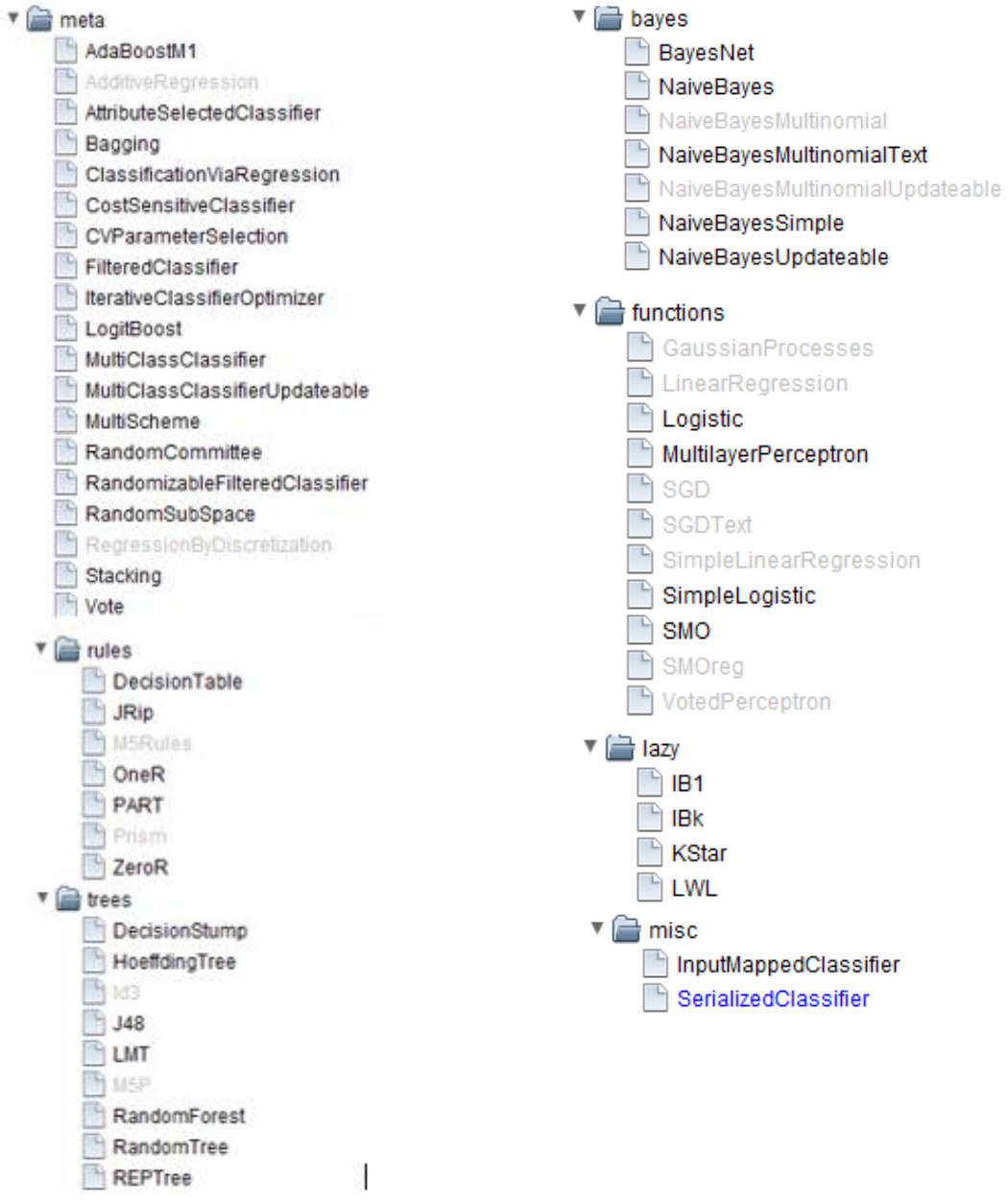
Figura 31 - Histograma valores dos atributos da base de dados



FONTE: Dados da pesquisa utilizando o software Weka (2016)

Para proceder a análise da base é necessário escolher os métodos que serão utilizados para dar suporte a análise dos dados. Primeiramente optou-se pelas heurísticas de classificação, considerando que são as mais conhecidas e utilizadas, pois consistem em associar objetos a um conjunto pré-definido de classes de acordo com as suas características. Na tarefa de classificação foram encontradas as heurísticas: *bayes, functions* (funções), *lazy, meta, misc, rules* (regras) e *trees* (árvores), conforme demonstra a Figura 32.

Figura 32 - Heurísticas e algoritmos de classificação ativados no Weka

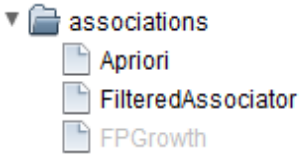


FONTE: Dados da pesquisa utilizando o software Weka (2016)

Para fins desse estudo são utilizadas apenas as heurísticas de regras e árvores, pois apresentam maior facilidade para compreensão dos resultados.

Na tarefa de associação foram ativados os algoritmos Apriori e FilteredAssociator, conforme demonstra a Figura 33. Para o estudo é utilizado apenas o algoritmo Apriori, considerando ser o método mais conhecido para mineração de regras de associação.

Figura 33 - Algoritmos de associação ativados no Weka



FONTE: Dados da pesquisa utilizando o software Weka (2016)

As heurísticas apresentadas são abordadas nas próximas subseções. Contudo, antes de analisar os resultados, torna-se necessário exemplificar alguns conceitos básicos descritos por Martinez, Casal e Janeiro (2009), considerando que aparecerão nos resultados fornecidos pelo software Weka:

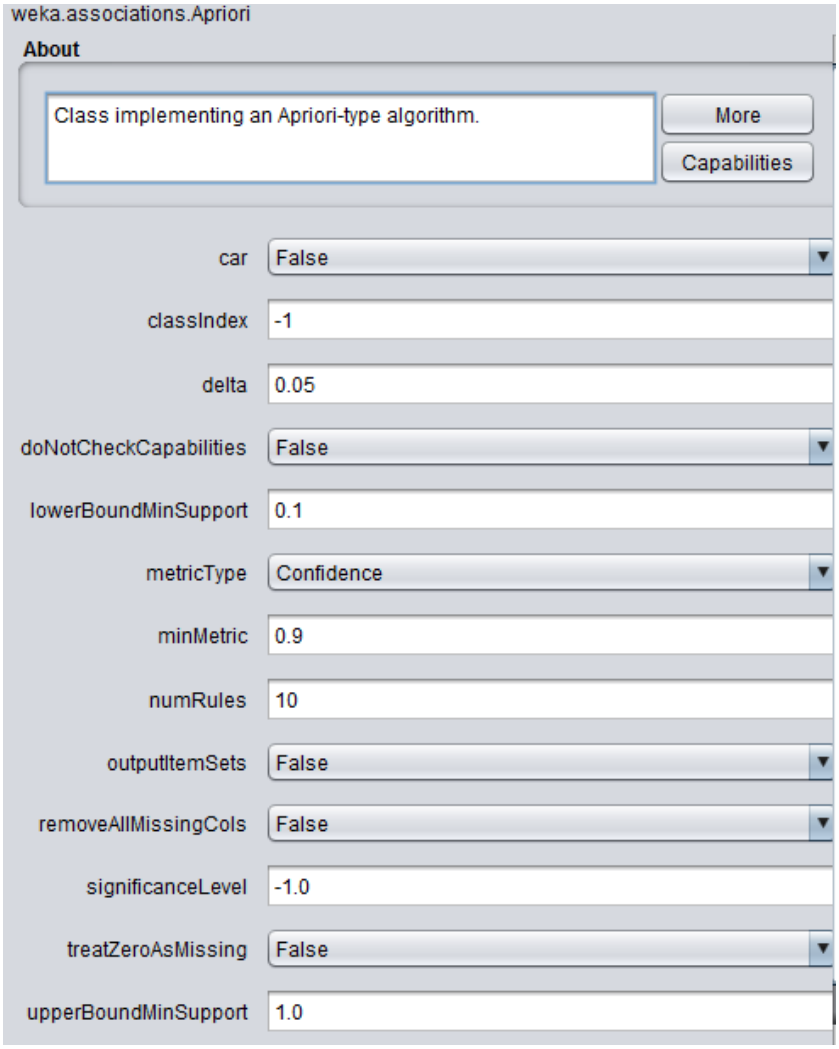
- *Kappa Statistic*: índice que compara o valor encontrado nas observações com aquele que se pode esperar do acaso. É o valor calculado dos resultados encontrados nas observações e relatado como um decimal (0 a 1). Quanto menor o valor de Kappa, menor a confiança de observação, o valor 1 implica a correlação perfeita.
- *mean absolute error*: média da diferença entre os valores atuais e os preditos em todos os casos, é a média do erro da predição.
- *true Positives (*TP): são os valores classificados verdadeiramente positivos.
- *false Positives* (FP): são os falsos positivos, são os dados classificados erroneamente como positivos pelo classificador.
- *precision* (Precisão): é o valor da predição positiva (número de casos positivos por total de casos cobertos), muito influenciada pela especificidade e pouco pela sensitividade. Sensitividade é o número de casos positivos que são verdadeiramente positivos e especificidade é o número de casos negativos que são verdadeiramente negativos.

- *recall* (Cobertura): é o valor da cobertura de casos muito influenciada pela sensitividade e pouco pela especificidade. É calculada por número de casos cobertos pelo número total de casos aplicáveis.
- *f-measure*: usada para medir o desempenho, pois combina valores de cobertura e precisão de uma regra numa única fórmula [2*Prec*Rec / (Prec+Rec)].
- *root relative squared error*: reduz o quadrado do erro relativo na mesma dimensão da quantidade sendo predita incluindo raiz quadrada. Assim como a raiz quadrada do erro significativo (*root mean-squared error*), este exagera nos casos em que o erro da predição foi significativamente maior do que o erro significativo.
- *relative absolute error*: é o erro total absoluto. Em todas as mensurações de erro, valores mais baixos significam maior precisão do modelo, com o valor próximo de zero temos o modelo estatisticamente perfeito.
- *root mean-squared error:* usado para medir o sucesso de uma predição numérica. Este valor é calculado pela média da raiz quadrada da diferença entre o valor calculado e o valor correto. O root *mean-squared error* é simplesmente a raiz quadrada do *mean-squared-error* (dá o valor do erro entre os valores atuais e os valores preditos).

4.3.1 Apriori

O algoritmo Apriori é definido pelo software Weka como “iterativamente reduz o suporte mínimo até encontrar o número necessário de regras com o dado de confiança mínimo. O algoritmo tem uma opção para minerar regras de associação de classe”. Apresenta como parâmetros default os valores dispostos na Figura 34.

Figura 34 - Parâmetros *default* Apriori



FONTE: Dados da pesquisa utilizando o software Weka (2016)

O Weka apresenta ainda um resumo sobre o significado das opções dos parâmetros:

- verbose - Se ativado o algoritmo será executado no modo detalhado. Parâmetro mantido como *default*.
- minMetric - pontuação métrica mínima. Considera apenas as regras com pontuações superiores a este valor. Corresponde ao valor mínimo para a métrica selecionada em metricType. Parâmetro mantido como *default*.
- num*Rules* - número máximo de regras que serão mostradas na tela de resultados. Parâmetro alterado para 100.000.
- lowerBoundMinSupport - um limite inferior para o suporte mínimo. Parâmetro mantido como *default*.

- classIndex - Índice do atributo de classe. Se for definido como -1, o último atributo é tomado como atributo de classe. Parâmetro mantido como *default*, pois o último atributo corresponde ao atributo meta motivo_arquivamento.
- outputItemSets - se configurado como “true”, na saída, além de exibir as regras mineradas, exibirá também os itemsets frequentes. Parâmetro mantido como *default*.
- car - Se as regras de associação de classe são extraídos em vez de regras de associação (geral). Parâmetro mantido como *default*.
- doNotCheckCapabilities - Se for definido, as capacidades do associador não são verificados antes do associados ser construído. Parâmetro mantido como *default*.
- removeAllMissingCols - Remover colunas com todos os valores em falta. Parâmetro mantido como *default*.
- significanceLevel - O nível de significância ou teste de significância (única métrica de confiança). Parâmetro mantido como *default*.
- treatZeroAsMissing - Se estiver ativado, a zero (isto é, o primeiro valor de um valor nominal) é tratado da mesma forma que um valor em falta. Parâmetro mantido como *default*.
- delta - reduz o suporte iterativamente por este valor, partindo do limite superior até que o limite inferior seja alcançado. Parâmetro mantido como *default*.
- metricType – trata-se da especificação da medida de interesse que irá determinar a validade da regra. O conjunto de resultados minerados será ordenado de acordo com essa medida. É possível escolher a medida: confiança, *lift, conviction* e *leverage*. Parâmetro mantido como *default*.
- upperBoundMinSupport - Limite superior para o apoio mínimo. Comece de forma iterativa diminuindo o apoio mínimo a partir deste valor. Parâmetro mantido como *default*.

No primeiro experimento foram considerados os parâmetros *default* mencionados acima, com exceção do número de linhas que foi alterado para apresentar o máximo de resultados possíveis. O algoritmo apresentou 2047 regras, gerando dificuldade para efetuar a análise. Dessa forma, foi realizado um segundo experimento retirando alguns atributos que não seriam interessantes para a análise,

mantendo apenas: acao; regiao; OJ; motivo_arquivamento. Como resultado, foram apresentadas 14 regras (Figura 35):

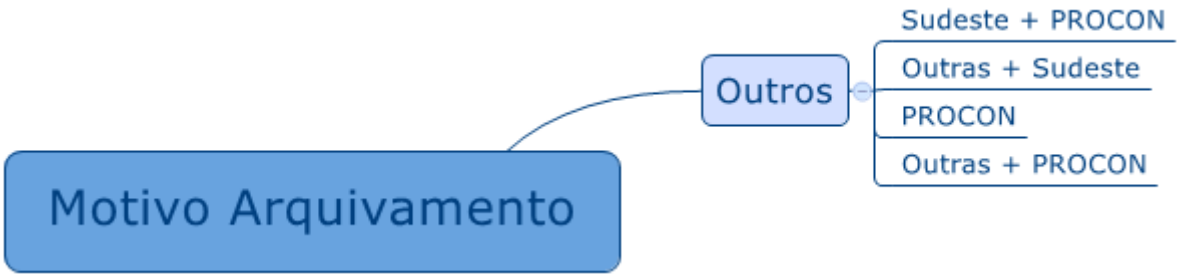
Figura 35 - Resultado Apriori Weka com Confiança 0.9

```
 1. REGIAO=SUDESTE OJ=PROCON 77 ==> MOTIVO_ARQUIVAMENTO=OUTROS 77    <conf:(1)> lift:(3.87) lev:(0.08) [57] conv:(57.12)
 2. ACAO=OUTRAS REGIAO=SUDESTE OJ=PROCON 72 ==> MOTIVO_ARQUIVAMENTO=OUTROS 72    <conf:(1)> lift:(3.87) lev:(0.08) [53] conv:(53.41)
 3. OJ=PROCON 126 ==> MOTIVO_ARQUIVAMENTO=OUTROS 124    <conf:(0.98)> lift:(3.81) lev:(0.13) [91] conv:(31.16)
 4. ACAO=OUTRAS OJ=PROCON 120 ==> MOTIVO_ARQUIVAMENTO=OUTROS 118    <conf:(0.98)> lift:(3.81) lev:(0.12) [87] conv:(29.67)
 5. ACAO=TARIFA E DANO MORAL 106 ==> OJ=JEC 103    <conf:(0.97)> lift:(1.73) lev:(0.06) [43] conv:(11.64)
 6. OJ=PROCON 126 ==> ACAO=OUTRAS 120    <conf:(0.95)> lift:(3.93) lev:(0.13) [89] conv:(13.63)
 7. OJ=PROCON MOTIVO_ARQUIVAMENTO=OUTROS 124 ==> ACAO=OUTRAS 118    <conf:(0.95)> lift:(3.92) lev:(0.13) [87] conv:(13.42)
 8. OJ=PROCON 126 ==> ACAO=OUTRAS MOTIVO_ARQUIVAMENTO=OUTROS 118    <conf:(0.94)> lift:(5.05) lev:(0.13) [94] conv:(11.4)
 9. ACAO=TARIFA MOTIVO_ARQUIVAMENTO=CONDENACAO 77 ==> OJ=JEC 72    <conf:(0.94)> lift:(1.67) lev:(0.04) [28] conv:(5.64)
10. REGIAO=SUDESTE OJ=PROCON 77 ==> ACAO=OUTRAS 72    <conf:(0.94)> lift:(3.86) lev:(0.08) [53] conv:(9.72)
11. REGIAO=SUDESTE OJ=PROCON MOTIVO_ARQUIVAMENTO=OUTROS 77 ==> ACAO=OUTRAS 72    <conf:(0.94)> lift:(3.86) lev:(0.08) [53] conv:(9.72)
12. REGIAO=SUDESTE OJ=PROCON 77 ==> ACAO=OUTRAS MOTIVO_ARQUIVAMENTO=OUTROS 72    <conf:(0.94)> lift:(5.04) lev:(0.08) [57] conv:(10.45)
13. ACAO=OUTRAS MOTIVO_ARQUIVAMENTO=OUTROS 130 ==> OJ=PROCON 118    <conf:(0.91)> lift:(5.05) lev:(0.13) [94] conv:(8.2)
14. ACAO=OUTRAS REGIAO=SUDESTE MOTIVO_ARQUIVAMENTO=OUTROS 80 ==> OJ=PROCON 72    <conf:(0.9)> lift:(5.01) lev:(0.08) [57] conv:(7.29)
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Como o foco da pesquisa consiste somente em verificar se há relações entre o Estado e o motivo do arquivamento, as relações que não atendiam esse requisito foram descartadas, restando apenas as regras 1, 2, 3, 4, 8 e 12. Com base nessas regras é possível analisar que ações classificadas como “outras” tramitando na região sudeste e pelo PROCON apresentam motivo de arquivamento “Outros”. Essas informações podem ser visualizadas no mapa conceitual disposto na Figura 36 criado para facilitar a análise.

Figura 36 - Mapa conceitual algoritmo Apriori



FONTE: Elaborado pela autora (2016)

Alterando a confiança para 0.8 foram geradas 20 regras (Figura 37). De maneira geral elas não apresentaram novas informações relevantes, com exceção da linha 12 que demonstra que se as ações tramitam na região Sudeste e no Procon, então elas serão classificadas como “Outras’ e terão como motivo de arquivamento

“Outros”. Além disso, a linha 17 acrescenta uma nova informação, indicando que se o motivo do arquivamento for por Condenação, então a ação tramita no JEC.

Figura 37 - Resultado Apriori Weka com Confiança 0.8

```
 1. REGIAO=SUDESTE OJ=PROCON 77 ==> MOTIVO_ARQUIVAMENTO=OUTROS 77    <conf:(1)> lift:(3.87) lev:(0.08) [57] conv:(57.12)
 2. ACAO=OUTRAS REGIAO=SUDESTE OJ=PROCON 72 ==> MOTIVO_ARQUIVAMENTO=OUTROS 72    <conf:(1)> lift:(3.87) lev:(0.08) [53] conv:(53.41)
 3. OJ=PROCON 126 ==> MOTIVO_ARQUIVAMENTO=OUTROS 124    <conf:(0.98)> lift:(3.81) lev:(0.13) [91] conv:(31.16)
 4. ACAO=OUTRAS OJ=PROCON 120 ==> MOTIVO_ARQUIVAMENTO=OUTROS 118    <conf:(0.98)> lift:(3.81) lev:(0.12) [87] conv:(29.67)
 5. ACAO=TARIFA E DANO MORAL 106 ==> OJ=JEC 103    <conf:(0.97)> lift:(1.73) lev:(0.06) [43] conv:(11.64)
 6. OJ=PROCON 126 ==> ACAO=OUTRAS 120    <conf:(0.95)> lift:(3.93) lev:(0.13) [89] conv:(13.63)
 7. OJ=PROCON MOTIVO_ARQUIVAMENTO=OUTROS 124 ==> ACAO=OUTRAS 118    <conf:(0.95)> lift:(3.92) lev:(0.13) [87] conv:(13.42)
 8. OJ=PROCON 126 ==> ACAO=OUTRAS MOTIVO_ARQUIVAMENTO=OUTROS 118    <conf:(0.94)> lift:(5.05) lev:(0.13) [94] conv:(11.4)
 9. ACAO=TARIFA MOTIVO_ARQUIVAMENTO=CONDENACAO 77 ==> OJ=JEC 72    <conf:(0.94)> lift:(1.67) lev:(0.04) [28] conv:(5.64)
10. REGIAO=SUDESTE OJ=PROCON 77 ==> ACAO=OUTRAS 72    <conf:(0.94)> lift:(3.86) lev:(0.08) [53] conv:(9.72)
11. REGIAO=SUDESTE OJ=PROCON MOTIVO_ARQUIVAMENTO=OUTROS 77 ==> ACAO=OUTRAS 72    <conf:(0.94)> lift:(3.86) lev:(0.08) [53] conv:(9.72)
12. REGIAO=SUDESTE OJ=PROCON 77 ==> ACAO=OUTRAS MOTIVO_ARQUIVAMENTO=OUTROS 72    <conf:(0.94)> lift:(5.04) lev:(0.08) [57] conv:(10.45)
13. ACAO=OUTRAS MOTIVO_ARQUIVAMENTO=OUTROS 130 ==> OJ=PROCON 118    <conf:(0.91)> lift:(5.05) lev:(0.13) [94] conv:(8.2)
14. ACAO=OUTRAS REGIAO=SUDESTE MOTIVO_ARQUIVAMENTO=OUTROS 80 ==> OJ=PROCON 72    <conf:(0.9)> lift:(5.01) lev:(0.08) [57] conv:(7.29)
15. REGIAO=SUDESTE MOTIVO_ARQUIVAMENTO=OUTROS 94 ==> ACAO=OUTRAS 80    <conf:(0.85)> lift:(3.51) lev:(0.08) [57] conv:(4.75)
16. ACAO=TARIFA 165 ==> OJ=JEC 140    <conf:(0.85)> lift:(1.51) lev:(0.07) [47] conv:(2.79)
17. MOTIVO_ARQUIVAMENTO=CONDENACAO 175 ==> OJ=JEC 148    <conf:(0.85)> lift:(1.51) lev:(0.07) [49] conv:(2.75)
18. ACAO=INDENIZATORIA 124 ==> OJ=JEC 102    <conf:(0.82)> lift:(1.47) lev:(0.05) [32] conv:(2.37)
19. REGIAO=SUDESTE MOTIVO_ARQUIVAMENTO=OUTROS 94 ==> OJ=PROCON 77    <conf:(0.82)> lift:(4.56) lev:(0.09) [60] conv:(4.28)
20. ACAO=OUTRAS REGIAO=SUDESTE 98 ==> MOTIVO_ARQUIVAMENTO=OUTROS 80    <conf:(0.82)> lift:(3.16) lev:(0.08) [54] conv:(3.83)
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Por fim, alterando a confiança para 1.0 foram obtidas apenas 2 regras, conforme demonstra a Figura 38. De maneira geral não houve nova descoberta, pois a regra indica as mesmas constatações identificadas previamente.

Figura 38 - Resultado Apriori Weka com Confiança 1.0

```
1. REGIAO=SUDESTE OJ=PROCON 77 ==> MOTIVO_ARQUIVAMENTO=OUTROS 77    <conf:(1)> lift:(3.87) lev:(0.08) [57] conv:(57.12)
2. ACAO=OUTRAS REGIAO=SUDESTE OJ=PROCON 72 ==> MOTIVO_ARQUIVAMENTO=OUTROS 72    <conf:(1)> lift:(3.87) lev:(0.08) [53] conv:(53.41)
```

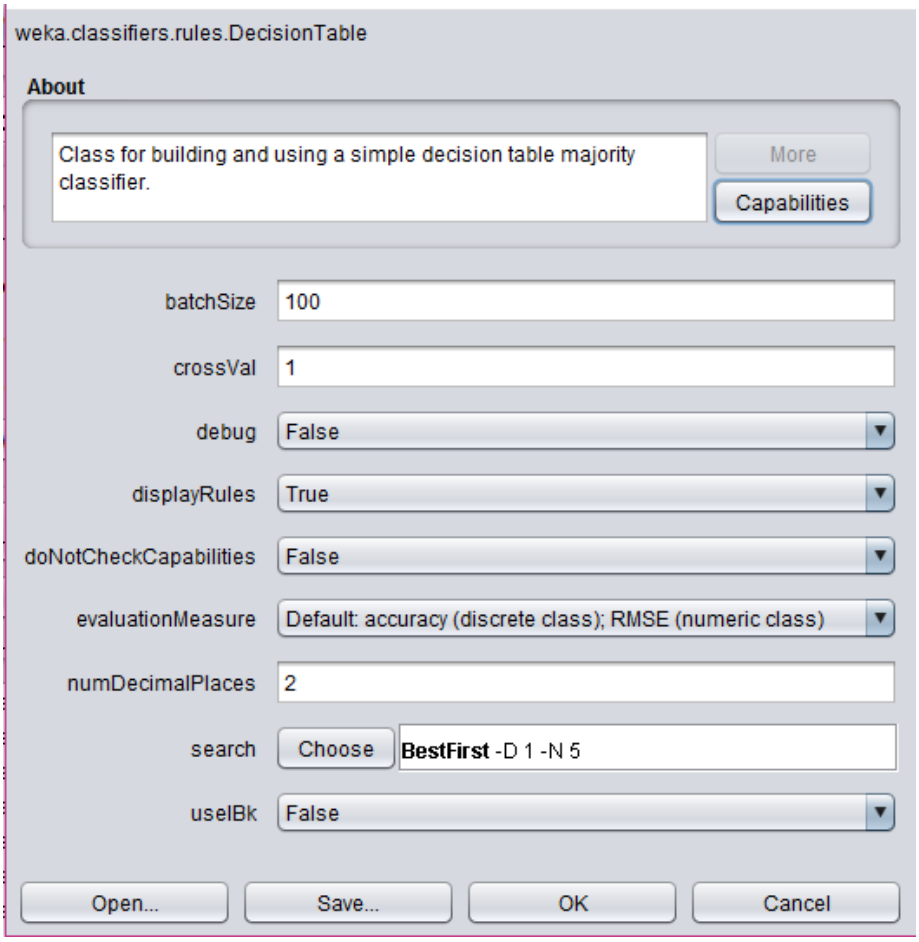
FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Ao concluir a análise dos resultados alcançados com a heurística de associação prosseguiu-se para a análise dos métodos de classificação.

4.3.2 PART

Na heurística de regras foi executado o algoritmo PART, o qual é definido pelo software Weka como: “classe para gerar uma lista de decisão PART. Constrói uma árvore de decisão C4.5 parcial em cada iteração e faz a "melhor" folha em uma regra”. Apresenta como parâmetros *default* os itens dispostos na Figura 39.

Figura 39 - Parâmetros *Default Decision Table*



FONTE: Dados da pesquisa utilizando o software Weka (2016)

Ao executar o algoritmo PART com os parâmetros *default* foram geradas 57 regras (Apêndice B) em 0,2 segundos. Mantendo apenas os quatro atributos principais (acao; regiao; OJ; motivo_arquivamento) foram geradas 17 regras em 0,08 segundos, conforme demonstra o Quadro 11.

Quadro 11 - Aplicação do algoritmo PART - Segundo Experimento

1. OJ = JEC AND ACAO = TARIFA: CONDENACAO (140.0/68.0)
2. OJ = PROCON: OUTROS (126.0/2.0)
3. ACAO = TARIFA E DANO MORAL: CONDENACAO (106.0/56.0)
4. ACAO = OUTRAS AND REGIAO = NORDESTE AND OJ = JEC: EXTINCAO_SEM_MERITO (3.0/1.0)
5. OJ = JEC: EXTINCAO_SEM_MERITO (147.0/91.0)
6. ACAO = OUTRAS AND REGIAO = SUDESTE: EXTINCAO_SEM_MERITO (19.0/11.0)
7. REGIAO = SUL AND ACAO = REVISIONAL: ACORDO (40.0/21.0)
8. REGIAO = SUDESTE: I MPROCEDENCIA (40.0/24.0)
9. ACAO = OUTRAS AND REGIAO = SUL: EXTINCAO_SEM_MERITO (6.0/3.0)
10. ACAO = OUTRAS AND REGIAO = CENTRO-OESTE: CONDENACAO (4.0/1.0)
11. ACAO = TARIFA AND REGIAO = SUL: IMPROCEDENCIA (4.0/2.0)
12. ACAO = TARIFA AND REGIAO = NORDESTE: CONDENACAO (4.0/1.0)
13. REGIAO = NORDESTE: OUTROS (24.0/16.0)
14. ACAO = REVISIONAL: IMPROCEDENCIA (23.0/13.0)
15. REGIAO = SUL: ACORDO (7.0/4.0)
16. ACAO = TARIFA: OUTROS (4.0/1.0)
17. : CONDENACAO (4.0/2.0)

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Para facilitar a visualização das informações os resultados foram separados em colunas, com base no motivo do arquivamento. No Quadro 12 é possível identificar a tendência de realização de acordos em ações revisionais que tramitam na região sul. Também é possível analisar que são comuns condenações em ações de cobrança de tarifa e dano moral, ações de tarifa tramitando na região nordeste e ações outras tramitando na região centro-oeste. Dessa forma, esses seriam os casos mais críticos que deveriam ser analisados pelo escritório de advocacia para conseguir reduzir a quantidade de condenações. Os casos de extinção sem mérito apresentaram maior variação, sendo que geralmente correspondem a reclamações de tarifa tramitando no JEC, ações outras tramitando no JEC da região nordeste e ações outras tramitando na região sudeste e sul. Foram identificados padrões de improcedência na região sudeste, em ações de tarifa tramitando na região sul e em ações revisionais. Por fim, as ações arquivadas por motivo Outros correspondem as ações “outras” tramitando no PROCON, na região nordeste e ações discutindo tarifas.

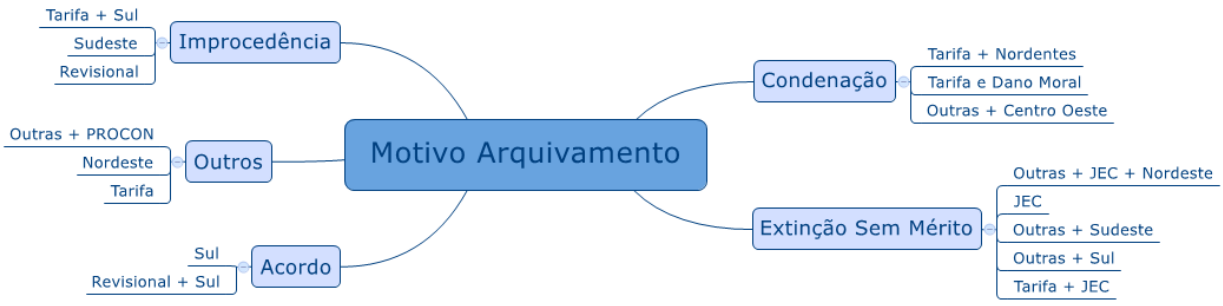
Quadro 12 - Resultado da aplicação do algoritmo PART para identificar o motivo arquivamento

MOTIVO ARQUIVAMENTO	ACAO	OJ	REGIÃO
ACORDO	REVISIONAL		SUL
			SUL
CONDENACAO	TARIFA E DANO MORAL		
	OUTRAS		CENTRO-OESTE
	TARIFA		NORDESTE
EXTINCAO_SEM_MERITO	TARIFA	JEC	
	OUTRAS	JEC	NORDESTE
		JEC	
	OUTRAS		SUDESTE
	OUTRAS		SUL
IMPROCEDENCIA			SUDESTE
	TARIFA		SUL
	REVISIONAL		
OUTROS	OUTRAS	PROCON	
			NORDESTE
	TARIFA		

FONTE: Elaborado pela autora (2016)

A Figura 40 apresenta o mapa conceitual com base nos resultados anteriormente apresentados.

Figura 40 - Mapa conceitual algoritmo PART



FONTE: Elaborado pela autora (2016)

O Quadro 13 demonstra resumidamente a comparação entre a precisão que os métodos foram capazes de prever com os experimentos:

Quadro 13 - Comparação dos resultados dos experimentos com aplicação do algoritmo PART.

	Experimento 1		**Experimento 2**	
Correctly Classified Instances	444	63,3381 %	349	49,786 %
Incorrectly Classified Instances	257	36,6619 %	352	50,214 %
Kappa Statistic	0,5285		0,3539	
Mean absolute error	0,1704		0,2338	
Root mean squared error	0,3278		0,3485	
Relative absolute error	54,4967 %		74,7582 %	
Root relative squared error	82,9222 %		88,1387 %	
Total Number of Instances	701		701	

FONTE: Elaborado pela autora (2016)

A partir dos resultados é possível identificar a porcentagem de instâncias que foram classificadas correta ou incorretamente. Do total de instâncias (701), no primeiro experimento 444 foram classificadas corretamente e 257 incorretamente. Já no segundo experimento 349 foram classificadas corretamente e 352 incorretamente. Com base nesse parâmetro, o experimento 1 apresentou resultados mais satisfatórios, pois teve maior porcentagem de assertividade na classificação das instâncias.

O *Kappa Statistic* (Estatística Kappa) indica o grau de concordância, sendo obtido o resultado de 0,5285 no primeiro experimento e 0,3539 no segundo. Apesar do primeiro experimento apresentar estatística superior quando comparado com o segundo, o resultado é inconclusivo, considerando que os valores ficaram na metade do intervalo entre 0 e 1, não permitindo afirmar que existe correlação.

O *Mean Absolute Error (*Erro Médio Absoluto) foi de 0,1704 no primeiro experimento e 0,2338 no segundo. Considerando esse parâmetro o primeiro experimento apresentou melhores resultados, considerando que gerou menor diferença entre os valores atuais e os preditos.

O *Root Mean Squared Error* (Erro Quadrado Médio) da base foi de 0,3278 no primeiro experimento e 0,3485 no segundo. Considerando esse parâmetro o primeiro experimento também apresentou melhores resultados, pois apresentou menor erro entre os valores atuais e os valores preditos.

O *Relative Absolute Error* (Erro Absoluto Relativo) da base foi de 54% no primeiro experimento e 74% no segundo. É possível verificar que o primeiro experimento apresentou resultados mais eficazes, considerando que obteve mais precisão para previsão numérica do que o primeiro experimento, contudo ambos os resultados apresentaram um erro absoluto relativo elevado.

Por fim, o *Root Relative Squared Error* (Raiz do Erro Quadrado Relativo) da base foi de 82% no primeiro experimento e 88% no segundo, o que indica que o primeiro experimento também apresentou menor erro quando comparado com o segundo.

A Figura 41 demonstra com mais detalhes alguns valores referentes a acurácia da base de dados no primeiro experimento realizado:

Figura 41 - Acurácia PART - Experimento 1

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,363	0,141	0,380	0,363	0,371	0,226	0,712	0,371	EXTINCAO_SEM_MERITO
	0,323	0,034	0,476	0,323	0,385	0,345	0,715	0,275	ACORDO
	0,863	0,101	0,740	0,863	0,797	0,726	0,927	0,784	CONDENACAO
	0,601	0,136	0,543	0,601	0,571	0,449	0,825	0,473	IMPROCEDENCIA
	0,746	0,052	0,833	0,746	0,787	0,720	0,901	0,824	OUTROS
Weighted Avg.	0,633	0,097	0,630	0,633	0,628	0,536	0,839	0,604	

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Considerando a quantidade de atributos meta da base de dados, optou-se por realizar a análise considerando os principais extremos: melhor e pior resultado.

Os valores obtidos no *True Positive* (TP) indicam uma probabilidade mais alta de um verdadeiro positivo na proporção de 86% em Condenação. Com isso, é possível identificar que o padrão detectado acerta mais vezes nos casos de Condenação do que em Acordo, que apresentou o menor TP representado por 32%

Os resultados obtidos em *False Positive* (FP) indicam os dados classificados erroneamente como positivos, sendo a maior probabilidade encontrada em Extinção sem mérito 14% e a menor em Acordo 0,3%.

A precisão indica o valor da predição positiva (número de casos positivos por total de casos cobertos). Os resultados foram melhores para a classificação de Outros apresentando 83% de precisão, enquanto Extinção sem Mérito apresentou somente 38%.

O resultado do *recall* (cobertura) é equivalente aos valores obtidos no verdadeiro positivo (TP), considerando que corresponde ao valor da cobertura de casos.

Os valores obtidos no *F-Measure* indicam um desempenho de 79% para a classificação de Condenação e de 36% para a classificação de Extinção sem Mérito.

Os resultados obtidos em MCC demonstram um desempenho de 72% para a classificação de Condenação e de 22% para a classificação de Extinção Sem Mérito.

O *ROC Area* do classificador Condenação ficou em 92%, representando um bom resultado, considerando que um classificador dito como ótimo terá valores próximos a 1. A pior classificação representada por 71% correspondeu a Extinção Sem Mérito.

Por fim, o *PRC Area* do classificador Outros ficou em 82%, enquanto a pior classificação foi Acordo com 27%.

A Figura 42 demonstra com mais detalhes alguns valores referentes a acurácia da base de dados no segundo experimento realizado:

Figura 42 - Acurácia PART - Experimento 2

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,378	0,177	0,338	0,378	0,357	0,193	0,695	0,341	EXTINCAO_SEM_MERITO
	0,323	0,045	0,408	0,323	0,360	0,309	0,736	0,258	ACORDO
	0,720	0,272	0,468	0,720	0,568	0,399	0,762	0,429	CONDENACAO
	0,182	0,116	0,297	0,182	0,226	0,081	0,630	0,274	IMPROCEDENCIA
	0,691	0,031	0,887	0,691	0,776	0,720	0,865	0,814	OUTROS
Weighted Avg.	0,498	0,138	0,510	0,498	0,490	0,367	0,746	0,464	

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Os valores obtidos no *True Positive* (TP) indicam uma probabilidade mais alta de um verdadeiro positivo na proporção de 72% em Condenação. Com isso, é possível identificar que o padrão detectado acerta mais vezes nos casos de Condenação do que em Improcedência, que apresentou o menor TP representado por 18%

Os resultados obtidos em *False Positive* (FP) indicam os dados classificados erroneamente como positivos, sendo a maior probabilidade representada por 27% encontrada em Condenação e a menor em Outros, com apenas 0,3%.

A precisão indica o valor da predição positiva (número de casos positivos por total de casos cobertos). Os resultados foram melhores para a classificação de Outros apresentando 88% de precisão, enquanto Improcedência apresentou apenas 29%.

O resultado do *recall* (cobertura) é equivalente aos valores obtidos no verdadeiro positivo (TP), considerando que corresponde ao valor da cobertura de casos.

Os valores obtidos no *F-Measure* indicam um desempenho de 77% para a classificação de Outros e de 22% para a classificação de Improcedência.

Os resultados obtidos em MCC demonstram um desempenho de 72% para a classificação de Outros e de 0,8% para a classificação de Improcedência.

O *ROC Area* do classificador Outros ficou em 86%, representando um bom resultado, considerando que um classificador dito como ótimo terá valores próximos a 1. A pior classificação representada por 0,6% correspondeu a Improcedência

Por fim, o *PRC Area* do classificador Outros ficou em 81%, enquanto a pior classificação foi Acordo com 25%.

Comparando a média ponderada obtida em ambos os experimentos é possível identificar que o experimento 1 apresentou maior porcentagem de verdadeiros positivos (63%) e menor porcentagem de falsos positivos (0,9%), o que indica que classificou melhor os atributos. Além disso, o experimento 1 também se mostrou mais efetivo quanto a precisão, apresentando 12% a mais quando comparado ao experimento 2. Quanto ao F-*Measure* o experimento 1 obteve 13% a mais de desempenho do que o experimento 2. Por fim, o ROC Area dos experimentos apresentaram uma média próxima, tendo o experimento 1 alcançado 0,9% a mais que o experimento 1, realizando uma melhor classificação dos atributos. Considerando esses aspectos, o experimento 1 mostrou-se mais eficiente em todos os parâmetros de acurácia. Contudo, o experimento 2 gerou como contribuição e facilidade a simplificação para análise dos resultados.

A

Figura 43 apresenta a matriz de confusão obtida com o experimento 1. Possui como objetivo mostrar o número de previsões corretas em relação às esperadas para cada regra, sendo que na diagonal principal é possível identificar as instâncias que foram corretamente classificadas.

Figura 43 - Matriz de Confusão PART - Experimento 1

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  49   5  13  52  16 |   a = EXTINCAO_SEM_MERITO
   7  20  17  12   6 |   b = ACORDO
  13   5 151   4   2 |   c = CONDENACAO
  40  12   4  89   3 |   d = IMPROCEDENCIA
  20   0  19   7 135 |   e = OUTROS
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Para facilitar a análise dos resultados foi elaborada a Tabela 10 mostrando a porcentagem de atributos que foram classificados corretamente. Nesse experimento é possível identificar que o motivo de arquivamento Outros obteve melhores resultados, realizando a classificação correta de 83% dos atributos, seguido de Condenação (74%), Improcedência (54%), Acordo (48%) e Extinção sem Mérito (38%).

Tabela 10 - Matriz de Confusão PART - Experimento 1

a	B	c	d	E	<--	classified as
49	5	13	52	16	|	a = EXTINCAO_SEM_MERITO
7	20	17	12	6	|	b = ACORDO
13	5	151	4	2	|	c = CONDENACAO
40	12	4	89	3	|	d = IMPROCEDENCIA
20	0	19	7	135	|	e = OUTROS
129	**42**	**204**	**164**	**162**		**TOTAL DE INSTÂNCIAS**
38%	**48%**	**74%**	**54%**	**83%**		

FONTE: Elaborado pela autora (2016)

Já a Figura 44 apresenta a matriz de confusão obtida com o segundo experimento.

Figura 44 - Matriz de Confusão PART - Experimento 2

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  51  10  43  25   6 |   a = EXTINCAO_SEM_MERITO
  14  20  11  11   6 |   b = ACORDO
  24   3 126  18   4 |   c = CONDENACAO
  41  11  69  27   0 |   d = IMPROCEDENCIA
  21   5  20  10 125 |   e = OUTROS
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Para facilitar a análise dos resultados foi elaborada a Tabela 11 mostrando a porcentagem de atributos que foram classificados corretamente. É possível identificar que o motivo de arquivamento Outros também obteve melhores resultados no segundo experimento, seguido de Condenação (47%), Acordo (41%), Extinção Sem Mérito (34%) e Improcedência (30%).

Tabela 11 - Matriz de Confusão PART - Experimento 2

A	b	c	d	E	<--	classified as
51	10	43	25	6	|	a = EXTINCAO_SEM_MERITO
14	20	11	11	6	|	b = ACORDO
24	3	126	18	4	|	c = CONDENACAO
41	11	69	27	0	|	d = IMPROCEDENCIA
21	5	20	10	125	|	e = OUTROS
151	**49**	**269**	**91**	**141**		**TOTAL DE INSTÂNCIAS**
34%	**41%**	**47%**	**30%**	**89%**		

FONTE: Elaborado pela autora (2016)

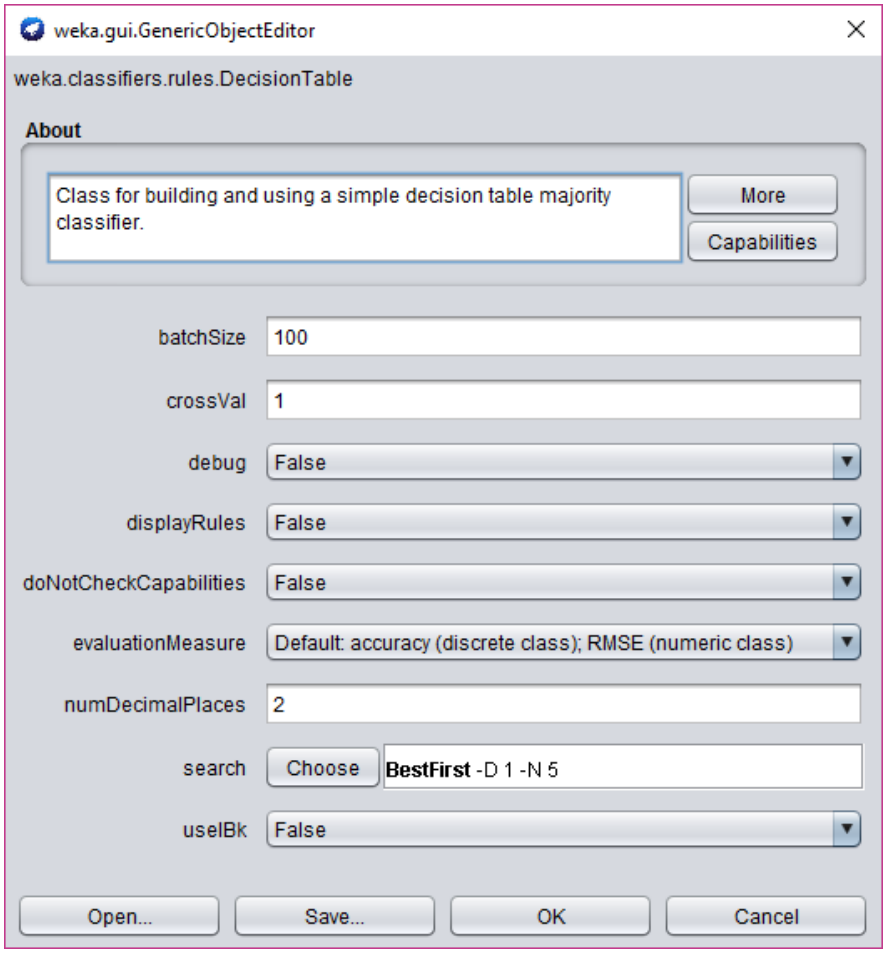
Ao comparar os resultados obtidos na matriz de confusão é possível identificar na categoria Outros o aumento de 5% nas instâncias classificadas corretamente no experimento 2. Em Improcedência o experimento 1 classificou 25% de instâncias a mais de que o experimento 2. Em condenação o experimento 1 apresentou melhores resultados, classificando corretamente 27% de instâncias a mais que o experimento 2. Em Acordo houve 7% a mais de instâncias classificadas corretamente no experimento 1 e 4% a mais em Extinção Sem Mérito.

Após concluir a análise dos resultados apresentados pela heurística PART, prosseguiu-se para a mineração de dados com outros algoritmos.

4.3.3 *Decision Table*

Na heurística de regras também foi utilizado o algoritmo Decision Table. O algoritmo é definido pelo software Weka como: “classe para a construção e utilização de uma simples tabela de decisão pela classificação da maioria. A Figura 45 apresenta a configuração dos parâmetros default:

Figura 45 - Parâmetros *Default Decision Table*



FONTE: Dados da pesquisa utilizando o software Weka (2016)

O primeiro experimento foi realizado alterando apenas o parâmetro "display*Rules*" para *True* para que nos resultados fosse apresentada a tabela de regras. O algoritmo demorou 0,17 segundos para ser executado e apresentou que os atributos ação, OJ, risco_atual são influenciadores para determinar o motivo_arquivamento. No total foram geradas 41 regras dispostas no Apêndice C. Contudo, como o atributo risco_atual não era interessante para a análise, foi realizado um novo experimento removendo todas as colunas que não exerceriam influência sobre o resultado almejado.

O segundo experimento foi realizado considerando apenas os atributos principais. Com isso, o tempo de execução reduziu para 0,01 segundos e foram encontradas 17 regras, conforme demonstra o Apêndice D.

Os resultados do Weka foram reorganizados em uma planilha, alterando a classificação do motivo_arquivamento para ordem alfabética para facilitar a análise, conforme demonstra o Quadro 14.

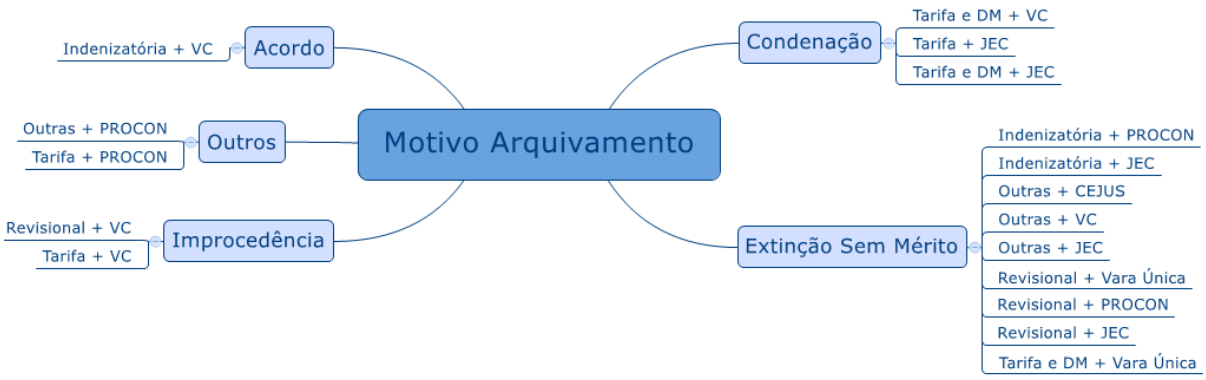
Quadro 14 - Resultado algoritmo *Decision Table*

MOTIVO_ARQUIVAMENTO	ACAO	OJ
ACORDO	INDENIZATORIA	VC
CONDENACAO	TARIFA	JEC
	TARIFA E DANO MORAL	VC
	TARIFA E DANO MORAL	JEC
EXTINCAO_SEM_MERITO	INDENIZATORIA	PROCON
	INDENIZATORIA	JEC
	OUTRAS	CEJUSC
	OUTRAS	VC
	OUTRAS	JEC
	REVISIONAL	VARA_UNICA
	REVISIONAL	PROCON
	REVISIONAL	JEC
	TARIFA E DANO MORAL	VARA_UNICA
IMPROCEDENCIA	REVISIONAL	VC
	TARIFA	VC
OUTROS	OUTRAS	PROCON
	TARIFA	PROCON

FONTE: Elaborado pela autora (2016)

Com base nos resultados apresentados foi criado um mapa conceitual, visando otimizar a visualização dos resultados, conforme demonstra a Figura 46:

Figura 46 - Mapa Conceitual algoritmo *Decision Table*



FONTE: Elaborado pela autora (2016)

Conforme demonstram os resultados, a classificação dos processos com Extinção Sem Mérito ainda são os que apresentam maior dificuldade no reconhecimento de padrões, encontrando quase todos os tipos de ações e órgãos julgadores. Contudo, ainda assim os resultados mostraram-se satisfatórios, pois

permitiram a identificação de padrões. No entanto, seria mais interessante se fosse apresentada na tabela a relação entre a região, o que não foi realizado automaticamente pela determinação da importância dos atributos pelo algoritmo.

O Quadro 15 demonstra resumidamente a comparação entre a precisão que os métodos foram capazes de prever com os experimentos:

Quadro 15 - Comparação dos resultados dos experimentos com aplicação do algoritmo *Decision Table*

	Experimento 1		**Experimento 2**	
Correctly Classified Instances	439	62,6248 %	361	49,786 %
Incorrectly Classified Instances	262	37,3752 %	340	50,214 %
Kappa Statistic	0,5173		0,3744	
Mean absolute error	0,1989		0,2427	
Root mean squared error	0,3101		0,3472	
Relative absolute error	63,6192 %		77,6157 %	
Root relative squared error	78,4263 %		87,8149 %	
Total Number of Instances	701		701	

FONTE: Elaborado pela autora (2016)

A partir dos resultados é possível identificar a porcentagem de instâncias que foram classificadas correta ou incorretamente. Do total de instâncias (701), no primeiro experimento 439 foram classificadas corretamente e 262 incorretamente. Já no segundo experimento 361 foram classificadas corretamente e 340 incorretamente. Com base nesse parâmetro, o experimento 1 apresentou resultados mais satisfatórios, pois teve maior porcentagem de assertividade na classificação das instâncias.

O *Kappa Statistic* (Estatística Kappa) indica o grau de concordância, sendo obtido o resultado de 0,5173 no primero experimento e 0,3744 no segundo. O *Kappa Statistic* (Estatística Kappa) indica o grau de concordância, sendo obtido o resultado de 0,5285 no primeiro experimento e 0,3539 no segundo. Apesar do primeiro experimento apresentar estatística superior quando comparado com o segundo, o resultado é inconclusivo, considerando que os valores ficaram na metade do intervalo entre 0 e 1, não permitindo afirmar que existe correlação.

O *Mean Absolute Error (*Erro Médio Absoluto) foi de 0,1989 no primeiro experimento e 0,2427 no segundo. Considerando esse parâmetro o primeiro experimento apresentou melhores resultados, considerando que gerou menor diferença entre os valores atuais e os preditos.

O *Root Mean Squared Error* (Erro Quadrado Médio) da base foi de 0,3101 no primeiro experimento e 0,3472 no segundo. Considerando esse parâmetro o primeiro experimento também apresentou melhores resultados, pois apresentou menor erro entre os valores atuais e os valores preditos.

O *Relative Absolute Error* (Erro Absoluto Relativo) da base foi de 63% no primeiro experimento e 77% no segundo. É possível verificar que o primeiro experimento apresentou resultados mais eficazes, considerando que obteve mais precisão para previsão numérica do que o primeiro experimento, contudo ambos os resultados apresentaram um erro absoluto relativo elevado.

Por fim, o *Root Relative Squared Error* (Raiz do Erro Quadrado Relativo) da base foi de 78% no primeiro experimento e 87% no segundo, o que indica que o primeiro experimento também apresentou menor erro quando comparado com o segundo.

A Figura 47 demonstra com mais detalhes alguns valores referentes a acurácia da base de dados no primeiro experimento realizado:

Figura 47 - Acurácia *Decision Table* - Experimento 1

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,378    0,141    0,389      0,378   0,383      0,239    0,776     0,367     EXTINCAO_SEM_MERITO
               0,129    0,020    0,381      0,129   0,193      0,181    0,759     0,260     ACORDO
               0,886    0,105    0,738      0,886   0,805      0,738    0,945     0,780     CONDENACAO
               0,649    0,166    0,511      0,649   0,571      0,444    0,854     0,512     IMPROCEDENCIA
               0,713    0,042    0,854      0,713   0,777      0,714    0,903     0,854     OUTROS
Weighted Avg.  0,626    0,101    0,621      0,626   0,613      0,524    0,866     0,617
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Considerando a quantidade de atributos meta da base de dados, optou-se por realizar a análise considerando os principais extremos: melhor e pior resultado.

Os valores obtidos no *True Positive* (TP) indicam uma probabilidade mais alta de um verdadeiro positivo na proporção de 88% em Condenação. Com isso, é possível identificar que o padrão detectado acerta mais vezes nos casos de Condenação do que em Acordo, que apresentou o menor TP representado por 12%

Os resultados obtidos em *False Positive* (FP) indicam os dados classificados erroneamente como positivos, sendo a maior probabilidade encontrada em Improcedência com 16% e a menor em Acordo com 0,2%.

A precisão indica o valor da predição positiva (número de casos positivos por total de casos cobertos). Os resultados foram melhores para a classificação de Outros apresentando 85% de precisão, enquanto Acordo apresentou somente 38%.

O resultado do *recall* (cobertura) é equivalente aos valores obtidos no verdadeiro positivo (TP), considerando que corresponde ao valor da cobertura de casos.

Os valores obtidos no *F-Measure* indicam um desempenho de 80% para a classificação de Condenação e de 19% para a classificação de Acordo.

Os resultados obtidos em MCC demonstram um desempenho de 73% para a classificação de Condenação e de 18% para a classificação de Acordo.

O *ROC Area* do classificador Condenação ficou em 95%, representando um bom resultado, considerando que um classificador dito como ótimo terá valores próximos a 1. A pior classificação representada por 75% correspondeu a Acordo.

Por fim, o *PRC Area* do classificador Outros ficou em 85%, enquanto a pior classificação foi Acordo com 26%.

A Figura 48 demonstra com mais detalhes alguns valores referentes a acurácia da base de dados no segundo experimento realizado:

Figura 48 - Acurácia *Decision Table* - Experimento 2

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,378	0,177	0,338	0,378	0,357	0,193	0,695	0,341	EXTINCAO_SEM_MERITO
	0,323	0,045	0,408	0,323	0,360	0,309	0,736	0,258	ACORDO
	0,720	0,272	0,468	0,720	0,568	0,399	0,762	0,429	CONDENACAO
	0,182	0,116	0,297	0,182	0,226	0,081	0,630	0,274	IMPROCEDENCIA
	0,691	0,031	0,887	0,691	0,776	0,720	0,865	0,814	OUTROS
Weighted Avg.	0,498	0,138	0,510	0,498	0,490	0,367	0,746	0,464	

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Os valores obtidos no *True Positive* (TP) indicam uma probabilidade mais alta de um verdadeiro positivo na proporção de 69% em Condenação. Com isso, é possível identificar que o padrão detectado acerta mais vezes nos casos de Condenação do que em Acordo, que apresentou o menor TP representado por 11%

Os resultados obtidos em *False Positive* (FP) indicam os dados classificados erroneamente como positivos, sendo a maior probabilidade representada por 24% encontrada em Condenação e a menor em Outros, com apenas 0,6%.

A precisão indica o valor da predição positiva (número de casos positivos por total de casos cobertos). Os resultados foram melhores para a classificação de Outros apresentando 97% de precisão, enquanto Improcedência apresentou apenas 32%.

O resultado do *recall* (cobertura) é equivalente aos valores obtidos no verdadeiro positivo (TP), considerando que corresponde ao valor da cobertura de casos.

Os valores obtidos no *F-Measure* indicam um desempenho de 77% para a classificação de Outros e de 22% para a classificação de Improcedência.

Por fim, o *ROC Area* do classificador Outros ficou em 80%, representando um bom resultado, considerando que um classificador dito como ótimo terá valores próximos a 1. A pior classificação representada por 16% correspondeu a Acordo.

Comparando a média ponderada obtida em ambos os experimentos é possível identificar que o experimento 1 apresentou 11% a mais de verdadeiros positivos e 3% a menos de falsos positivos, o que indica que classificou melhor os atributos. Além disso, o experimento 1 também se mostrou mais efetivo quanto a precisão, apresentando 8% a mais quando comparado ao experimento 2. Quanto ao F-Measure, o experimento 1 obteve 10% a mais de desempenho do que o experimento 2 e no MCC 13%. Por fim, no ROC Area o experimento 1 alcançou 12% a mais que o experimento 1 e no PRC Área 16% a mais, realizando uma melhor classificação dos atributos. Considerando esses aspectos, o experimento 1 mostrou-se mais eficiente em todos os parâmetros de acurácia. Contudo, o experimento 2 gerou como contribuição e facilidade a simplificação para análise dos resultados, reduzindo o número de regras.

A Figura 49 apresenta a matriz de confusão obtida com o experimento 1, demonstrando na diagonal principal as instâncias que foram corretamente classificadas.

Figura 49 - Matriz de Confusão *Decision Table* - Experimento 1

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  51   1  13  58  12 |   a = EXTINCAO_SEM_MERITO
  10   8  17  23   4 |   b = ACORDO
  11   2 155   3   4 |   c = CONDENACAO
  38   8   4  96   2 |   d = IMPROCEDENCIA
  21   2  21   8 129 |   e = OUTROS
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Para facilitar a análise dos resultados foi elaborada a Tabela 12 mostrando a porcentagem de atributos que foram classificados corretamente. Nesse experimento é possível identificar que o motivo de arquivamento Outros obteve melhores resultados, realizando a classificação correta de 85% dos atributos, seguido de Condenação (74%), Improcedência (51%), Extinção sem Mérito (39%) e Acordo (38%).

Tabela 12 - Matriz de Confusão *Decision Table* - Experimento 1

a	b	C	d	E	<--	classified as
51	1	13	58	12	|	a = EXTINCAO_SEM_MERITO
10	8	17	23	4	|	b = ACORDO
11	2	155	3	4	|	c = CONDENACAO
38	8	4	96	2	|	d = IMPROCEDENCIA
21	2	21	8	129	|	e = OUTROS
131	21	210	188	151		**TOTAL DE INSTÂNCIAS**
39%	**38%**	**74%**	**51%**	**85%**		

FONTE: Elaborado pela autora (2016)

Já a Figura 50 apresenta a matriz de confusão obtida com o segundo experimento.

Figura 50 - Matriz de Confusão *Decision Table* - Experimento 2

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  51  10  43  25   6 |   a = EXTINCAO_SEM_MERITO
  14  20  11  11   6 |   b = ACORDO
  24   3 126  18   4 |   c = CONDENACAO
  41  11  69  27   0 |   d = IMPROCEDENCIA
  21   5  20  10 125 |   e = OUTROS
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Para facilitar a análise dos resultados foi elaborada a

Tabela 13 mostrando a porcentagem de atributos que foram classificados corretamente. É possível identificar que o motivo de arquivamento Outros também obteve melhores resultados no segundo experimento, seguido de Condenação (49%), Extinção Sem Mérito (37%), Acordo (33%), e Improcedência (32%).

Tabela 13 - Matriz de Confusão *Decision Table* - Experimento 2

a	b	C	D	e	<--	classified as
70	3	39	23	0	|	a = EXTINCAO_SEM_MERITO
15	7	8	30	2	|	b = ACORDO
34	4	122	15	0	|	c = CONDENACAO
46	2	62	37	1	|	d = IMPROCEDENCIA
24	5	17	10	125	|	e = OUTROS
189	21	248	115	128		**TOTAL**
37%	**33%**	**49%**	**32%**	**98%**		

FONTE: Elaborado pela autora (2016)

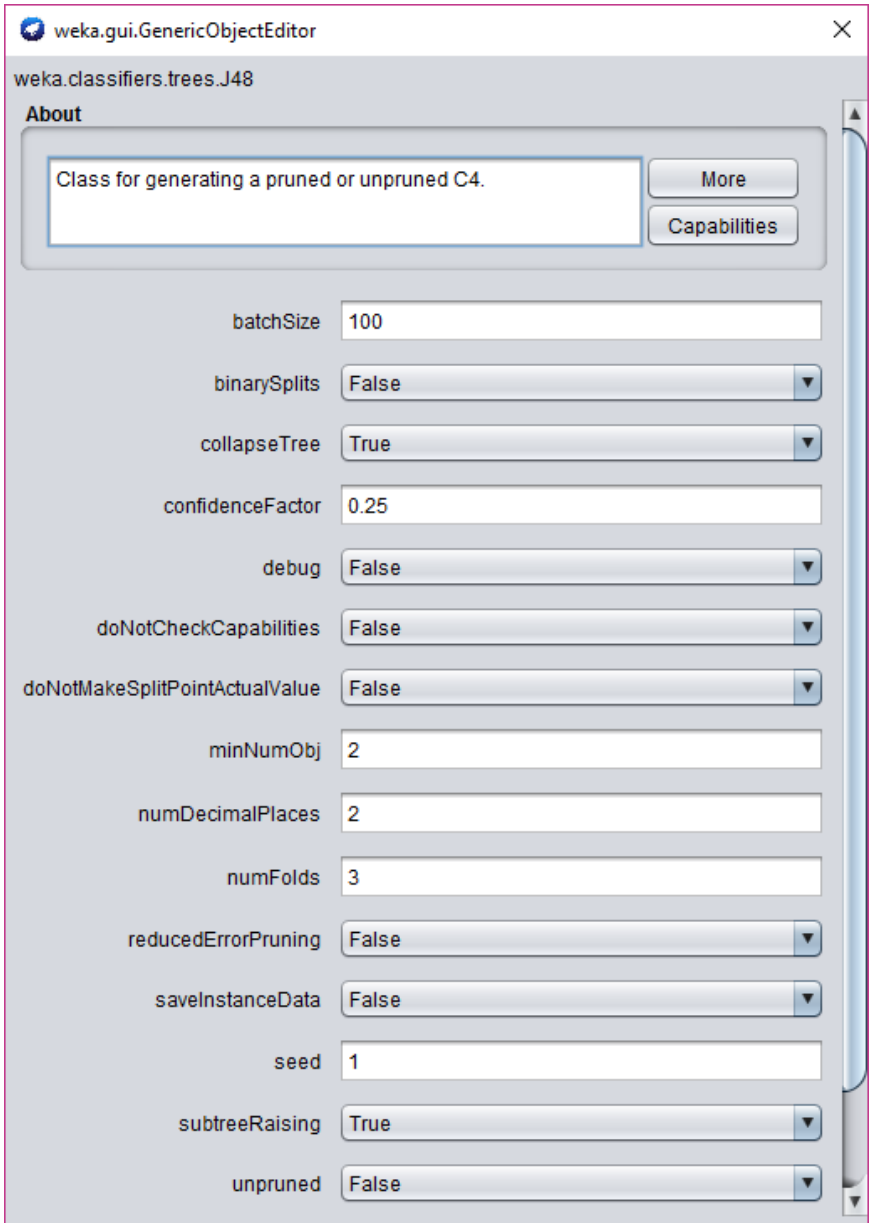
Ao comparar os resultados obtidos na matriz de confusão é possível identificar na categoria Outros o aumento de 12% nas instâncias classificadas corretamente no experimento 2. Em Improcedência o experimento 1 classificou 19% de instâncias a mais de que o experimento 2. Em condenação o experimento 1 apresentou melhores resultados, classificando corretamente 25% de instâncias a mais que o experimento 2. Em Acordo houve 5% a mais de instâncias classificadas corretamente no experimento 1 e 2% a mais em Extinção Sem Mérito.

Após concluir a análise dos resultados apresentados pela heurística *Decision Table*, prosseguiu-se para a mineração de dados com outros algoritmos.

4.3.4 J48

Na heurística de árvores foi executado o algoritmo J48, o qual é definido pelo software Weka como: “Classe para gerar uma árvore de decisão C4.5 podada ou não podada”. Apresenta como parâmetros default os itens dispostos na Figura 51.

Figura 51 - Parâmetros *Default* J48



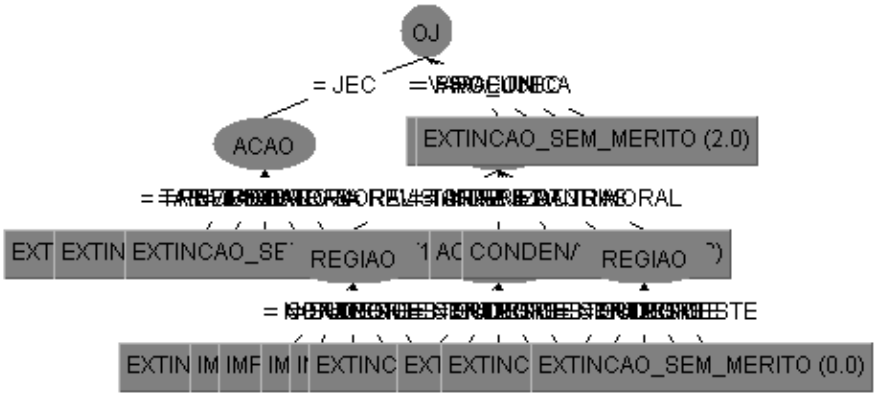
FONTE: Dados da pesquisa utilizando o software Weka (2016)

Ao executar o algoritmo J48 com os parâmetros *default* foi gerada uma árvore com 69 folhas e tamanho 94 em 0 segundos. Porém devido ao seu tamanho, as folhas ficaram sobrepostas, impossibilitando a análise.

Para facilitar a análise e compreensão dos dados foi realizado um novo experimento conforme aplicado anteriormente nos algoritmos PART e *Decision Table* considerando apenas os quatro atributos principais. Na nova execução o algoritmo também demorou 0 segundos, porém reduziu o tamanho da árvore para 31 com 25 folhas. A árvore de decisão gerada encontra-se disposta na Figura 52. No entanto, o

software não permite a edição da árvore gerada, mantendo também as folhas sobrepostas. Dessa forma, para facilitar a análise foi gerada manualmente a árvore de decisão com base nos resultados apresentados no Weka. O resultado encontra-se disposto no Apêndice E.

Figura 52 - Árvore de Decisão J48 - Experimento 2



FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Pela leitura da árvore de decisão é possível identificar que a raiz da árvore corresponde ao Órgão Julgador, sendo este, portanto, o atributo com maior influência. O segundo atributo com maior influência corresponde ao tipo de ação, seguido de região. Para PROCON, Vara Única e Cejusc o resultado foi simplificado, mostrando diretamente o motivo do arquivamento. Já pra VC e JEC existem outros atributos que exercem influência sobre o motivo do arquivamento.

No PROCON foi identificado o motivo de arquivamento Outros, enquanto na Vara Única Improcedência e no Cejusc Extinção Sem Mérito. No JEC o tipo de ação exerce influência sobre o resultado, sendo para ações revisionais, indenizatórias e outras obtido Extinção Sem Mérito e para Tarifas e Tarifa e Dano Moral Condenação. Por fim, na Vara Cível (VC), além do tipo de ação, a região também exerce influência sobre o resultado das ações revisionais, tarifas e outras. Em ações indenizatórias foi obtido como resultado Acordo e em ações de tarifa e dano moral Condenação. Em ações revisionais que tramitam no sul foi obtido Acordo, no nordeste Extinção Sem Mérito, no sudeste, centro-oeste e norte Improcedência. Em ações de tarifa que tramitam no sul foi obtido Improcedência, no nordeste Condenação, no Sudeste Extinção sem Mérito, no centro-oeste Outros e no norte Condenação.

O Quadro 16 demonstra resumidamente a comparação entre a precisão que os métodos foram capazes de prever com os experimentos:

Quadro 16 - Comparação dos resultados dos experimentos com aplicação do algoritmo J48

	Experimento 1		**Experimento 2**	
Correctly Classified Instances	465	66,3338 %	354	5,4993%
Incorrectly Classified Instances	236	33,6662 %	347	49,5007 %
Kappa Statistic	0,5678		0,3649	
Mean absolute error	0,1714		0,233	
Root mean squared error	0,3118		0,3511	
Relative absolute error	54,8111 %		74,514 %	
Root relative squared error	78,8722 %		88,8055 %	
Total Number of Instances	701		701	

FONTE: Elaborado pela autora (2016)

A partir dos resultados é possível identificar a porcentagem de instâncias que foram classificadas correta ou incorretamente. Do total de instâncias (701), no primeiro experimento 465 foram classificadas corretamente e 236 incorretamente. Já no segundo experimento 354 foram classificadas corretamente e 347 incorretamente. Com base nesse parâmetro, o experimento 1 apresentou resultados mais satisfatórios, pois teve maior porcentagem de assertividade na classificação das instâncias.

O *Kappa Statistic* (Estatística Kappa) indica o grau de concordância, sendo obtido o resultado de 0,5678 no primero experimento e 0,3649 no segundo. O *Kappa Statistic* (Estatística Kappa) indica o grau de concordância, sendo obtido o resultado de 0,5285 no primeiro experimento e 0,3539 no segundo. Apesar do primeiro experimento apresentar valor superior quando comparado com o segundo, o resultado é inconclusivo, considerando que ficaram na metade do intervalo entre 0 e 1, não permitindo afirmar que existe correlação.

O *Mean Absolute Error (*Erro Médio Absoluto) foi de 0,1714 no primeiro experimento e 0,233 no segundo. Considerando esse parâmetro o primeiro experimento apresentou melhores resultados, considerando que gerou menor diferença entre os valores atuais e os preditos.

O *Root Mean Squared Error* (Erro Quadrado Médio) da base foi de 0,3118 no primeiro experimento e 0,3511no segundo. Considerando esse parâmetro o primeiro experimento também apresentou melhores resultados, pois apresentou menor erro entre os valores atuais e os valores preditos.

O *Relative Absolute Error* (Erro Absoluto Relativo) da base foi de 54% no primeiro experimento e 74% no segundo. É possível verificar que o primeiro experimento apresentou resultados mais eficazes, considerando que obteve mais precisão para previsão numérica do que o primeiro experimento, contudo ambos os resultados apresentaram um erro absoluto relativo elevado.

Por fim, o *Root Relative Squared Error* (Raiz do Erro Quadrado Relativo) da base foi de 78% no primeiro experimento e 88% no segundo, o que indica que o primeiro experimento também apresentou menor erro quando comparado com o segundo.

A Figura 53 demonstra com mais detalhes alguns valores referentes a acurácia da base de dados no primeiro experimento realizado:

Figura 53 - Acurácia J48 - Experimento 1

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,430    0,136    0,430      0,430    0,430      0,294    0,748     0,404     EXTINCAO_SEM_MERITO
                 0,306    0,042    0,413      0,306    0,352      0,303    0,695     0,260     ACORDO
                 0,903    0,089    0,771      0,903    0,832      0,774    0,915     0,695     CONDENACAO
                 0,669    0,116    0,607      0,669    0,637      0,534    0,858     0,561     IMPROCEDENCIA
                 0,724    0,040    0,862      0,724    0,787      0,726    0,892     0,796     OUTROS
Weighted Avg.    0,663    0,087    0,662      0,663    0,659      0,577    0,845     0,598
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Considerando a quantidade de atributos meta da base de dados, optou-se por realizar a análise considerando os principais extremos: melhor e pior resultado.

Os valores obtidos no *True Positive* (TP) indicam uma probabilidade mais alta de um verdadeiro positivo na proporção de 90% em Condenação. Com isso, é possível identificar que o padrão detectado acerta mais vezes nos casos de Condenação do que em Acordo, que apresentou o menor TP representado por 30%

Os resultados obtidos em *False Positive* (FP) indicam os dados classificados erroneamente como positivos, sendo a maior probabilidade encontrada em Extinção Sem Mérito com 13% e a menor em Outros com 0,4%.

A precisão indica o valor da predição positiva (número de casos positivos por total de casos cobertos). Os resultados foram melhores para a classificação de Outros apresentando 86% de precisão, enquanto Acordo apresentou somente 41%.

O resultado do *recall* (cobertura) é equivalente aos valores obtidos no verdadeiro positivo (TP), considerando que corresponde ao valor da cobertura de casos.

Os valores obtidos no *F-Measure* indicam um desempenho de 83% para a classificação de Condenação e de 35% para a classificação de Acordo.

Os resultados obtidos em MCC demonstram um desempenho de 77% para a classificação de Condenação e de 29% para a classificação de Extinção Sem Mérito.

O *ROC Area* do classificador Condenação ficou em 91%, representando um bom resultado, considerando que um classificador dito como ótimo terá valores próximos a 1. A pior classificação representada por 69% correspondeu a Acordo.

Por fim, o *PRC Area* do classificador Outros ficou em 79%, enquanto a pior classificação foi Acordo com 0,2%.

A Figura 54 demonstra com mais detalhes alguns valores referentes a acurácia da base de dados no segundo experimento realizado:

Figura 54 - Acurácia J48 - Experimento 2

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,452    0,200    0,351      0,452    0,395      0,230    0,678     0,331     EXTINCAO_SEM_MERITO
                 0,306    0,053    0,358      0,306    0,330      0,272    0,708     0,232     ACORDO
                 0,709    0,255    0,481      0,709    0,573      0,407    0,761     0,423     CONDENACAO
                 0,155    0,101    0,291      0,155    0,203      0,070    0,645     0,283     IMPROCEDENCIA
                 0,702    0,019    0,927      0,702    0,799      0,753    0,878     0,812     OUTROS
Weighted Avg.    0,505    0,133    0,520      0,505    0,497      0,379    0,746     0,459
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Os valores obtidos no *True Positive* (TP) indicam uma probabilidade mais alta de um verdadeiro positivo na proporção de 70% em Condenação. Com isso, é possível identificar que o padrão detectado acerta mais vezes nos casos de Condenação do que em Improcedência, que apresentou o menor TP representado por 15%

Os resultados obtidos em *False Positive* (FP) indicam os dados classificados erroneamente como positivos, sendo a maior probabilidade representada por 25% encontrada em Condenação e a menor em Outros, com apenas 0,1%.

A precisão indica o valor da predição positiva (número de casos positivos por total de casos cobertos). Os resultados foram melhores para a classificação de Outros apresentando 92% de precisão, enquanto Improcedência apresentou apenas 29%.

O resultado do *recall* (cobertura) é equivalente aos valores obtidos no verdadeiro positivo (TP), considerando que corresponde ao valor da cobertura de casos.

Os valores obtidos no *F-Measure* indicam um desempenho de 79% para a classificação de Outros e de 20% para a classificação de Improcedência.

Os resultados obtidos em MCC demonstram um desempenho de 75% para a classificação de Condenação e de 0,07% para a classificação de Improcedêcia.

O *ROC Area* do classificador Outros ficou em 87%, representando um bom resultado, considerando que um classificador dito como ótimo terá valores próximos a 1. A pior classificação representada por 0,7% correspondeu a Improcedência.

Por fim, o *PRC Area* do classificador Outros ficou em 82%, enquanto a pior classificação foi Acordo com 0,2%.

Comparando a média ponderada obtida em ambos os experimentos é possível identificar que o experimento 1 apresentou 16% a mais de verdadeiros positivos e 5% a menos de falsos positivos, o que indica que classificou melhor os atributos. Além disso, o experimento 1 também se mostrou mais efetivo quanto a precisão, apresentando 14% a mais quando comparado ao experimento 2. Quanto ao F-Measure, o experimento 1 obteve 16% a mais de desempenho do que o experimento 2 e no MCC 20%. Por fim, no ROC Area o experimento 1 alcançou 10% a mais que o experimento 1 e no PRC Área 14% a mais, realizando uma melhor classificação dos atributos. Considerando esses aspectos, o experimento 1 mostrou-se mais eficiente em todos os parâmetros de acurácia. Contudo, o experimento 2 gerou como contribuição e facilidade a simplificação para análise dos resultados, permitindo visualizar a árvore de decisão.

A Figura 55 apresenta a matriz de confusão obtida com o experimento 1, demonstrando na diagonal principal as instâncias que foram corretamente classificadas.

Figura 55 - Matriz de Confusão J48 - Experimento 1

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  51   1  13  58  12 |   a = EXTINCAO_SEM_MERITO
  10   8  17  23   4 |   b = ACORDO
  11   2 155   3   4 |   c = CONDENACAO
  38   8   4  96   2 |   d = IMPROCEDENCIA
  21   2  21   8 129 |   e = OUTROS
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Para facilitar a análise dos resultados foi elaborada a Tabela 14 mostrando a porcentagem de atributos que foram classificados corretamente. Nesse experimento é possível identificar que o motivo de arquivamento Outros obteve melhores resultados, realizando a classificação correta de 86% dos atributos, seguido de Condenação (77%), Improcedência (61%), Extinção sem Mérito (43%) e Acordo (41%).

Tabela 14 - Matriz de Confusão J48 - Experimento 1

a	b	C	d	e	<--	classified as
58	11	6	45	15	|	a = EXTINCAO_SEM_MERITO
9	19	20	11	3	|	b = ACORDO
14	0	158	1	2	|	c = CONDENACAO
29	15	4	99	1	|	d = IMPROCEDENCIA
25	1	17	7	131	|	e = OUTROS
135	46	205	163	152		**TOTAL DE INSTÂNCIAS**
43%	**41%**	**77%**	**61%**	**86%**		

FONTE: Elaborado pela autora (2016)

Já a Figura 56 apresenta a matriz de confusão obtida com o segundo experimento.

Figura 56 - Matriz de Confusão J48 - Experimento 2

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  61  10  41  20   3 |   a = EXTINCAO_SEM_MERITO
  15  19  12  11   5 |   b = ACORDO
  29   5 124  17   0 |   c = CONDENACAO
  45  14  64  23   2 |   d = IMPROCEDENCIA
  24   5  17   8 127 |   e = OUTROS
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Para facilitar a análise dos resultados foi elaborada a Tabela 15Tabela 15 mostrando a porcentagem de atributos que foram classificados corretamente. É possível identificar que o motivo de arquivamento Outros também obteve melhores resultados no segundo experimento, realizando a classificação correta de 93% dos atributos, seguido de Condenação (48%), Acordo (36%), Extinção Sem Mérito (35%), e Improcedência (29%).

Tabela 15 - Matriz de Confusão J48 - Experimento 2

a	b	C	D	e	<--	classified as
61	10	41	20	3	|	a = EXTINCAO_SEM_MERITO
15	19	12	11	5	|	b = ACORDO
29	5	124	17	0	|	c = CONDENACAO
45	14	64	23	2	|	d = IMPROCEDENCIA
24	5	17	8	127	|	e = OUTROS
174	53	258	79	137		**TOTAL**
35%	**36%**	**48%**	**29%**	**93%**		

FONTE: Elaborado pela autora (2016)

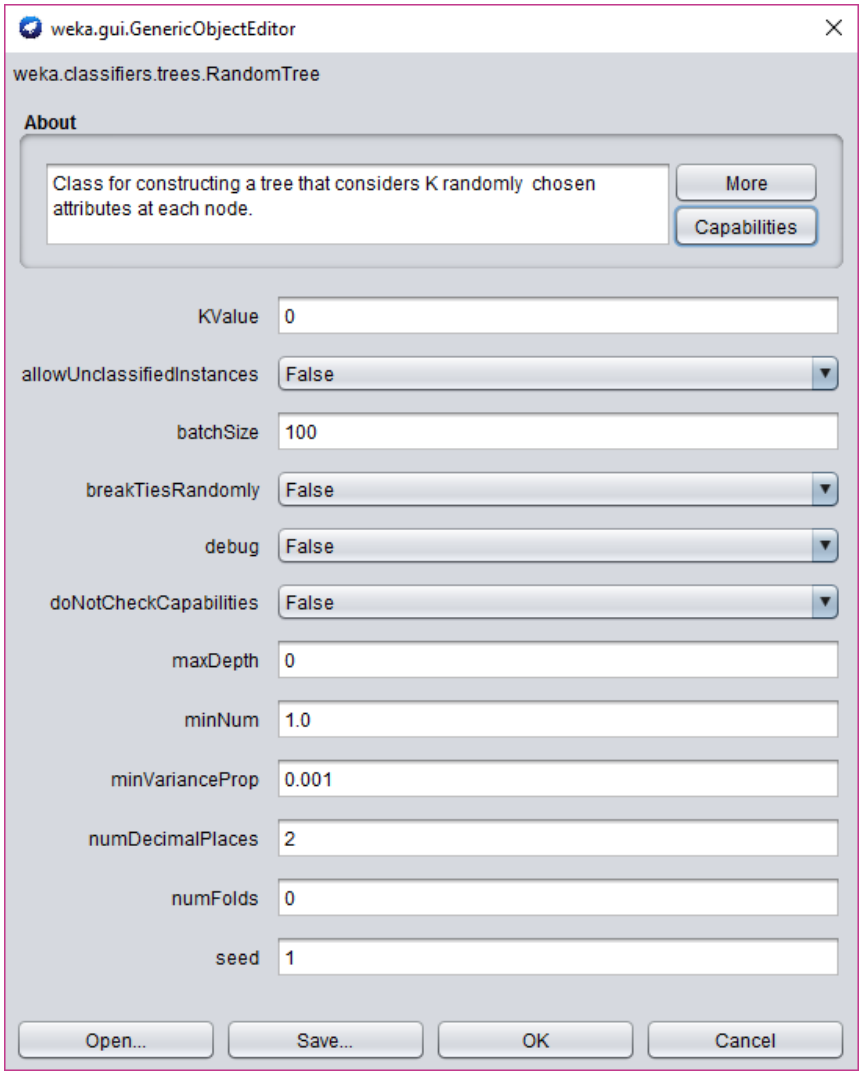
Ao comparar os resultados obtidos na matriz de confusão é possível identificar na categoria Outros houve o aumento de 7% nas instâncias classificadas corretamente no experimento 2. Em Improcedência o experimento 1 classificou 32% de instâncias a mais de que o experimento 2. Em condenação o experimento 1 apresentou melhores resultados, classificando corretamente 29% de instâncias a mais que o experimento 2. Em Acordo houve 5% a mais de instâncias classificadas corretamente no experimento 1 e 8% a mais em Extinção Sem Mérito.

Após concluir a análise dos resultados apresentados pela heurística J48 prosseguiu-se para a mineração de dados com outros algoritmos.

4.3.5 REPTree

Ainda na heurística de árvores foi executado o algoritmo REPTree, o qual é definido pelo software Weka como: “classe para a construção de uma árvore que considera atributos K escolhidos aleatoriamente em cada nó. Não executa nenhuma poda. Também tem uma opção para permitir a estimativa de probabilidades de classe”. Apresenta como parâmetros *default* os itens dispostos na Figura 57.

Figura 57 - Parâmetros *Default* REPTree



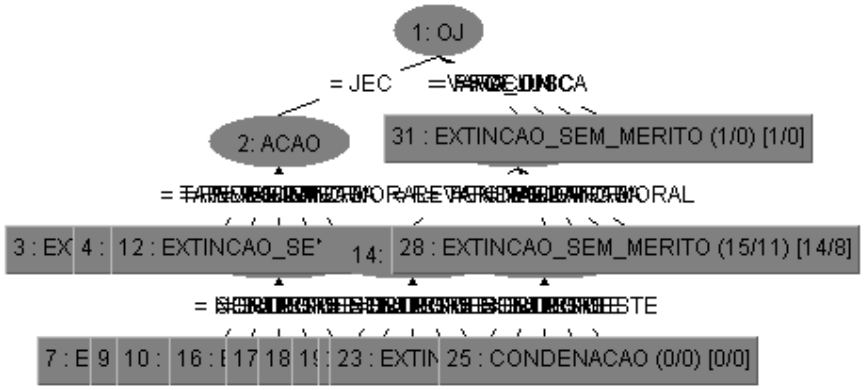
FONTE: Dados da pesquisa utilizando o software Weka (2016)

Ao executar o algoritmo REPTree com os parâmetros *default* foi gerada uma árvore com tamanho 43 (Apêndice F) em 0,4 segundos. Porém devido ao seu tamanho, as folhas ficaram sobrepostas, impossibilitando a análise.

Para facilitar a análise e compreensão dos dados foi realizado um novo experimento conforme aplicado anteriormente nos algoritmos PART, Decision Table e J48, considerando apenas os quatro principais atributos. Na nova execução o algoritmo demorou 0,01 segundos e reduziu o tamanho da árvore para 31 (Apêndice G). A árvore de decisão gerada encontra-se disposta na Figura 58. É possível verificar uma redução no tamanho, contudo ainda não é possível realizar a leitura. Dessa forma, para facilitar a análise foi gerada manualmente a árvore de decisão com base

nos resultados apresentados no Weka. O resultado encontra-se disposto no Apêndice H.

Figura 58 - Árvore de Decisão REPTree - Experimento 2



FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Pela leitura da árvore de decisão é possível identificar que a raiz da árvore corresponde ao Órgão Julgador, sendo este, portanto, o atributo com maior influência. O segundo atributo com maior influência corresponde ao tipo de ação, seguido de região. Para PROCON, Vara Única e Cejusc o resultado foi simplificado, mostrando diretamente o motivo do arquivamento. Já pra VC e JEC existem outros atributos que exercem influência sobre o motivo do arquivamento.

No PROCON foi identificado o motivo de arquivamento Outros, enquanto na Vara Única Improcedência e no Cejusc Extinção Sem Mérito. No JEC o tipo de ação e a região exercem influência sobre o resultado, sendo para ações revisionais e outras obtido Extinção Sem Mérito, para tarifas e tarifa e dano moral Condenação e para indenizatórias que tramitam no sul Condenação, no nordeste, centro-oeste e norte Extinção Sem Mérito e no sudeste Condenação. Por fim, na Vara Cível (VC) também recebeu influência do tipo de ação e da região nas ações revisionais e de tarifas. Em ações indenizatórias foi obtido como resultado Acordo, em ações de tarifa e dano moral Condenação e em Outras Extinção Sem Mérito. Em ações revisionais que tramitam no sul foi obtido Acordo, no nordeste Extinção Sem Mérito, no sudeste, centro-oeste e norte Improcedência. Em ações de tarifa que tramitam no sul foi obtido Improcedência, no nordeste Condenação, no Sudeste Extinção sem Mérito, no centro-oeste Outros e no norte Condenação.

O Quadro 17 demonstra resumidamente a comparação entre a precisão que os métodos foram capazes de prever com os experimentos:

Quadro 17 - Comparação dos resultados dos experimentos com aplicação do algoritmo REPTree

	Experimento 1		Experimento 2	
Correctly Classified Instances	461	65,7632 %	350	49,9287 %
Incorrectly Classified Instances	240	34,2368 %	351	50,0713 %
Kappa Statistic	0,5587		0,3564	
Mean absolute error	0,1776		0,2349	
Root mean squared error	0,3102		0,3509	
Relative absolute error	56,7864 %		75,1386 %	
Root relative squared error	78,4613 %		88,7495 %	
Total Number of Instances	701		701	

FONTE: Elaborado pela autora (2016)

A partir dos resultados é possível identificar a porcentagem de instâncias que foram classificadas correta ou incorretamente. Do total de instâncias (701), no primeiro experimento 461 foram classificadas corretamente e 240 incorretamente. Já no segundo experimento 350 foram classificadas corretamente e 351 incorretamente. Com base nesse parâmetro, o experimento 1 apresentou resultados mais satisfatórios, pois teve maior porcentagem de assertividade na classificação das instâncias.

O *Kappa Statistic* (Estatística Kappa) indica o grau de concordância, sendo obtido o resultado de 0,5587 no primero experimento e 0,3564 no segundo. Apesar do primeiro experimento apresentar estatística superior quando comparado com o segundo, o resultado é inconclusivo, considerando que os valores ficaram na metade do intervalo entre 0 e 1, não permitindo afirmar que existe correlação.

O *Mean Absolute Error (*Erro Médio Absoluto) foi de 0,1776 no primeiro experimento e 0,2349 no segundo. Considerando esse parâmetro o primeiro experimento apresentou melhores resultados, considerando que gerou menor diferença entre os valores atuais e os preditos.

O *Root Mean Squared Error* (Erro Quadrado Médio) da base foi de 0,3102 no primeiro experimento e 0,3509 no segundo. Considerando esse parâmetro o primeiro experimento também apresentou melhores resultados, pois apresentou menor erro entre os valores atuais e os valores preditos.

O *Relative Absolute Error* (Erro Absoluto Relativo) da base foi de 56% no primeiro experimento e 75% no segundo. É possível verificar que o primeiro experimento apresentou resultados mais eficazes, considerando que obteve mais precisão para previsão numérica do que o primeiro experimento, contudo ambos os resultados apresentaram um erro absoluto relativo elevado.

Por fim, o *Root Relative Squared Error* (Raiz do Erro Quadrado Relativo) da base foi de 78% no primeiro experimento e 88% no segundo, o que indica que o primeiro experimento também apresentou menor erro quando comparado com o segundo.

A Figura 59 demonstra com mais detalhes alguns valores referentes a acurácia da base de dados no primeiro experimento realizado:

Figura 59 - Acurácia REPTree - Experimento 1

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,444    0,124    0,462      0,444    0,453      0,325    0,789     0,418     EXTINCAO_SEM_MERITO
                 0,210    0,027    0,433      0,210    0,283      0,257    0,707     0,245     ACORDO
                 0,891    0,080    0,788      0,891    0,836      0,780    0,928     0,759     CONDENACAO
                 0,655    0,150    0,539      0,655    0,591      0,472    0,851     0,548     IMPROCEDENCIA
                 0,746    0,054    0,828      0,746    0,785      0,717    0,909     0,805     OUTROS
Weighted Avg.    0,658    0,092    0,652      0,658    0,649      0,565    0,861     0,615
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Considerando a quantidade de atributos meta da base de dados, optou-se por realizar a análise considerando os principais extremos: melhor e pior resultado.

Os valores obtidos no *True Positive* (TP) indicam uma probabilidade mais alta de um verdadeiro positivo na proporção de 89% em Condenação. Com isso, é possível identificar que o padrão detectado acerta mais vezes nos casos de Condenação do que em Acordo, que apresentou o menor TP representado por 21%

Os resultados obtidos em *False Positive* (FP) indicam os dados classificados erroneamente como positivos, sendo a maior probabilidade encontrada em Improcedência com 15% e a menor em Outros com 0,2%.

A precisão indica o valor da predição positiva (número de casos positivos por total de casos cobertos). Os resultados foram melhores para a classificação de Outros apresentando 74% de precisão, enquanto Acordo apresentou somente 21%.

O resultado do *recall* (cobertura) é equivalente aos valores obtidos no verdadeiro positivo (TP), considerando que corresponde ao valor da cobertura de casos.

Os valores obtidos no *F-Measure* indicam um desempenho de 83% para a classificação de Condenação e de 28% para a classificação de Acordo.

Os resultados obtidos em MCC demonstram um desempenho de 78% para a classificação de Condenação e de 25% para a classificação de Acordo.

O *ROC Area* do classificador Condenação ficou em 92%, representando um bom resultado, considerando que um classificador dito como ótimo terá valores próximos a 1. A pior classificação representada por 70% correspondeu a Acordo.

Por fim, o *PRC Area* do classificador Outros ficou em 80%, enquanto a pior classificação foi Acordo com 24%.

A Figura 60 demonstra com mais detalhes alguns valores referentes a acurácia da base de dados no segundo experimento realizado:

Figura 60 - Acurácia REPTree - Experimento 2

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,459    0,180    0,378      0,459    0,415      0,260    0,700     0,347     EXTINCAO_SEM_MERITO
                 0,258    0,050    0,333      0,258    0,291      0,234    0,712     0,227     ACORDO
                 0,714    0,276    0,463      0,714    0,562      0,390    0,761     0,427     CONDENACAO
                 0,155    0,119    0,258      0,155    0,194      0,044    0,642     0,288     IMPROCEDENCIA
                 0,685    0,012    0,954      0,685    0,797      0,758    0,857     0,798     OUTROS
Weighted Avg.    0,499    0,136    0,519      0,499    0,493      0,373    0,745     0,460
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Os valores obtidos no *True Positive* (TP) indicam uma probabilidade mais alta de um verdadeiro positivo na proporção de 71% em Condenação. Com isso, é possível identificar que o padrão detectado acerta mais vezes nos casos de Condenação do que em Improcedência, que apresentou o menor TP representado por 15%

Os resultados obtidos em *False Positive* (FP) indicam os dados classificados erroneamente como positivos, sendo a maior probabilidade representada por 27% encontrada em Condenação e a menor em Outros, com apenas 0,1%.

A precisão indica o valor da predição positiva (número de casos positivos por total de casos cobertos). Os resultados foram melhores para a classificação de Outros apresentando 95% de precisão, enquanto Improcedência apresentou apenas 25%.

O resultado do *recall* (cobertura) é equivalente aos valores obtidos no verdadeiro positivo (TP), considerando que corresponde ao valor da cobertura de casos.

Os valores obtidos no *F-Measure* indicam um desempenho de 79% para a classificação de Outros e de 19% para a classificação de Improcedência.

Os resultados obtidos em MCC demonstram um desempenho de 75% para a classificação de Outros e de 0,04% para a classificação de Improcedência.

O *ROC Area* do classificador Outros ficou em 85%, representando um bom resultado, considerando que um classificador dito como ótimo terá valores próximos a 1. A pior classificação representada por 0,6% correspondeu a Improcedência.

Por fim, o *PRC Area* do classificador Outros ficou em 79%, enquanto a pior classificação foi Acordo com 0,2%.

Comparando a média ponderada obtida em ambos os experimentos é possível identificar que o experimento 1 apresentou 16% a mais de verdadeiros positivos e 4% a menos de falsos positivos, o que indica que classificou melhor os atributos. Além disso, o experimento 1 também se mostrou mais efetivo quanto a precisão, presentando 13% a mais quando comparado ao experimento 2. Quanto ao F-Measure, o experimento 1 obteve 16% a mais de desempenho do que o experimento 2 e no MCC 19%. Por fim, no ROC Area o experimento 1 alcançou 12% a mais que o experimento 1 e no PRC Área 16% a mais, realizando uma melhor classificação dos atributos. Considerando esses aspectos, o experimento 1 mostrou-se mais eficiente em todos os parâmetros de acurácia. Contudo, o experimento 2 gerou como contribuição e facilidade a simplificação para análise dos resultados, permitindo gerar a árvore de decisão.

A Figura 61 apresenta a matriz de confusão obtida com o experimento 1, demonstrando na diagonal principal as instâncias que foram corretamente classificadas.

Figura 61 - Matriz de Confusão REPTree - Experimento 1

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  60   3   6  52  14 |   a = EXTINCAO_SEM_MERITO
   6  13  17  21   5 |   b = ACORDO
   8   1 156   5   5 |   c = CONDENACAO
  31  12   4  97   4 |   d = IMPROCEDENCIA
  25   1  15   5 135 |   e = OUTROS
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Para facilitar a análise dos resultados foi elaborada a Tabela 16 apresentando a porcentagem de atributos que foram classificados corretamente. Nesse experimento é possível identificar que o motivo de arquivamento Outros obteve melhores resultados, realizando a classificação correta de 83% dos atributos, seguido de

Condenação (79%), Improcedência (54%), Extinção sem Mérito (46%) e Acordo (43%).

Tabela 16 - Matriz de Confusão REPTree- Experimento 1

a	b	c	d	e	<--	classified as
60	3	6	52	14	|	a = EXTINCAO_SEM_MERITO
6	13	17	21	5	|	b = ACORDO
8	1	156	5	5	|	c = CONDENACAO
31	12	4	97	4	|	d = IMPROCEDENCIA
25	1	15	5	135	|	e = OUTROS
130	30	198	180	163		**TOTAL DE INSTÂNCIAS**
46%	**43%**	**79%**	**54%**	**83%**		

FONTE: Elaborado pela autora (2016)

Já a Figura 62 apresenta a matriz de confusão obtida com o segundo experimento.

Figura 62 - Matriz de Confusão REPTree - Experimento 2

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  61  10  41  20   3 |   a = EXTINCAO_SEM_MERITO
  15  19  12  11   5 |   b = ACORDO
  29   5 124  17   0 |   c = CONDENACAO
  45  14  64  23   2 |   d = IMPROCEDENCIA
  24   5  17   8 127 |   e = OUTROS
```

FONTE: Resultados da pesquisa utilizando o software Weka (2016)

Para facilitar a análise dos resultados foi elaborada a Tabela 17 demostrando a porcentagem de atributos que foram classificados corretamente. É possível identificar que o motivo de arquivamento Outros também obteve melhores resultados no segundo experimento, realizando a classificação correta de 95% dos atributos, seguido de Condenação (46), Extinção Sem Mérito (38%), Acordo (33%) e Improcedência (26%).

Tabela 17 - Matriz de Confusão REPTree - Experimento 2

A	b	c	D	e	<--	classified as
62	9	42	22	0	|	a = EXTINCAO_SEM_MERITO
15	16	13	14	4	|	b = ACORDO
24	5	125	21	0	|	c = CONDENACAO
40	13	70	23	2	|	d = IMPROCEDENCIA
23	5	20	9	124	|	e = OUTROS
164	48	270	89	130		**TOTAL**
38%	**33%**	**46%**	**26%**	**95%**		

FONTE: Elaborado pela autora (2016)

Ao comparar os resultados obtidos na matriz de confusão é possível identificar na categoria Outros houve o aumento de 13% nas instâncias classificadas corretamente no experimento 2. Em Improcedência o experimento 1 classificou 28% de instâncias a mais de que o experimento 2. Em condenação o experimento 1 apresentou melhores resultados, classificando corretamente 32% de instâncias a mais que o experimento 2. Em Acordo houve 10% a mais de instâncias classificadas corretamente no experimento 1 e 8% a mais em Extinção Sem Mérito.

Após concluir a análise dos resultados apresentados pela heurística REPTree prosseguiu-se para a análise dos resultados obtidos com os experimentos.

4.3.6 Análise dos Resultados Obtidos

A partir dos testes realizados com os algoritmos Apriori, PART, *Decision Table*, J48 e REPTree foi possível identificar que, de maneira geral, o experimento 1 apresentou melhores resultados quanto a acurácia da base de dados. No entanto, quanto a análise dos resultados, o experimento 2 mostrou-se mais eficiente, considerando a apresentação mais simplificada e compreensível os resultados.

A Figura 63 apresenta uma comparação entre o desempenho de classificação dos algoritmos no primeiro experimento. Nele é possível identificar o tempo de execução e os parâmetros de validação cruzada estratificada apresentados pelo Weka.

Figura 63 - Desempenho de Classificação - Experimento 1

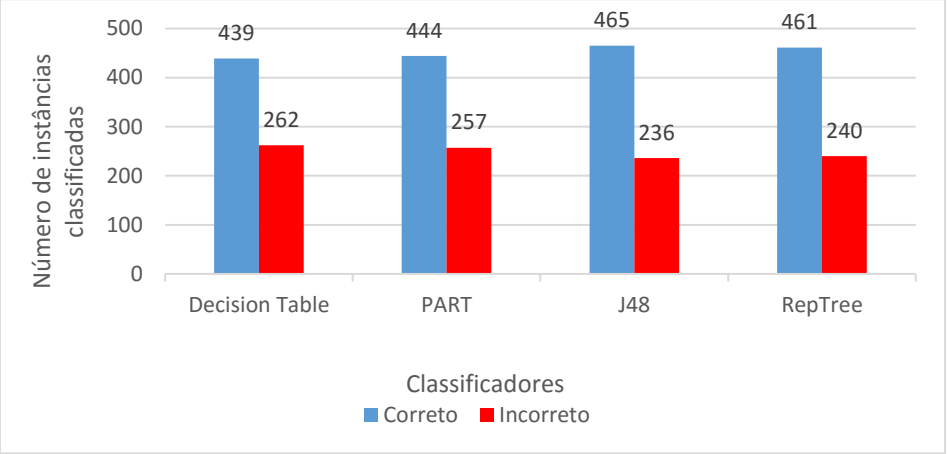
Parâmetros	*Decision Table*	PART	J48	REPTree	Média	Desvio Padrão
Correto	439,0000	444,0000	465,0000	461,0000	452,2500	12,6853
Incorreto	262,0000	257,0000	236,0000	240,0000	248,7500	12,6853
Tempo de Execução	0,5300	0,2000	0,0000	0,0800	0,2025	0,2333
Kappa Statistic	0,5173	0,5285	0,5678	0,5587	0,5431	0,0240
Mean absolute error	0,1989	0,1704	0,1714	0,1776	0,1796	0,0133
Root mean squared error	0,3101	0,3278	0,3118	0,3102	0,3150	0,0086
Relative absolute error (%)	63,6192	54,4967	54,8111	56,7864	57,4284	4,2498
Root relative squared error (%)	78,4263	82,9222	78,8722	78,4613	79,6705	2,1772

FONTE: Elaborado pela autora (2016)

O tempo de execução foi bom para todos os experimentos, tendo em vista que a base de dados contém apenas 701 instâncias. Contudo, é possível identificar que o *Decision Table* foi o que demorou mais tempo para ser executado. Quanto ao *Kappa Statistic* (Estatística Kappa) o resultado foi inconclusivo, considerando que os valores ficaram na metade do intervalo entre 0 e 1, não permitindo afirmar que existe correlação. Quanto ao *Mean absolute error* (Erro Absoluto Médio) o algoritmo PART foi o que apresentou melhor resultados, gerando menor erro na classificação dos atributos. Quanto ao *Root Mean Squared Erro*r (Erro Quadrado Médio) o algoritmo *Decision Table* apresentou menor erro entre os valores atuais e os valores preditos, porém ambos os experimentos tiveram valores aproximados. Quanto ao *Relative Absolute Erro*r (Erro Absoluto Relativo) o algoritmo PART obteve mais precisão para previsão numérica do que os demais experimentos. Por fim, quanto ao *Root Relative Squared Error* (Raiz do Erro Quadrado Relativo) o algoritmo *Decision Table* obteve melhores resultados, pois apresentou menor erro que os demais experimentos.

A Figura 64 apresenta a comparação entre os resultados dos algoritmos com base nas instâncias classificas correta e incorretamente. Analisando os dados é possível identificar que o algoritmo J48 realizou uma classificação mais eficiente dos atributos, seguido de REPTree, PART e *Decision Table*.

Figura 64 - Gráfico Desempenho de Classificação - Experimento 1



FONTE: Elaborado pela autora (2016)

A Figura 65 apresenta uma comparação entre o desempenho de classificação dos algoritmos no segundo experimento. Nele é possível identificar o tempo de execução e os parâmetros de validação cruzada estratificada apresentados pelo Weka.

Figura 65 - Desempenho de Classificação - Experimento 2

Parâmetros	*Decision Table*	PART	J48	REPTree	Média	Desvio Padrão
Correto	361,0000	349,0000	354,0000	350,0000	353,5000	5,4467
Incorreto	340,0000	352,0000	347,0000	351,0000	347,5000	5,4467
Tempo de Execução	0,0100	0,0800	0,0000	0,01	0,0300	0,0436
Kappa Statistic	0,3744	0,3539	0,3649	0,3564	0,3624	0,0093
Mean absolute error	0,2427	0,2338	0,2330	0,2349	0,2361	0,0045
Root mean squared error	0,3472	0,3485	0,3511	0,3509	0,3494	0,0019
Relative absolute error (%)	77,6157	74,7582	74,5140	75,1386	75,5066	1,4293
***Root relative squared error* (%)**	87,8149	88,1387	88,8055	88,7495	88,3772	0,4814

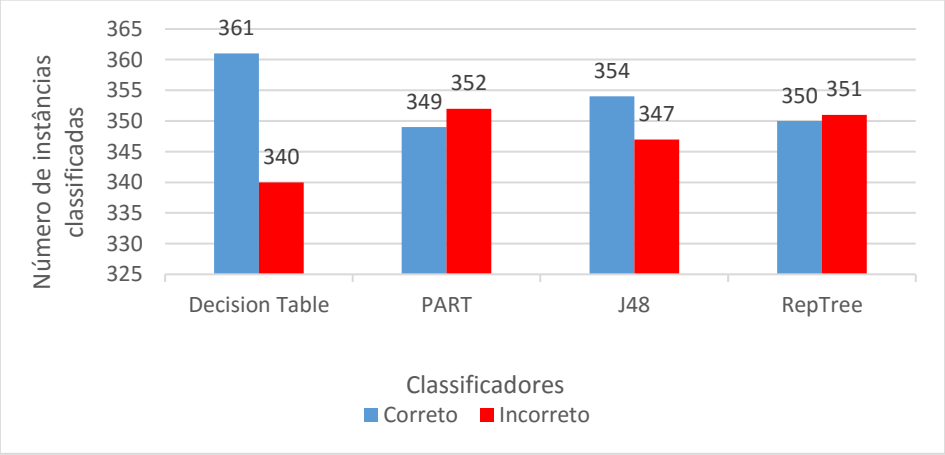
FONTE: Elaborado pela autora (2016)

O tempo de execução foi bom para todos os experimentos, tendo o mais lento demorado 0,08 (PART) segundos e o mais rápido 0. (J48). Houve também uma redução no tempo de execução quando comparado ao primeiro experimento, considerando a redução do número de atributos de 10 para 4. Quanto ao *Kappa Statistic* (Estatística Kappa) o resultado foi inconclusivo, considerando que os valores ficaram na metade do intervalo entre 0 e 1, não permitindo afirmar que existe correlação. Quanto ao *Mean absolute error* (Erro Absoluto Médio) o algoritmo J48 foi o que apresentou melhor resultados, gerando menor erro na classificação dos

atributos. Quanto ao *Root Mean Squared Erro*r (Erro Quadrado Médio) o algoritmo *Decision Table* apresentou menor erro entre os valores atuais e os valores preditos. Quanto ao *Relative Absolute Erro*r (Erro Absoluto Relativo) o algoritmo J48 obteve mais precisão para previsão numérica do que os demais experimentos. Por fim, quanto ao *Root Relative Squared Error* (Raiz do Erro Quadrado Relativo) o algoritmo *Decision Table* obteve melhores resultados, pois apresentou menor erro que os demais experimentos.

A Figura 66 apresenta a comparação entre os resultados dos algoritmos com base nas instâncias classificas correta e incorretamente. Analisando os dados é possível identificar que o algoritmo *Decision Table* realizou uma classificação mais eficiente dos atributos, seguido de J48, REPTree e Part.

Figura 66 - Gráfico Desempenho de Classificação - Experimento 2

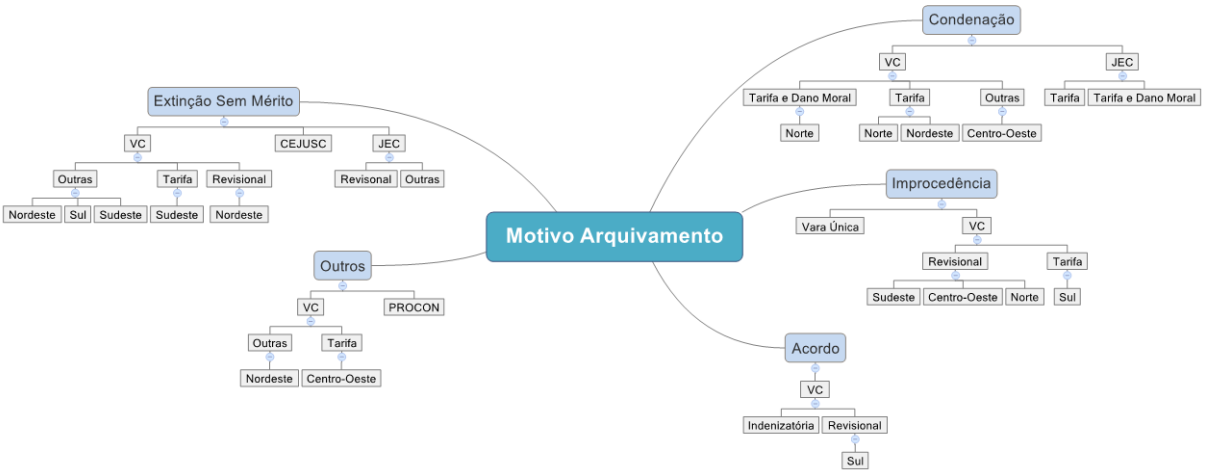


FONTE: Elaborado pela autora (2016)

Analisando os resultados e desempenhos obtidos pelos experimentos os algoritmos J48 e *Decision Table* atenderam melhor as características da base de dados. Eles conseguiram apresentar resultados satisfatórios para análise, mantendo a acurácia da base de dados e tornando os resultados relevantes para a análise. O J48 apresenta como vantagem gerar a árvore de decisão que facilita a análise para a tomada de decisão. O *Decision Table*, por sua vez, gera uma tabela de decisão que também permite analisar condições, contudo, torna a análise mais demorada por não gerar uma representação gráfica. Além disso, o algoritmo gera resultados mais simplificados, tendo em vista que não considera todas as hipóteses possíveis, sendo que na árvore de decisão podem ser analisados todos os caminhos possíveis.

Na árvore de decisão gerada pelo algoritmo J48 e REPTree é possível identificar padrões referentes ao motivo de arquivamento. A Figura 67 sintetiza os dados apresentados pela árvore de decisão gerada pelo algoritmo J48 através do agrupamento dos resultados com base no motivo do arquivamento.

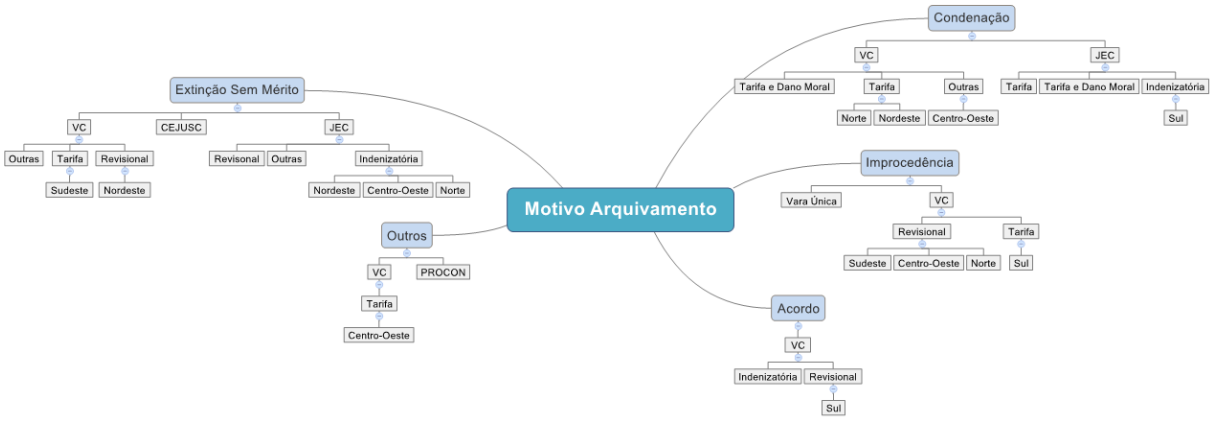
Figura 67 - Mapa conceitual - Motivo Arquivamento J48



FONTE: Elaborado pela autora (2016)

Já a Figura 68 apresenta os resultados da árvore de decisão gerada pelo algoritmo REPTree agrupados de acordo com o motivo de arquivamento.

Figura 68 - Mapa conceitual - Motivo Arquivamento REPTree



FONTE: Elaborado pela autora (2016)

Com base nos dois modelos é possível identificar muitas similaridades e algumas diferenças. Entre as principais diferenças observadas o REPTree apresentou Condenação para as ações de tarifa e dano moral que tramitam na vara cível, enquanto o J48 classificou com base na região. Já para as ações que tramitam no JEC o algoritmo REPTree criou um novo subnível, indicando Condenação para ações indenizatórias que tramitam na região Sul, enquanto pelo algoritmo J48 estas foram classificadas como Extinção sem Mérito. Em vara cível com tipo de ação Outras o REPTree classificou o motivo de arquivamento como Extinção Sem Mérito, enquanto o J48 como Outros, criando um subnível para a região nordeste. Em JEC com tipo de ação indenizatória o REPTree e o J48 classificaram o motivo de arquivamento como Condenação, porém o REPTree apresentou mais um nível da árvore de decisão para região Sul. Já em Extinção Sem Mérito é possível verificar o encurtamento da árvore pelo REPTree, apresentando o resultado diretamente, enquanto no J48 houve o subnível da região Sudeste, Sul e Nordeste.

As árvores permitiram identificar alguns padrões com base no motivo do arquivamento. Em ações improcedentes foram identificados processos que tramitam na Vara Única; ações revisionais que tramitam na vara cível da região sudeste, centro-oeste e norte; e ações que discutem tarifas que tramitam na vara cível da região sul.

Em Condenação foi identificado no JEC processos de tarifa e tarifa e dano moral e na Vara Cível ações de tarifa e dano moral que tramitam na região norte, ações de tarifa na região norte e nordeste e ações outras na região centro-oeste.

Em Acordo foi identificado na vara cível ações indenizatórias e ações revisionais que tramitam na região sul.

Em Extinção Sem Mérito houve maior dificuldade para reconhecimento dos padrões para os dois algoritmos, com exceção de ações que tramitam no CEJUSC, por representarem menor porcentagem. No JEC foram identificadas ações revisionais e outras e ações indenizatórias no nordeste, centro-oeste e norte. Na vara cível foram identificadas ações outras, tarifa na região sudeste e revisional na região nordeste.

Em Outros foram identificadas ações que tramitam no PROCON e na vara cível ações de tarifa da região centro-oeste.

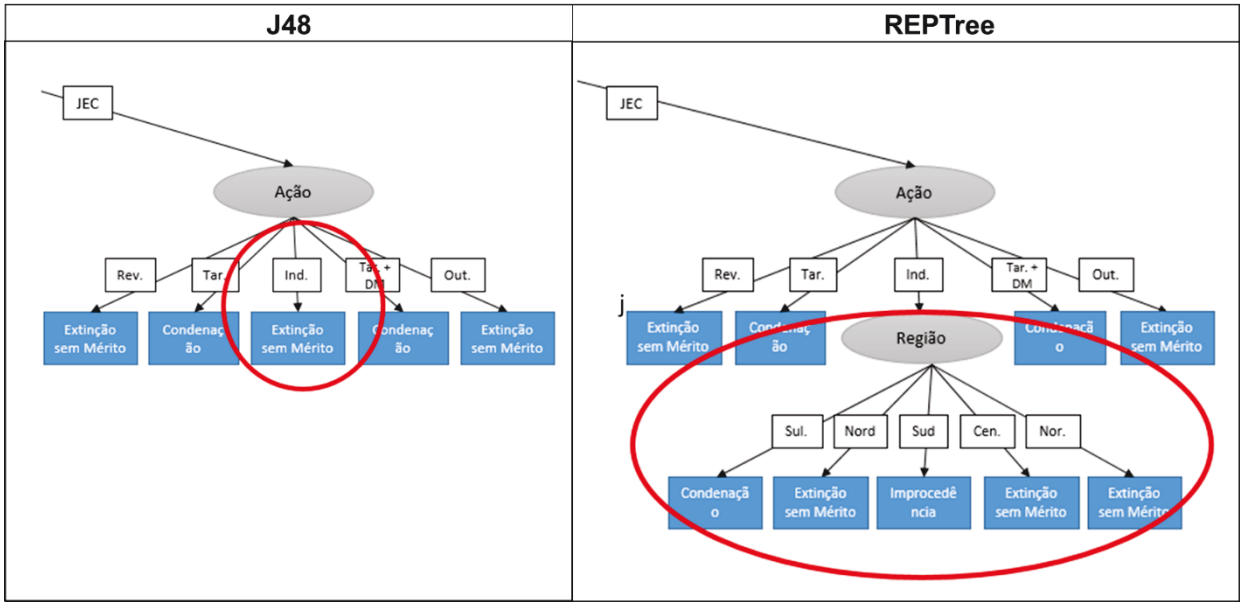
Após a análise dos padrões identificados, a próxima subseção aborda a validação dos resultados por um profissional da área jurídica.

4.3.7 Validação dos Resultados

Conforme mencionado na metodologia (seção 3), para realizar a validação dos resultados foi realizada uma entrevista com duração de aproximadamente uma hora com dois profissionais da área jurídica. Os resultados alcançados foram submetidos a análise, a fim de verificar se foram satisfatórios para identificação de padrões de decisões judiciais, bem como questionar melhorias.

Para facilitar a análise foram apresentadas as árvores de decisão geradas pelos algoritmos J48 e REPTree, pois tornam mais fácil a compreensão dos resultados através da representação gráfica. Após a apresentação e discussão dos resultados foi questionada qual árvore de decisão atende melhor a realidade da área jurídica, com base no conhecimento e experiência que eles possuem sobre o assunto. Ambos concluíram que a árvore gerada pelo REPTree foi mais satisfatória, pois além de estar mais simplificada, apresentou resultados mais próximos da realidade. Os entrevistados apresentaram um exemplo, explicando que ações indenizatórias que tramitam no JEC podem conter diferentes motivos de arquivamento, não se restringindo apenas a Extinção Sem Mérito, conforme resultado apresentado pelo algoritmo J48 (Figura 69).

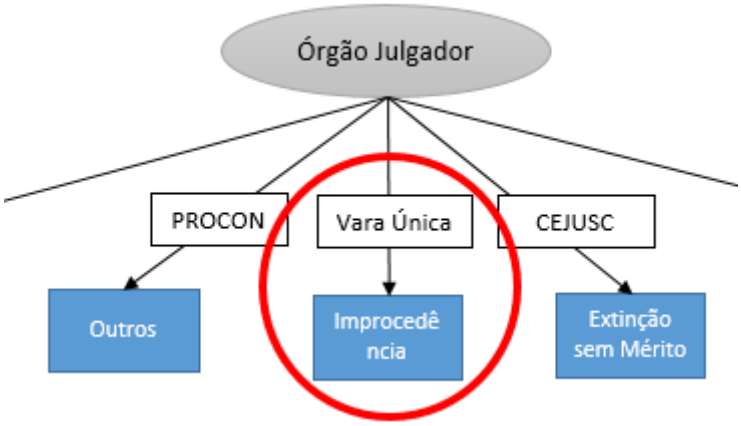
Figura 69 - Comparação árvore de decisão J48 e REPTree



FONTE: Elaborado pela autora (2016)

Outra observação realizada pelos entrevistados foi em relação a Vara Única (Figura 70) que apresentou como resultado Improcedência nas duas árvores de decisões. Segundo os entrevistados, essas ações correspondem à Vara Cível, porém foram classificadas incorretamente na base de dados e poderiam ser desconsideradas.

Figura 70 - Resultado órgão julgador apresentado pela árvore de decisão



FONTE: Elaborado pela autora (2016)

Os entrevistados mostraram-se satisfeitos com os resultados apresentados e destacaram a relevância da pesquisa para a área, pois declararam não ter conhecimento de uma proposta no mercado que realize esse tipo de análise. Além disso, os entrevistados explicaram que tomavam as decisões com base no conhecimento tácito e experiências inerentes a cada um, sem realizar um efetivo estudo. Porém, com a falta de uniformização das decisões judiciais, a identificação de padrões é de suma importância para estabelecimento das estratégias a serem adotadas.

Os entrevistados também comentaram sobre a importância da análise das decisões ao longo do tempo, sendo relevante realizar um planejamento anual para acompanhamento dos novos resultados a partir da mudança nas estratégias. Pela mineração de dados é possível realizar esse tipo de análise, pois ela demonstra o histórico dos processos da organização ao longo do tempo, podendo aumentar ou reduzir a quantidade de condenações de acordo com o posicionamento adotado pelo escritório.

5 CONSIDERAÇÕES FINAIS

Os avanços tecnológicos proporcionaram um crescimento exponencial no volume de dados devido ao aumento de usuários na *internet* e de conteúdos publicadas diariamente. Nesse contexto, tornou-se um desafio gerenciar as informações de forma a extrair conhecimento para a tomada de decisão. Esse crescimento afetou também o ambiente organizacional, no qual as bases de dados cresceram tanto que dificultaram a análise manual, sendo necessárias novas técnicas e ferramentas capazes de analisar grandes volumes de dados de forma inteligente, visando gerar vantagem competitiva.

Nesse cenário a mineração de dados tem sido uma ferramenta de apoio com papel fundamental na gestão da informação dentro das organizações. No entanto, a escolha do método de mineração de dados não é uma tarefa fácil, pois não existe um padrão para a escolha, variando de acordo com os tipos de atributos da base de dados. Assim, é destacada a importância do pré-processamento, considerando que as atividades de limpeza, integração, redução, transformação e discretização são essenciais para realizar uma efetiva mineração de dados. A escolha do software de mineração de dados é outra atividade importante, considerando que o mesmo deve atender ao método de mineração anteriormente definido.

5.1 VERIFICAÇÃO DOS OBJETIVOS PROPOSTOS

Para atingir o objetivo da aplicação de técnicas de mineração de dados em uma base de dados jurídica para identificação de padrões de decisões judiciais, foi necessário alcançar três objetivos específicos.

O primeiro objetivo específico - pesquisar e definir o(s) método(s) de mineração de dados que será(ão) utilizado(s) na base de dados jurídica - foi alcançado primeiramente por meio do levantamento bibliográfico para identificar os principais métodos utilizados na literatura e que apresentassem maior facilidade na compreensão dos resultados. Destaca-se nessa etapa a dificuldade no levantamento de material bibliográfico, pois no Brasil ainda não existem muitas pesquisas desenvolvidas na área e os principais estudos são desenvolvidos na área de Ciência da Computação, apresentando, portanto, uma linguagem mais técnica e de difícil compreensão. Concluindo o levantamento teórico a base de dados foi importada no

software Weka. Primeiramente optou-se pela escolha das tarefas, sendo utilizadas associação e classificação, considerando que atendem melhor o resultado almejado com os experimentos, pois classificam e categorizam os atributos em classes e determinam as correlações entre os itens. Em seguida foram escolhidas as heurísticas e os respectivos algoritmos que foram habilitados pelo Weka de acordo com as características da base de dados.

Na tarefa de associação optou-se pela utilização do algoritmo Apriori, considerando ser o mais utilizado e mencionado na literatura sobre o tema. Contudo, os resultados não foram muito satisfatórios, pois o algoritmo não permite escolher o atributo meta, gerando poucas regras com importância para o enfoque da pesquisa. Como contribuição, o algoritmo conseguiu identificar padrões para processos arquivados por motivo “Outros”, não gerando regras para os demais casos.

Na tarefa de classificação optou-se pela utilização das heurísticas árvores de decisão e regras, tendo em vista que apresentam resultados mais compreensíveis, permitindo a análise por pessoas que não dominem as técnicas de mineração de dados. Para cada algoritmo foi realizado dois experimentos: o primeiro contendo todos os atributos da base de dados e o segundo somente com os quatro principais. O segundo experimento foi realizado para reduzir a quantidade de informações e facilitar a compreensão dos resultados, pois a base de dados continha atributos que não exerciam impacto direto sobre o atributo meta analisado. De maneira geral o primeiro experimento mostrou-se mais efetivo, apresentando melhor acurácia e classificação das instâncias. Contudo, para análise e compreensão dos resultados, o segundo experimento foi mais eficaz, pois conseguiu gerar resultados compreensíveis. Houve apenas o retrabalho de gerar as árvores de decisão manualmente por deficiência do software que não permite o redimensionamento das folhas.

O segundo objetivo específico - classificar os Estados com base nos padrões de decisões identificadas - foi atingido parcialmente, pois para realizar a análise foi necessário separar os Estados em regiões e, a partir disso, verificar os principais motivos de arquivamento. Contudo, essa análise foi possível a partir da árvore de decisão, que demonstra para cada região o motivo de arquivamento, com base no tipo de ação e órgão julgador em que tramita o processo.

O terceiro objetivo específico - identificar o formato da base de dados que viabilize a aplicação de métodos de mineração de dados – foi alcançado a partir da alteração do formato da base de dados, discretizando os atributos numéricos em

intervalos e mantendo apenas os atributos nominais, visando aumentar a possibilidade de algoritmos disponíveis para a execução. Ao realizar os ajustes alguns atributos que não foram considerados interessantes para a análise foram retirados, resultado na base de dados com dez atributos.

5.2 CONTRIBUIÇÕES

Devido à falta de conhecimento sobre a área jurídica, foi elaborado um glossário de termos para facilitar a compreensão de alguns conceitos recorrentes. Além disso, os resultados precisaram ser submetidos a um profissional da área para validação da proposta. Em uma entrevista com duração de aproximadamente uma hora, dois advogados do escritório que cedeu a base de dados para o estudo analisaram a veracidade dos resultados apresentados. Na entrevista foi possível identificar a importância da criação manual das árvores de decisão, pois a apresentação gráfica permitiu aos advogados compreenderem cada situação, percorrendo os caminhos possíveis da árvore para identificar possíveis inconsistências. O *feedback* foi muito interessante e contribuiu para o estudo, pois demonstrou a importância crescente da aplicação de tecnologias da informação na área jurídica, sendo um ramo interessante para pesquisas aprofundadas. Além disso, conforme a pesquisa já apontava, não existem muitos estudos desenvolvidos que auxiliem esses profissionais a realizarem a tomada de decisão com base em informações fundamentadas. Conforme apontado pelos entrevistados e afirmado por Surden (2014), esses profissionais utilizam uma mistura de experiências e habilidades para fazerem as avaliações e conseguirem possíveis resultados. Nesse contexto, a mineração de dados permite apresentar os resultados com base no histórico identificado nos processos, apoiando a análise dos advogados. A partir disso, os mesmos podem estudar a estratégia de atuação, propondo mudanças para os casos em que são identificadas sentenças desfavoráveis, a fim de melhorar o resultado a seu favor. Um exemplo evidenciado pela árvore de decisão consiste nas condenações recebidas em ações que discutem tarifa e tarifa e dano moral no JEC e na Vara Cível, sendo uma das possíveis soluções a proposição de acordos para o encerramento dessas ações.

5.3 TRABALHOS FUTUROS

Para trabalhos futuros sugere-se a aplicação do estudo em outras bases de dados jurídicas, de forma a validar a proposta e comparar as mudanças nos resultados obtidos. Além disso, é recomendada a aplicação das técnicas em outras áreas do direito, a fim de verificar se também ocorre a falta de uniformização das decisões jurídicas. Por fim, é sugerido testar a aplicação de outras técnicas de mineração de dados para verificar a descoberta de novos conhecimentos.

REFERÊNCIAS

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Recuperação da informação:** conceitos e tecnologia das máquinas de busca. 2. ed. Porto Alegre: Bookman, 2013. 590 p.

BRAGA, Ryon. **O excesso de informação**: a neurose do século XXI. Disponível em: <http://www.mettodo.com.br/pdf/O%20Excesso%20de%20Informacao.pdf>. Acesso em: 29 mar. 2016.

BRASIL. Constituição (2015). Lei nº 13.105, de 16 de março de 2015. **Código de Processo Civil.** Brasília, DF, 16 mar. 2015. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13105.htm>. Acesso em: 10 maio 2016.

CAETANO, A.G.L.S. **Sistemas de supervisão de chão-de-fábrica**: uma contribuição para implantação em indústrias de usinagem. Dissertação de Mestrado, Escola de Engenharia de São Carlos da Universidade de São Paulo, SP, 2000.

CALIL, L. A. A. *et al.* **Mineração de dados e pós-processamento em padrões descobertos.** Publ. UEPG Ci. Exatas Terra, Ci. Agr. Eng., Ponta Grossa, 14 (3): 207-215, dez. 2008. Disponível em: <http://www.revistas2.uepg.br/index.php/exatas/article/view/946>. Acesso em: 27 mai. 2016.

CASTRO, Leandro Nunes; FERRARI, Daniel Gomes. **Introdução à mineração de dados:** conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de dados:** conceitos, tarefas, métodos e ferramentas. Goiás: Instituto de Informática - Universidade Federal de Goiás, 2009. 29 p. (RT-INF_001-09). Disponível em: <http://www.inf.ufg.br/sites/*default*/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 8 mai. 2016.

CHAGAS, Cadu. **Extinção do processo com julgamento de mérito.** Disponível em: < https://caduchagas.blogspot.com.br/2012/05/extincao-do-processo-com-julgamento-de.html>. Acesso em: 23 ago. 2016.

CHEN, M. S.; HAN, J.; YU, P. S. ***Data mining***: an overview from database perspective. IEEE Transactions on Knowledge and Data Engineering, 1996.

CHIAVENATO, Idalberto. **Introdução à teoria geral da administração.** 7. ed. Rio de Janeiro: Campus, 2003. 630 p.

CHOO, Chun Wei. **A organização do conhecimento**: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. São Paulo: Editora Senac, 2003.

DAVENPORT, Thomas H. **Ecologia da informação**: por que só a tecnologia não basta para o sucesso na era da informação. São Paulo: Futura, 1998. 312 p.

DELGADO, José Augusto. **A imprevisibilidade das decisões jurídicas e seus reflexos na segurança jurídica.** Disponível em: <http://goo.gl/0IS7qn>. Acesso em: 02 mai. 2016.

ESTRADA, M. M. P. **A criação do direito pela inteligência artificial.** Disponível em: <http://direitoeti.com.br/artigos/a-criacao-do-direito-pela-inteligencia-artificial/>. Acesso em: 25 fev. 2016.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From *data mining* to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996. Disponível em: <http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf>. Acesso em: 03 mar. 2016.

FERNANDES, Luciane Alves; GOMES, José Mario Matsumura. Relatórios de pesquisa nas ciências sociais: características e modalidades de investigação. **ConTexto**, Porto Alegre, v. 3, n. 4, 1º semestre 2003. Disponível em: <http://www.praticadapesquisa.com.br/2013/02/relatorios-de-pesquisa-nas-ciencias.html>. Acesso em: 09 abr. 2016.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2009.

Glossário de Termos Jurídicos. Ministério Público Federal. Disponível em: <http://www.prba.mpf.mp.br/sala-de-imprensa/glossario>. Acesso em: 15 abr. 2016.

Lei nº 3.071 de 01 de Janeiro de 1916. Disponível em: <http://www.jusbrasil.com.br/topicos/11482313/artigo-159-da-lei-n-3071-de-01-de-janeiro-de-1916>. Acesso em: 16 ago. 2016.

MARTINEZ, Luís; CASAL, Ricardo; JANEIRO, João. **Sistemas de apoio à decisão clínica.** Porto: Faculdade de Medicina do Porto, 2009. 16 p.

MAXIMIANO, A. C. A. **Introdução à administração.** 5. ed. rev. e ampl. – São Paulo: Atlas, 20

MORAES, G. D. A.; FILHO, E. E. A gestão da informação diante das especificidades das pequenas empresas. **Ci. Inf**., Brasília, v. 35, n. 3, p. 124-132, set./dez. 2006. Disponível em: <http://www.scielo.br/pdf/ci/v35n3/v35n3a12.pdf>. Acesso em: 8 mai. 2016.

NEVES, R. C. D. **Pré-processamento no processo de descoberta de conhecimento em banco de dados.** 2003. 137 f. DISSERTAÇÃO (Mestrado) - Programa de Pós-graduação em Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul.

NOGUEIRA, E. D. A. **Avaliação de *brand equity* sob a perspectiva do consumidor nas mídias sociais por meio da mineração de opinião e análise de redes sociais.** 2015. 237 f. DISSERTAÇÃO (Mestrado) – Programa de Pós-graduação em Ciência, Gestão e Tecnologia da Informação, Ciências Sociais Aplicadas, Universidade Federal do Paraná.

RADDATZA, Joice. **A ação revisional de contratos.** Disponível em: < http://joiceraddatz.jusbrasil.com.br/artigos/117586531/a-acao-revisional-de-contratos>. Acesso em: 16 ago. 2016

ROVER, Tadeu. **Controle de processos:** conheça os Softwares jurídicos mais usados por escritórios e empresas. Disponível em: < http://www.conjur.com.br/2015-jun-24/conheca-Softwares-juridicos-usados-advogados>. Acesso em: 20 abr. 2016.

SIDNEY, Christiane Faleiro. **Aplicação de mineração de dados no banco de dados do zoneamento ecológico econômico de minas gerais**. 2010. 60 f. TCC (Graduação) – Sistemas de Informação, Departamento de Ciência da Computação, Universidade Federal de Lavras, Lavras, 2010. Disponível em: <http://goo.gl/zZk0ds>. Acesso em: 08 mar. 2016.

SILVA, Bruno Mattos e. **A súmula vinculante para a administração pública aprovada pela reforma do judiciário.** Disponível em: <http://www.egov.ufsc.br/portal/conteudo/s%C3%BAmula-vinculante-para-administra%C3%A7%C3%A3o-p%C3%BAblica-aprovada-pela-reforma-do-judici%C3%A1rio>. Acesso em: 15 mar. 2016.

STJ. **Previstas no contrato, tarifas em financiamento são legais.** Disponível em: < http://www.migalhas.com.br/Quentes/17,MI168272,71043-Previstas+no+contrato+tarifas+em+financiamento+sao+legais>. Acesso em: 15 mar. 2016

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao *data mining***: mineração de dados. Rio de Janeiro: Editora Ciência Moderna Ltda., 2009. 900 p.

TARAPANOFF, Kira. Inteligência, informação e conhecimento em corporações. Brasília: **IBICT**, UNESCO, 2006. 456 p. Universidade de São Paulo, SP, 2000.

VIEIRA, Eliane Aparecida. **A gestão da informação na tomada das decisões gerenciais:** estudo de caso na organização multinacional de Reflorestamento - v & m florestal. 2011. 81 f. DISSERTAÇÃO (Mestrado) - Curso de Administração, Fundação Cultural Dr. Pedro Leopoldo, Pedro Leopoldo, 2011. Disponível em: <http://www.fpl.edu.br/2013/media/pdfs/mestrado/dissertacoes_2011/dissertacao_eliane_aparecida_vieira_2011.pdf>. Acesso em: 23 abr. 2016.

YIN, R. K. **Estudo de caso**: planejamento e métodos. Porto Alegre: Bookman, 2005.

APÊNDICE A - Glossário de Termos Jurídicos

(continua)

TERMO	DESCRIÇÃO
Ação	Direito que tem qualquer cidadão para buscar uma decisão judicial, por meio de um processo.
Ação cível	É toda aquela em que se pleiteia em juízo um direito de natureza civil, ou seja, não-criminal. Trata de conflitos de natureza civil, ou seja, pertencente às áreas familiar, sucessória, obrigacional ou real.
Autos	É o nome que se dá ao conjunto das peças que compõem um processo, incluindo todos os anexos e volumes.
Busca e apreensão	É a diligência policial ou judicial que tem por fim procurar coisa ou pessoa que se deseja encontrar, para trazê-la à presença da autoridade que a determinou. A busca e apreensão se faz para procurar e trazer a coisa litigiosa, a pedido de uma das partes, para procurar e apreender a coisa roubada ou sonegada. Também se procede a diligência para procurar e trazer à presença da autoridade, que a ordenou, o menor, que saiu do poder de seus pais ou tutores, para recolocá-lo sob o poder destes. Em regra, a busca e apreensão, é de natureza criminal. Mas admite-se em juízo civil e comercial, para trazer as coisas à custódia do juízo, onde se discute quanto ao direito sobre elas.
Comarca	A circunscrição territorial, compreendida pelos limites em que se encerra a jurisdição de um juiz de Direito.
Competência	É a medida ou extensão do poder de jurisdição de um juiz. Ou seja, a competência diz que causas, que pessoas, de que lugar, devem ser julgadas por determinado juiz.
Decisão	Denominação genérica dos atos do juízo, provocada por petições das partes ou do julgamento do pedido. Em sentido estrito, pronunciamento do juiz que resolve questão incidente.
Decisão judicial	Todo e qualquer despacho proferido por um juiz ou tribunal, em qualquer processo ou ato submetido a sua apreciação e veredito.
Deferir	Acolher um requerimento, um pedido, uma pretensão.
Demanda	É todo pedido feito em juízo.
Denegar	Indeferir, negar uma pretensão formulada em juízo.
Instância	Grau da hierarquia do Poder Judiciário. A primeira instância, onde em geral começam as ações, é composta pelo juiz de direito de cada comarca, pelo juiz federal, eleitoral e do trabalho. A segunda instância, onde são julgados recursos, é formada pelos tribunais de Justiça e de Alçada, e pelos tribunais regionais federais, eleitorais e do trabalho. A terceira instância são os tribunais superiores (STF, STJ, TST, TSE) que julgam recursos contra decisões dos tribunais de segunda instância.

(conclusão)

Liminar	**Pedido de antecipação dos efeitos da decisão, antes do seu julgamento. É concedido quando a demora da decisão causar prejuízos. Ao examinar a liminar, o ministro relator também avalia se o pedido apresentado tem fundamentos jurídicos aceitáveis.**
Petição	De forma geral, é um pedido escrito dirigido ao tribunal. A petição Inicial é o pedido para que se comece um processo. Outras petições podem ser apresentadas durante o processo para requerer o que é de interesse ou de direito das partes.
Preliminar	São questões que devem ser decididas antes do mérito, porque dizem respeito à própria formação da relação processual. Por exemplo, a discussão sobre a competência de um juiz para julgamento de uma causa constitui espécie de preliminar; assim também a legitimidade da parte para fazer aquele pedido. Por isso, o julgamento das preliminares pode impedir o próprio julgamento do mérito, caso sejam julgadas procedentes
Processo	Atividade por meio da qual se exerce concretamente, em relação a determinado caso, a função jurisdicional, e que é instrumento de composição das lides; pleito judicial; litígio; conjunto de peças que documentam o exercício da atividade jurisdicional em um caso concreto; autos.
Recurso	Instrumento para pedir a mudança de uma decisão, na mesma instância ou em instância superior.
Sentença	Decisão do juiz que põe fim a um processo.
Tutela	Encargo ou autoridade que se confere a alguém, por lei ou por testamento, para administrar os bens e dirigir e proteger um menor que se acha fora do pátrio poder, bem como para representá-lo ou assistir-lhe nos atos da vida civil; defesa, amparo, proteção; tutoria; dependência ou sujeição vexatória.
Tutela antecipada	É a antecipação de um ou mais pedidos feitos pelo autor na ação. Exige alguns requisitos, como a possibilidade de que a demora no julgamento da causa resulte em prejuízo irreparável à parte, bem como a existência de provas que convençam o juiz da veracidade da alegação. Ver artigo 273 e parágrafos do Código de Processo Civil.

FONTE: Elaborado pela autora (2016). Adaptado do Glossário do STF.

APÊNDICE B - Resultado aplicação algoritmo PART – Experimento 1

6 RISCO_ATUAL = PROVAVEL AND OJ = JEC AND ACAO = TARIFA: CONDENACAO (68.0/7.0)
7 RISCO_ATUAL = POSSIVEL AND OJ = PROCON: OUTROS (119.64/2.0)
8 RISCO_ATUAL = PROVAVEL AND OJ = JEC AND ACAO = TARIFA E DANO MORAL: CONDENACAO (54.0/8.0)
9 RISCO_ATUAL = PROVAVEL AND OJ = JEC AND ACAO = INDENIZATORIA: CONDENACAO (27.0/7.0)
10 RISCO_ATUAL = PROVAVEL AND OJ = JEC AND VALOR_PROVISAO = I1: EXTINCAO_SEM_MERITO (5.0/2.0)
11 RISCO_ATUAL = PROVAVEL AND ACAO = OUTRAS: CONDENACAO (8.07/2.07)
12 RISCO_ATUAL = PROVAVEL AND VALOR_PROVISAO = I3: CONDENACAO (9.0/1.0)
13 RISCO_ATUAL = PROVAVEL AND ACAO = INDENIZATORIA AND REGIAO = SUL: ACORDO (3.0)
14 RISCO_ATUAL = PROVAVEL AND VALOR_CAUSA = I1 AND ACAO = REVISIONAL: CONDENACAO (12.0/5.0)
15 RISCO_ATUAL = PROVAVEL AND REGIAO = SUL: CONDENACAO (2.0)
16 RISCO_ATUAL = PROVAVEL AND REGIAO = CENTRO-OESTE AND ACAO = TARIFA: OUTROS (2.0)
17 RISCO_ATUAL = PROVAVEL AND REGIAO = CENTRO-OESTE: CONDENACAO (3.0)
18 RISCO_ATUAL = REMOTO AND OJ = JEC AND REGIAO = NORTE: IMPROCEDENCIA (12.0/4.0)
19 RISCO_ATUAL = REMOTO AND OJ = JEC AND REGIAO = NORDESTE AND VALOR_PROVISAO = I1 AND ACAO = INDENIZATORIA AND VALOR_CAUSA = I1 AND
20 REU = Companhia de Credito Financiamento e Investimento RCI Brasil: XTINCAO_SEM_MERITO (6.0)
21 RISCO_ATUAL = REMOTO AND OJ = JEC AND REGIAO = NORDESTE AND VALOR_PROVISAO = I3: IMPROCEDENCIA (10.0/3.0)
22 RISCO_ATUAL = REMOTO AND OJ = PROCON: OUTROS (4.95)
23 RISCO_ATUAL = REMOTO AND OJ = JEC AND REGIAO = NORDESTE AND RISCO_ANTERIOR = POSSIVEL AND REU = Companhia de Credito Financiamento e Investimento RCI Brasil AND VALOR_RISCO = I1: IMPROCEDENCIA (4.0/1.0)
24 RISCO_ATUAL = REMOTO AND OJ = JEC AND REGIAO = NORDESTE AND RISCO_ANTERIOR = POSSIVEL: EXTINCAO_SEM_MERITO (7.0/1.0)
25 RISCO_ATUAL = REMOTO AND OJ = JEC AND REGIAO = NORDESTE AND RISCO_ANTERIOR = PROVAVEL: IMPROCEDENCIA (3.0)
26 RISCO_ATUAL = REMOTO AND OJ = JEC AND REGIAO = NORDESTE AND VALOR_PROVISAO = I1 AND ACAO = INDENIZATORIA AND VALOR_CAUSA = I1: IMPROCEDENCIA (3.0/1.0)
27 RISCO_ATUAL = REMOTO AND OJ = JEC AND REGIAO = NORDESTE AND ACAO = TARIFA E DANO MORAL AND REU = Companhia de Credito Financiamento e Investimento RCI Brasil: EXTINCAO_SEM_MERITO (5.0/1.0)
28 RISCO_ATUAL = REMOTO AND OJ = JEC AND REGIAO = NORDESTE AND VALOR_PROVISAO = I1 AND ACAO = TARIFA AND REU = Companhia de Credito Financiamento e Investimento RCI Brasil: IMPROCEDENCIA (7.0/3.0)
29 RISCO_ATUAL = REMOTO AND ACAO = TARIFA E DANO MORAL: IMPROCEDENCIA (22.0/6.0)
30 RISCO_ATUAL = REMOTO AND ACAO = TARIFA AND REGIAO = SUDESTE AND OJ = JEC: IMPROCEDENCIA (18.0/1.0)
31 RISCO_ATUAL = REMOTO AND ACAO = REVISIONAL AND OJ = JEC AND REGIAO = SUL AND REU = Companhia de Arrendamento Mercantil RCI Brasil: IMPROCEDENCIA (5.0/1.0)
32 RISCO_ATUAL = REMOTO AND ACAO = REVISIONAL AND OJ = JEC AND REGIAO = SUDESTE: IMPROCEDENCIA (5.0/1.0)
33 RISCO_ATUAL = POSSIVEL AND ACAO = OUTRAS AND VALOR_CAUSA = I1: EXTINCAO_SEM_MERITO (3.0)
34 RISCO_ATUAL = REMOTO AND ACAO = REVISIONAL AND OJ = JEC AND REGIAO = SUL: EXTINCAO_SEM_MERITO (3.0)

35 RISCO_ATUAL = REMOTO AND ACAO = REVISIONAL AND REGIAO = SUDESTE AND RISCO_ANTERIOR = REMOTO: IMPROCEDENCIA (15.0/3.0)
36 RISCO_ATUAL = PROVAVEL AND ACAO = TARIFA: CONDENACAO (3.01/1.01)
37 RISCO_ATUAL = PROVAVEL AND VALOR_CAUSA = I1: ACORDO (2.0/0.0)
38 RISCO_ATUAL = POSSIVEL AND ACAO = REVISIONAL: EXTINCAO_SEM_MERITO (14.0/5.0)
39 RISCO_ATUAL = REMOTO AND ACAO = REVISIONAL AND REGIAO = SUL: ACORDO (33.0/14.0)
40 RISCO_ATUAL = REMOTO AND ACAO = REVISIONAL AND REU = Companhia de Arrendamento Mercantil RCI Brasil: EXTINCAO_SEM_MERITO (7.0/2.0)
41 RISCO_ATUAL = REMOTO AND ACAO = REVISIONAL AND VALOR_PROVISAO = I3: IMPROCEDENCIA (5.0)
42 RISCO_ATUAL = REMOTO AND ACAO = REVISIONAL AND VALOR_RISCO = I1: IMPROCEDENCIA (13.0/6.0)
43 RISCO_ATUAL = REMOTO AND ACAO = REVISIONAL AND VALOR_RISCO = I3: EXTINCAO_SEM_MERITO (4.0/1.0)
44 RISCO_ATUAL = REMOTO AND ACAO = OUTRAS AND VALOR_CAUSA = I3: EXTINCAO_SEM_MERITO (15.09/4.09)
45 RISCO_ATUAL = POSSIVEL AND VALOR_PROVISAO = I3 AND REGIAO = SUDESTE: EXTINCAO_SEM_MERITO (8.0/3.0)
46 RISCO_ATUAL = POSSIVEL AND VALOR_PROVISAO = I3: OUTROS (7.0/2.0)
47 RISCO_ATUAL = POSSIVEL AND OJ = VC: OUTROS (11.0/3.0)
48 RISCO_ATUAL = POSSIVEL AND REGIAO = NORDESTE AND ACAO = INDENIZATORIA: EXTINCAO_SEM_MERITO (4.0/2.0)
49 RISCO_ATUAL = POSSIVEL AND REGIAO = CENTRO-OESTE: EXTINCAO_SEM_MERITO (7.0/3.0)
50 RISCO_ATUAL = POSSIVEL AND VALOR_RISCO = I1 AND VALOR_CAUSA = I2: EXTINCAO_SEM_MERITO (8.0/3.0)
51 RISCO_ATUAL = POSSIVEL AND REGIAO = NORDESTE AND ACAO = TARIFA: OUTROS (4.0/1.0)
52 RISCO_ATUAL = POSSIVEL: CONDENACAO (18.04/7.04)
53 ACAO = INDENIZATORIA AND VALOR_RISCO = I1 AND RISCO_ANTERIOR = REMOTO AND VALOR_CAUSA = I1 AND REGIAO = SUDESTE: IMPROCEDENCIA (9.0/2.0)
54 ACAO = INDENIZATORIA AND VALOR_RISCO = I2: OUTROS (6.0/2.0)
55 ACAO = INDENIZATORIA AND RISCO_ANTERIOR = REMOTO AND VALOR_PROVISAO = I1 AND REGIAO = SUL: IMPROCEDENCIA (6.0/2.0)
56 ACAO = INDENIZATORIA AND RISCO_ANTERIOR = REMOTO AND VALOR_PROVISAO = I1: EXTINCAO_SEM_MERITO (11.0/2.0)
57 VALOR_RISCO = I3 AND REU = Companhia de Credito Financiamento e Investimento RCI Brasil: IMPROCEDENCIA (16.0/7.0)
58 ACAO = INDENIZATORIA: EXTINCAO_SEM_MERITO (8.0/3.0)
59 RISCO_ATUAL = REMOTO AND ACAO = TARIFA AND REGIAO = SUL: IMPROCEDENCIA (5.0/1.0)
60 RISCO_ATUAL = REMOTO AND OJ = JEC: EXTINCAO_SEM_MERITO (12.0/4.0)
61 REGIAO = NORDESTE AND RISCO_ANTERIOR = REMOTO: ACORDO (3.0/1.0)
62 RISCO_ANTERIOR = REMOTO AND REU = Companhia de Credito Financiamento e Investimento RCI Brasil: IMPROCEDENCIA (5.0/1.0) : OUTRO S (11.19/6.0)

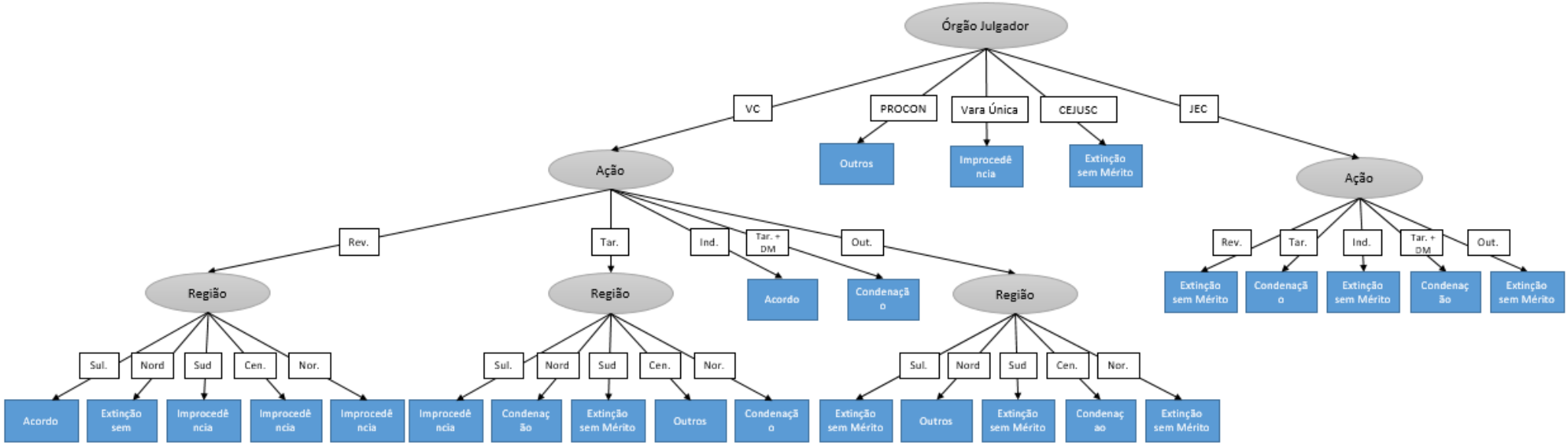
APÊNDICE C - Resultado Experimento 1 *Decision Table*

ACAO	OJ	RISCO_ATUAL	MOTIVO_ARQUIVAMENTO
OUTRAS	CEJUSC	REMOTO	EXTINCAO_SEM_MERITO
TARIFA E DANO MORAL	VARA_UNICA	REMOTO	EXTINCAO_SEM_MERITO
OUTRAS	PROCON	REMOTO	OUTROS
OUTRAS	PROCON	?	EXTINCAO_SEM_MERITO
TARIFA	PROCON	?	EXTINCAO_SEM_MERITO
TARIFA E DANO MORAL	VC	REMOTO	EXTINCAO_SEM_MERITO
OUTRAS	VC	REMOTO	EXTINCAO_SEM_MERITO
INDENIZATORIA	VC	REMOTO	EXTINCAO_SEM_MERITO
REVISIONAL	VC	REMOTO	IMPROCEDENCIA
TARIFA	VC	REMOTO	ACORDO
OUTRAS	CEJUSC	POSSIVEL	EXTINCAO_SEM_MERITO
OUTRAS	JEC	REMOTO	EXTINCAO_SEM_MERITO
OUTRAS	PROCON	PROVAVEL	EXTINCAO_SEM_MERITO
INDENIZATORIA	JEC	REMOTO	EXTINCAO_SEM_MERITO
REVISIONAL	JEC	REMOTO	IMPROCEDENCIA
TARIFA E DANO MORAL	JEC	REMOTO	IMPROCEDENCIA
TARIFA	JEC	REMOTO	IMPROCEDENCIA
TARIFA E DANO MORAL	VC	PROVAVEL	EXTINCAO_SEM_MERITO
INDENIZATORIA	VC	PROVAVEL	ACORDO
REVISIONAL	VARA_UNICA	POSSIVEL	EXTINCAO_SEM_MERITO
OUTRAS	VC	PROVAVEL	CONDENACAO
REVISIONAL	VC	PROVAVEL	CONDENACAO
TARIFA	VC	PROVAVEL	CONDENACAO
OUTRAS	JEC	PROVAVEL	EXTINCAO_SEM_MERITO
INDENIZATORIA	PROCON	POSSIVEL	EXTINCAO_SEM_MERITO
REVISIONAL	PROCON	POSSIVEL	EXTINCAO_SEM_MERITO
TARIFA	PROCON	POSSIVEL	OUTROS
INDENIZATORIA	JEC	PROVAVEL	CONDENACAO
OUTRAS	PROCON	POSSIVEL	OUTROS
REVISIONAL	JEC	PROVAVEL	CONDENACAO
TARIFA E DANO MORAL	JEC	PROVAVEL	CONDENACAO
TARIFA	JEC	PROVAVEL	CONDENACAO
INDENIZATORIA	VC	POSSIVEL	OUTROS
OUTRAS	VC	POSSIVEL	EXTINCAO_SEM_MERITO
TARIFA	VC	POSSIVEL	OUTROS
REVISIONAL	VC	POSSIVEL	EXTINCAO_SEM_MERITO
OUTRAS	JEC	POSSIVEL	EXTINCAO_SEM_MERITO
TARIFA E DANO MORAL	JEC	POSSIVEL	EXTINCAO_SEM_MERITO
INDENIZATORIA	JEC	POSSIVEL	EXTINCAO_SEM_MERITO
TARIFA	JEC	POSSIVEL	CONDENACAO
REVISIONAL	JEC	POSSIVEL	EXTINCAO_SEM_MERITO

APÊNDICE D - Resultado Experimento 2 *Decision Table*

ACAO	OJ	MOTIVO_ARQUIVAMENTO
OUTRAS	CEJUSC	EXTINCAO_SEM_MERITO
TARIFA E DANO MORAL	VARA_UNICA	EXTINCAO_SEM_MERITO
REVISIONAL	VARA_UNICA	EXTINCAO_SEM_MERITO
INDENIZATORIA	PROCON	EXTINCAO_SEM_MERITO
REVISIONAL	PROCON	EXTINCAO_SEM_MERITO
TARIFA	PROCON	OUTROS
OUTRAS	PROCON	OUTROS
TARIFA E DANO MORAL	VC	CONDENACAO
INDENIZATORIA	VC	ACORDO
OUTRAS	VC	EXTINCAO_SEM_MERITO
TARIFA	VC	IMPROCEDENCIA
REVISIONAL	VC	IMPROCEDENCIA
OUTRAS	JEC	EXTINCAO_SEM_MERITO
TARIFA E DANO MORAL	JEC	CONDENACAO
INDENIZATORIA	JEC	EXTINCAO_SEM_MERITO
TARIFA	JEC	CONDENACAO
REVISIONAL	JEC	EXTINCAO_SEM_MERITO

APÊNDICE E - Experimento 2 J48 – Árvore de decisão



APÊNDICE F - Resultado Experimento 1 REPTree

REPTree

==========

```
RISCO_ATUAL = POSSIVEL
|  OJ = JEC : EXTINCAO_SEM_MERITO (40/23) [18/11]
|  OJ = VC
|  |  ACAO = REVISIONAL : EXTINCAO_SEM_MERITO (6/2) [4/2]
|  |  ACAO = TARIFA : EXTINCAO_SEM_MERITO (5/3) [0/0]
|  |  ACAO = INDENIZATORIA : OUTROS (4/0) [0/0]
|  |  ACAO = TARIFA E DANO MORAL : EXTINCAO_SEM_MERITO (0/0) [0/0]
|  |  ACAO = OUTRAS : EXTINCAO_SEM_MERITO (4/2) [1/0]
|  OJ = PROCON : OUTROS (79.3/2) [40.3/0]
|  OJ = VARA_UNICA : OUTROS (0/0) [1/0]
|  OJ = CEJUSC : EXTINCAO_SEM_MERITO (1/0) [0/0]
RISCO_ATUAL = PROVAVEL
|  ACAO = REVISIONAL : CONDENACAO (20/7) [6/4]
|  ACAO = TARIFA : CONDENACAO (48.29/5.29) [27/5]
|  ACAO = INDENIZATORIA
|  |  VALOR_RISCO = I1 : ACORDO (15/7) [3/1]
|  |  VALOR_RISCO = I2 : CONDENACAO (1/0) [2/0]
|  |  VALOR_RISCO = I3 : CONDENACAO (10/0) [5/1]
|  ACAO = TARIFA E DANO MORAL : CONDENACAO (34/6) [21/2]
|  ACAO = OUTRAS : CONDENACAO (6/2) [3.29/1.29]
RISCO_ATUAL = REMOTO
|  ACAO = REVISIONAL
|  |  REGIAO = SUL : ACORDO (29/17) [12/5]
|  |  REGIAO = NORDESTE : IMPROCEDENCIA (8/4) [5/3]
|  |  REGIAO = SUDESTE : IMPROCEDENCIA (16/4) [7/3]
|  |  REGIAO = CENTRO-OESTE : IMPROCEDENCIA (9/4) [6/3]
|  |  REGIAO = NORTE : IMPROCEDENCIA (3/1) [0/0]
|  ACAO = TARIFA : IMPROCEDENCIA (42.41/12.41) [17/7]
|  ACAO = INDENIZATORIA
|  |  VALOR_RISCO = I1
|  |  |  REGIAO = SUL : IMPROCEDENCIA (5/2) [3/1]
|  |  |  REGIAO = NORDESTE : EXTINCAO_SEM_MERITO (11/4) [2/0]
|  |  |  REGIAO = SUDESTE
|  |  |  |  VALOR_CAUSA = I1 : IMPROCEDENCIA (5/2) [7/2]
|  |  |  |  VALOR_CAUSA = I3 : EXTINCAO_SEM_MERITO (2/0) [0/0]
|  |  |  |  VALOR_CAUSA = I2 : EXTINCAO_SEM_MERITO (1/0) [0/0]
|  |  |  REGIAO = CENTRO-OESTE : EXTINCAO_SEM_MERITO (3/1) [0/0]
|  |  |  REGIAO = NORTE : EXTINCAO_SEM_MERITO (2/1) [0/0]
|  |  VALOR_RISCO = I2 : OUTROS (6/3) [1/0]
|  |  VALOR_RISCO = I3 : EXTINCAO_SEM_MERITO (8/3) [7/4]
|  ACAO = TARIFA E DANO MORAL : IMPROCEDENCIA (23/9) [16/6]
|  ACAO = OUTRAS : EXTINCAO_SEM_MERITO (20/11) [19.41/9.41]
```

APÊNDICE G - Resultado Experimento 2 REPTree

```
REPTree
============

OJ = JEC
|   ACAO = REVISIONAL : EXTINCAO_SEM_MERITO (21/12) [8/3]
|   ACAO = TARIFA : CONDENACAO (95/48) [45/20]
|   ACAO = INDENIZATORIA
|   |   REGIAO = SUL : CONDENACAO (18/13) [10/7]
|   |   REGIAO = NORDESTE : EXTINCAO_SEM_MERITO (20/8) [5/3]
|   |   REGIAO = SUDESTE : IMPROCEDENCIA (18/15) [11/5]
|   |   REGIAO = CENTRO-OESTE : EXTINCAO_SEM_MERITO (11/6) [3/3]
|   |   REGIAO = NORTE : EXTINCAO_SEM_MERITO (5/4) [1/0]
|   ACAO = TARIFA E DANO MORAL : CONDENACAO (61/32) [42/22]
|   ACAO = OUTRAS : EXTINCAO_SEM_MERITO (13/6) [6/2]
OJ = VC
|   ACAO = REVISIONAL
|   |   REGIAO = SUL : ACORDO (29/17) [11/4]
|   |   REGIAO = NORDESTE : EXTINCAO_SEM_MERITO (8/5) [7/5]
|   |   REGIAO = SUDESTE : IMPROCEDENCIA (20/11) [7/3]
|   |   REGIAO = CENTRO-OESTE : IMPROCEDENCIA (12/7) [8/5]
|   |   REGIAO = NORTE : IMPROCEDENCIA (3/1) [0/0]
|   ACAO = TARIFA
|   |   REGIAO = SUL : IMPROCEDENCIA (4/2) [0/0]
|   |   REGIAO = NORDESTE : CONDENACAO (2/1) [2/0]
|   |   REGIAO = SUDESTE : EXTINCAO_SEM_MERITO (5/2) [4/4]
|   |   REGIAO = CENTRO-OESTE : OUTROS (3/1) [1/0]
|   |   REGIAO = NORTE : CONDENACAO (0/0) [0/0]
|   ACAO = INDENIZATORIA : ACORDO (17/12) [4/2]
|   ACAO = TARIFA E DANO MORAL : CONDENACAO (2/1) [0/0]
|   ACAO = OUTRAS : EXTINCAO_SEM_MERITO (15/11) [14/8]
OJ = PROCON : OUTROS (83/2) [43/0]
OJ = VARA_UNICA : IMPROCEDENCIA (1/0) [1/1]
OJ = CEJUSC : EXTINCAO_SEM_MERITO (1/0) [1/0]
```

APÊNDICE H - Experimento 2 REPTree – Árvore de decisão