UNIVERSIDADE FEDERAL DO PAMPA

THALES VAZ MACIEL

DESCOBERTA DE CONHECIMENTO COM MINERAÇÃO DE DADOS APLICADA À OVINOCULTURA

Bagé
2013

THALES VAZ MACIEL

DESCOBERTA DE CONHECIMENTO COM MINERAÇÃO DE DADOS APLICADA À OVINOCULTURA

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Sistemas Distribuídos com Ênfase em Banco de Dados da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Especialista.

Orientador: Prof. Dr. Sandro da Silva Camargo

Coorientador: Prof. Dr. Milton Roberto Heinen

Bagé
2013

THALES VAZ MACIEL

DESCOBERTA DE CONHECIMENTO COM MINERAÇÃO DE DADOS APLICADA À OVINOCULTURA

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Sistemas Distribuídos com Ênfase em Banco de Dados da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Especialista.

Trabalho de Conclusão de Curso defendido e aprovado em: 8 de agosto de 2013.

Banca examinadora:

_____

Prof. Dr. Sandro da Silva Camargo
Orientador
UNIPAMPA

_____

Prof. Dr. Milton Roberto Heinen
UNIPAMPA

_____

Prof. MSc. Sandra Dutra Piovesan
UNIPAMPA

Dedico este trabalho, primeiramente, a Deus, por iluminar meu caminho com diversas oportunidades. Dedico à minha família, mas, em especial, ao meu filho, Thael, luz na minha vida e, também, neste momento, ao meu avô, Ten. Vaz, de eterna influência no meu ser.

Dedico a todos meus entes queridos que entendem por quê eu fico "na frente do computador".

RESUMO

Atualmente, as empresas tendem à manter grande e crescente volume de dados. Eventualmente, estes conjuntos de dados podem chegar à quantidades massivas e este cenário é conveniente para a descoberta de conhecimento em bancos de dados. Neste contexto, a mineração de dados é aplicada através de suas funcionalidades e implementações de algoritmos especializados em tais propósitos. Este trabalho versa sobre os resultados de testes de descoberta de conhecimento em ovinocultura com a aplicação de mineração de dados, utilizando pré-processamento do conjunto de dados em estudo em favor de cada experimento em específico. Atividades de associação, classificação e *clustering* foram desempenhadas com os algoritmos *Apriori*, 1R, C4.5, CART e *k-means* através da ferramenta *Waikato Environment for Knowledge Analysis* (WEKA) e foram descobertos indícios de atividade fora de padrão, assim como informações potencialmente relevantes para o domínio de negócio da ovinocultura.

Palavras-chave: mineração de dados, descoberta de conhecimento, ovinocultura, KDD, WEKA.

ABSTRACT

Nowadays, companies tend to maintain great and growing volumes of data. Eventually, these datasets may reach massive amounts and this scenario is convenient for the discovery of knowledge in databases. In this context, data mining is applied through its functionalities and algorithmic implementations that are specific for that purpose. This research is about the results of knowledge discovery tests in sheep breeding with the application of data mining, utilizing pre-processing of the provided dataset in each specific experiment. Association, classification and clustering tasks were performed with the 1R, C4.5, CART and k-means algorithms with the Waikato Environment for Knowledge Analysis (WEKA) tool and enabled the discovery of out-of-pattern activity evidence, as well as information that is potentially relevant for the sheep breeding business domain.

Keywords: data mining, discovery of knowledge, sheep breeding, KDD, WEKA.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

API – *Application Programming Interface*

ARCO – Assistência aos Rebanhos de Criadores de Ovinos

ARFF – *Attribute-Relation File Format*

CART – *Classification and Regression Trees*

CLI – *Command-Line Interface*

CSV – *Comma-Separated Values*

GNU – *GNU is Not Unix*

IDE – *Integrated Development Environment*

KDD – *Knowledge Discovery from Data*

MAPA – Ministério da Agricultura, Pecuária e Abastecimento

PHP – *PHP: Hypertext Preprocessor*

SGBD – Sistema Gerenciador de Banco de Dados

SQL – *Structured Query Language*

WEKA – *Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

A mineração de dados trata, primariamente, da descoberta de conhecimento em grandes volumes de dados. A criação regulamentada de ovinos é uma atividade de pleno potencial para geração de serviços que, por sua vez geram grande tráfego e controle de dados. No Brasil a atividade de criação de ovinos é regulamentada e executada pela Associação Brasileira de Criadores de Ovinos (ARCO).

De acordo com o Regulamento do Registro Genealógico de Ovinos no Brasil (ARCO, 2012), o domínio deste negócio deve realizar como requisito, dentro outras atividades, a comunicação, análise, processamento e arquivamento de dados.

Han, Kamber e Pei (2009) explicam que a mineração de dados é uma tecnologia bastante abrangente no que se refere à viabilidade de sua aplicação e o único requisito apresentado, para tal, é que a fonte de dados em estudo tenha significado em relação à aplicação em domínio específico ou tenha relevância para este.

Além disto, entende-se que os conjuntos de dados mais frequentemente utilizados em mineração de dados são provenientes de bancos de dados reais, ou seja, de domínios de aplicações reais. Portanto, percebe-se uma situação propícia para a descoberta de conhecimento com mineração de dados no domínio da criação de ovinos.

1.1 Objetivos

- Geral:
    - Descobrir conhecimento em conjuntos de dados oriundos do domínio da ovinocultura com o auxílio da mineração de dados para obter sugestões de melhorias no processo de criação de ovinos.
- Específicos:
    - Buscar conhecimento novo que tenha aplicabilidade prática à ovinocultura;
    - Tratar o conjunto de dados desde sua fonte até torná-la passível da execução de atividades de mineração de dados;
    - Estudar a ferramenta WEKA e suas funcionalidades, de modo a utilizá-la no desempenho do objetivo geral;
    - Aplicar funcionalidades de mineração de dados, sendo capaz de interpretar os dados colhidos, de maneira à gerar conhecimento novo.

1.2 Metodologia

Este trabalho foi desenvolvido a partir da revisão bibliográfica sobre descoberta de conhecimento, mineração de dados e suas funcionalidades. Foi, também, estudado o domínio de negócio da ovinocultura, assim como o conjunto de dados, conforme disponibilizado e as utilidades disponibilizadas pela ferramenta WEKA.

Foram desenvolvidas projeções do conjunto de dados em estudo e, em cinco experimentos distintos, foram aplicados algoritmos de associação, classificação e *clustering* sobre estas projeções para fins de geração de artefatos de saída dos algoritmos, interpretação dos resultados e mineração de conhecimento.

1.3 Estrutura do Trabalho

Após esta introdução ao presente trabalho, é apresentada uma revisão bibliográfica sobre descoberta de conhecimento com mineração de dados no Capítulo 2. O Capítulo 3 trata da mineração de dados, onde suas funcionalidades são estudadas e abordagens algorítmicas são apresentadas. A ferramenta de mineração de dados utilizada neste estudo é apresentada no Capítulo 4, onde são brevemente consideradas as aplicações de seus quatro ambientes.

O Capítulo 5, após revisar as regras de negócio inerentes à ovinocultura no Brasil, versa sobre os experimentos de mineração de dados efetuados neste estudo e, finalmente, o Capítulo 6 apresenta conclusões sobre eles e as considerações finais sobre o presente estudo.

2 DESCOBERTA DE CONHECIMENTO COM MINERAÇÃO DE DADOS

De acordo com Han, Kamber e Pei (2009), "*knowledge discovery from data*" (KDD), ou descoberta de conhecimento a partir de dados, é frequentemente tratada como sinônimo de outro termo, também utilizado no mesmo contexto. Porém, a mineração de dados, pode ser considerada, simplesmente, uma fase essencial da descoberta de conhecimento.

Sumathi e Esakkirajan (2006) explicam que a mineração de dados, como descoberta de conhecimento, é referente à extração de conhecimento a partir de grandes quantidades de dados. Neste contexto, também é conhecida como mineração de conhecimento, mineração de conhecimento a partir de dados, extração de conhecimento, análise de dados e padrões e, até mesmo, arqueologia de dados. Estes autores deixam explícito que o conceito de mineração de dados existe no sentido amplo e no sentido estrito. No seu sentido amplo, tem significação referente à descoberta de conhecimento como um processo completo, admitindo, portanto, todas as etapas envolvidas. No seu sentido estrito, por outro lado, representa apenas uma das etapas, entretanto essencial, do processo de descoberta de conhecimento, onde métodos de inteligência artificial e estatística são aplicados, tendo seu escopo de definição mais limitado em relação ao seu sentido amplo.

De maneira geral, existe a definição das etapas de pré-processamento de dados, antecedente à mineração de dados. Após a mineração de dados acorre a avaliação dos padrões descobertos e a apresentação de conhecimento, como modelo do processo de descoberta de conhecimento.

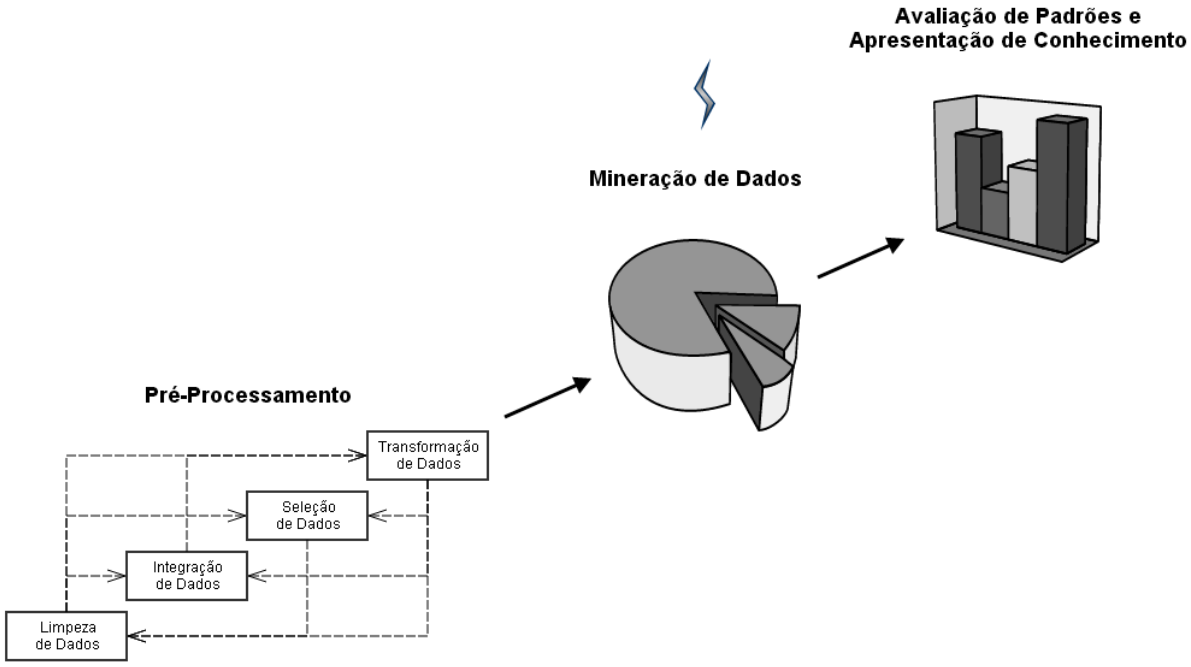
Na obra de Sumathi e Esakkirajan (2006), é exposto que dados armazenados em bancos de dados podem conter o que é chamado de "ruído", casos onde podem ser encontrados dados não esperados ou, até mesmo, objetos ou tuplas incompletas. Quando na atividade de mineração de dados, estas ocorrências podem confundir o processo e, portanto, levando à construção errônea de modelos de dados. Como resultado, a qualidade na descoberta de padrões pode ser pequena.

Enquanto Ramakrishnan e Gehrke (2000) tratam a descoberta de conhecimento como um processo cuja divisão e especificação de etapas pode ocorrer de forma homogênea e confusa, na abordagem proposta por Han, Kamber e Pei (2009), o processo é, efetivamente, apresentado em etapas com escopo bem definido, conforme a sequência apresentada abaixo.

1. Limpeza de Dados;
2. Integração de Dados;
3. Seleção de Dados;
4. Transformação de Dados;
5. Mineração de Dados;
6. Avaliação de Padrões;
7. Apresentação de Conhecimento.

A figura 1, abaixo, ilustra a evolução entre as fases apresentadas.

Figura 1 - Ilustração do processo de descoberta de conhecimento com mineração de dados.



Fonte: Dados primários.

Devido à definição coerente no escopo de cada etapa do processo, esta abordagem é adotada para fins de aplicação no presente estudo.

2.1 Limpeza de Dados

A limpeza de dados, conforme Han, Kamber e Pei (2009) trata da remoção de dados inconsistentes do conjunto de dados à ser utilizado para fins da descoberta de conhecimento.

Date (2004) explica que não é incomum que fontes de dados tenham controle de qualidade inadequado e, como resultado disto, é observada a necessidade de avaliação, adequação e filtragem do conjunto de dados em estudo. Esta etapa é, geralmente, realizada com o auxílio de processos em lote, também conhecidos como *batch*, no pré-processamento dos dados, antecedendo sua utilização na etapa de mineração de dados.

Operações típicas de limpeza de dados podem incluir o preenchimento de valores ausentes, a correção de erros tipográficos, a definição de abreviaturas e formatação de valores (de acordo com o que seja julgado conveniente), e a substituição de sinônimos por identificadores padrão.

Além disto, dados que são julgados inválidos, ou seja, não considerados passíveis de correção, são descartados. Informações obtidas durante o processo de limpeza dos dados têm o potencial de auxiliar na identificação da causa de erros em outras etapas de pré-processamento na descoberta de conhecimento, possibilitando o melhoramento do processo.

Ramakrishnan e Gehrke (2000) expõem a limpeza de dados como uma etapa que, além de realizar atividades conforme descrito anteriormente, teria, em seu escopo, a transformação de valores de dados, a geração de novos valores, calculados com base em valores existentes, a integração dos dados em um esquema relacional e, até mesmo, a desnormalização de dados, em função de relacionamentos aparentes. Este tipo de agrupamento de atividades não é validado em abordagens apresentadas por outros autores, como Han, Kamber e Pei (2009), onde as fases de integração, seleção e transformação de dados não são agrupadas em uma única fase, justamente pela confusão que isto pode causar.

Considerando que dados tipicamente disponibilizados para a atividade da mineração de dados são associados à aplicações de *software* de domínios específicos de negócio e/ou conhecimento, entende-se que a ocorrência da inconsistência nestes dados em sua forma bruta, conforme disponibilizados, se deve à múltiplos fatores. Dentre eles, inconsistências ou falhas no próprio produto de *software* que alimenta a fonte de dados e/ou, até mesmo, atividades de natureza maliciosa e/ou fraudulenta por parte do elemento humano que age sobre o elemento de *software*, tanto sob a função de usuário ou na função de desenvolvimento do mesmo.

2.2 Integração de Dados

Conforme Han, Kamber e Pei (2009), a integração de dados é uma fase onde, no caso da existência de múltiplas fontes de dados, estas são integradas, tendo seus dados dispostos em formato consolidado e unificado.

De acordo com Date (2004), a integração de dados ou consolidação, como esta etapa também é conhecida, é um conceito inerente à prática de *data warehousing*, onde a prática da consolidação de múltiplos conjuntos de dados é realizada por ferramentas utilizadas no processo de apoio à tomada de decisão. Nesta situação, quaisquer relacionamentos implícitos em quaisquer conjuntos de dados, ou dados de diferentes conjuntos de dados, devem ser feitos explícitos. Isto pode acontecer através da posição de valores de dados em lugar de relacionamentos gerenciados por um sistema de controle, como um sistema gerenciador de bancos de dados (SGBD).

Uma atividade que pode fazer parte da integração de dados, ou consolidação, é a sincronização temporal (DATE, 2004), pois não, necessariamente, os processos executados dentro do domínio de um negócio e/ou aplicação irão armazenar dados temporais na mesma escala ou precisão. Enquanto, por exemplo, para um processo de negócio, a recuperação de dados em formato de data com indicadores de dia, mês e ano é relevante, no caso de outro processo qualquer, somente o armazenamento do ano pode ser suficiente para sua execução. Isto pode trazer discrepâncias de processamento de dados e é onde a sincronização temporal deve ser aplicada como parte da etapa de consolidação. Entende-se que, dependentemente da complexidade dos ajustes à serem desempenhados, tal atividade passe a não fazer parte do escopo da integração de dados, devendo, portanto, ser executada durante a etapa de transformação de dados.

Além disto, nesta etapa, da mesma maneira como na etapa de limpeza de dados, dados considerados impassíveis de correção podem ser descartados e informações obtidas como parte da etapa de integração de dados podem ser utilizadas no melhoramento do processo nas etapas subsequentes.

2.3 Seleção de Dados

A seleção de dados, de acordo com Ramakrishnan e Gehrke (2000), como pré-processamento de dados, é a etapa no processo de descoberta de conhecimento onde é selecionado um subconjunto de dados, de acordo com sua relevância para a tarefa de mineração de dados a ser executada, ou seja, são selecionadas as instâncias e atributos a serem analisados durante o processo.

2.4 Transformação de Dados

De acordo com Han, Kamber e Pei (2009), fase de transformação de dados tem em seu escopo de realização a consolidação dos dados em formatos devidamente apropriados para a atividade de mineração de dados a ser aplicada. Pode ser considerada necessidade de executar operações de síntese (resumo) e/ou agregação dos dados presentes no conjunto de dados em estudo, e isto pode acontecer mesmo antes da fase de seleção de dados.

Witten, Frank e Hall (2011) entendem que o sucesso em atividades de mineração de dados é envolve mais do que a simples seleção de um algoritmo e sua aplicação sobre um conjunto de dados, justificando, assim, a execução deste tipo de atividade.

Existe uma atividade que não é entendida como uma fase adicional ao processo de descoberta de conhecimento, mas parte da fase de transformação de dados e de natureza opcional (dependentemente da tarefa de mineração de dados a ser executada e da dimensão do conjunto de dados identificado como relevante para o estudo). Trata-se da redução de dados (HAN, KAMBER e PEI, 2009), onde, sem alterar a integridade deste conjunto de dados, pode ser obtida uma representação em menor escala do mesmo, caso seja considerado apropriado.

Além da redução de dados, conforme mencionado anteriormente, atividades de sincronização temporal, também consideradas na etapa de integração de dados, podem ser aplicadas sobre o conjunto de dados nesta transformação de dados, caso tal sincronização represente complexidade em escala suficiente para estar caracterizada no escopo da transformação de dados.

2.5 Mineração de Dados

Entende-se que as fases de limpeza, integração, seleção e transformação de dados, conforme apresentadas anteriormente, são realizadas na fase de pré-processamento do conjunto de dados utilizados no estudo da descoberta de conhecimento.

De acordo com Date (2004), a mineração de dados pode ser descrita como uma análise exploratória de um conjunto de dados, onde o objetivo é a descoberta de padrões nestes. Uma vez que determinada a relevância dos padrões descobertos, estes podem ser utilizados como base para tomada de decisão em nível estratégico em uma organização ou, até mesmo, a identificação de atividades não convencionais em um sistema.

Entende-se que tendo como entrada um conjunto de dados livre de inconsistências, produto da consolidação de múltiplas fontes de dados, relevantes para a tarefa a ser

executada, além de resumidos e agregados ou reduzidos a uma dimensão apropriada, são executadas tarefas de mineração de dados. Algoritmos baseados em métodos de inteligência artificial e estatísticos são aplicados com o propósito da identificação de padrões para que estes sejam utilizados como entrada à próxima fase da descoberta de conhecimento, justamente, a avaliação desses padrões.

Conforme exposto posteriormente, a etapa de mineração de dados é exposta em mais detalhes no capítulo 3.

2.6 Avaliação de Padrões

A fase de avaliação de padrões, posterior à descoberta de padrões, o que, sistematicamente, ocorre na fase de mineração de dados, trata da classificação dos padrões descobertos, de forma que sejam identificados os padrões que são de interesse ou que representem relevância para estudo em determinado domínio de conhecimento, aplicação ou negócio (HAN, KAMBER e PEI, 2009).

2.7 Apresentação de Conhecimento

A última fase do processo de descoberta de conhecimento, conforme apresentado anteriormente, é a apresentação de conhecimento, onde são utilizadas técnicas de representação de conhecimento para efetuar a apresentação dos padrões e conhecimento descobertos durante a mineração de dados e posteriormente avaliados. Esta apresentação representa a visualização, pelo usuário, dos resultados obtidos. Uma alternativa seria a alimentação de uma base de conhecimento.

De acordo com Sumathi e Esakkirajan (2006), a descoberta de conhecimento deve ser expressa em linguagens de alto nível, representações visuais ou outras formas alternativas e expressivas, para que o conhecimento possa ser facilmente entendido e diretamente utilizado por humanos. Isto é especialmente crucial quando o processo de mineração de dados toma natureza interativa com os usuários.

Este capítulo apresentou a descoberta de conhecimento como um processo de maior complexidade em relação à mineração de dados, sendo esta apenas uma fase do processo. Foram, juntamente, citadas etapas referentes à pré-processamento de dados e pós-processamento que também são utilizadas na descoberta por conhecimento a partir de dados.

O capítulo 3 trata da mineração de dados em bancos de dados dentro deste escopo, como uma etapa de um processo bem definido, apresentando o quê sua execução propõe, especificamente.

3 MINERAÇÃO DE DADOS EM BANCOS DE DADOS

Conforme Ramakrishnan e Gehrke (2000), a mineração de dados consiste na descoberta de tendências ou padrões de dados em grandes conjuntos de dados. Neste contexto, existe o pressuposto de que ferramentas de mineração de dados sejam capazes de identificar estes padrões com mínima intervenção dos usuários.

Entende-se que a característica mais importante e o diferencial da mineração de dados, em relação a outras áreas de pesquisa semelhantes, é que o volume de dados à ser analisado é, de fato, de quantidade massiva. Conforme exposto anteriormente, ferramentas de mineração de dados são projetadas para realizar a aplicação de técnicas de natureza estatística sobre quantidades massivas de dados em busca de determinados padrões.

Rajaraman, Leskovec e Ullman (2013) explicam uma definição bastante aceita neste contexto, onde a mineração de dados é tratada como a descoberta de "modelos" para os dados. A construção destes "modelos" pode seguir estratégias oriundas das áreas de modelagem estatística, aprendizado de máquina, e/ou outras abordagens computacionais.

De acordo com estes autores, a mineração de dados baseada em modelagem estatística era considerada, originalmente, um termo depreciativo referente à tentativas de extração de informação que não era suportada pelos dados em seu formato disponível, onde irregularidades eram apresentadas devido à limitações da própria natureza estatística. Atualmente, com uma mudança positiva nesta perspectiva, a utilização de modelagem estatística na mineração de dados trata, especificamente, da construção de modelos estatísticos, ou seja, a consideração de modelos de dados não explícitos, a partir dos quais os dados explícitos tenham sido gerados ou sejam fundamentados.

O aprendizado de máquina, apesar de ser, por muitos, considerado um sinônimo à mineração de dados, de acordo com Rajaraman, Leskovec e Ullman (2013), não é um conceito adequado, pois o fato é que algumas formas de mineração de dados, efetivamente, utilizam algoritmos provenientes da área de aprendizado de máquina, mas não exclusivamente. Além disto, a interpretação destes autores é que a utilização de aprendizado de máquina não deve ser considerada uma solução global na mineração de dados, pois a razão em sua utilização depende no conhecimento dos usuários sobre o seu objetivo na busca de padrões no conjunto de dados. Entende-se que o aprendizado de máquina é uma abordagem eficiente à mineração de dados quando se tem pouco conhecimento dos padrões a serem descobertos. Portanto, não tem obtido os mesmos resultados positivos em situações cujos objetivos são descritos de maneira mais clara e direta. Isto se deve ao fato de não haver vantagem em desempenho na utilização de

aprendizado de máquina sobre algoritmos de projeto específico para a realização de uma determinada tarefa quando se tem conhecimento dos objetivos próprios da mesma.

Ramakrishnan e Gehrke (2000) explicam que a mineração de dados não é, somente, relacionada com a análise exploratória de dados, classificando esta como uma subárea da estatística que, apesar de possuir objetivos bastante similares aos da mineração de dados, é puramente fundamentada em metodologias estatísticas. De acordo com estes autores, esta etapa é, também, relacionada com duas subáreas da inteligência artificial, sendo elas a descoberta de conhecimento e o aprendizado de máquina, tratando a mineração de dados como um contexto multidisciplinar.

Rajaraman, Leskovec e Ullman (2013), explicam que a maioria das abordagens computacionais à criação de modelos de dados trata ou do resumo dos dados, de forma sucinta e aproximada ou da extração de características mais salientes e o descarte posterior do que não estiver contido no conjunto obtido. O resumo de dados trata da possibilidade de utilizar um único valor para resumir a relevância de um item dentro de um determinado contexto. Um exemplo disto é a tarefa de análise de agrupamentos (c*lustering*), onde, conforme é exposto à seguir, o resumo dos dados pode ser visualizado como entidades posicionadas dentro de um plano multidimensional. Entidades posicionadas em proximidade umas das outras podem ser consideradas parte do mesmo agrupamento, compartilhando, portanto atributos considerados relevantes. Estes agrupamentos podem representar um resumo do conjunto de dados por inteiro.

No que se trata da extração de características de dados, segundo estes autores, a procura por modelos de dados ocorre em função da busca de ocorrências mais extremas, ou seja, determinados fenômenos ou anomalias no conjunto de dados. Este é, então, representado por estes exemplos, como modelos.

Algumas abordagens importantes à extração de características de conjuntos de dados em larga escala são a busca por conjuntos de itens frequentes e a busca por itens similares. Em definições breves, a busca por conjuntos de itens frequentes é feita a busca por itens que estejam relacionados em conjunto dentro de determinada transação ou processo. A busca por itens similares, também conhecida como filtragem colaborativa, tem o objetivo de identificar conjuntos de itens que tenham elementos em comum com a premissa de que isto venha a ser associado à semelhanças entre os itens.

De acordo com Han, Kamber e Pei (2009), as tarefas de mineração de dados são executadas através funcionalidades de mineração de dados, que são responsáveis pela

determinação dos tipos e os próprios padrões a serem encontrados durante a execução de uma tarefa desta natureza.

3.1 Funcionalidades de Mineração de Dados

Em relação a funcionalidades de mineração de dados, nota-se que, dentre os autores desta área, nem todos utilizam algum tipo de classificação ou taxonomia para tais funcionalidades.

Han, Kamber e Pei (2009) incluem na lista de funcionalidades de mineração de dados, descritas em sua obra, a caracterização e discriminação, a descobertas de padrões frequentes, associações e correlações, classificação e regressão, análise de agrupamentos (*clustering*) e a análise de casos isolados. Estes autores descrevem uma taxonomia sobre estas funcionalidades, onde elas podem ser categorizadas em descritivas ou preditivas. Tarefas descritivas caracterizam propriedades dos dados em um determinado conjunto de dados. Tarefas preditivas, por sua vez, desempenham a indução sobre os dados em análise para a predição de dados futuros.

De acordo com Sumathi e Esakkirajan (2006), as funcionalidade de mineração de dados são utilizadas para determinar os tipos de padrões à serem descobertos nas tarefas de mineração de dados. Segundo estes autores, estas funcionalidades podem ser classificadas em duas categorias, sendo descritivas ou preditivas e, ainda tipificadas como caracterização e discriminação, associação, classificação e predição, *clustering*, análise evolutiva e análise de casos isolados, mostrando uma abordagem bastante similar à de Han, Kamber e Pai (2006), conforme exposto anteriormente.

Ramakrishnan e Gehrke (2000), em sua obra, versam sobre tarefas de mineração de dados baseadas nas funcionalidades de contagem de co-ocorrências, procura por regras de associação (também para tarefas de predição de dados) e padrões sequenciais, classificação e regressão, estruturação de árvores de decisão, e a funcionalidade de *clustering*, de forma individualizada.

Rajaraman, Leskovec e Ullman (2013) abordam, de forma diversificada, técnicas para identificação de entidades similares, identificação conjuntos de dados frequentes, e a funcionalidade de *clustering*, mas sem uma apresentação explícita de uma taxonomia para estas.

Camargo (2002), por sua vez, cita, como referência de tipologia de tarefas de mineração de dados, Berry e Linnof (1997), cuja obra descreve seis grupos distintos, sendo eles a classificação, regressão, predição, regras de associação, agrupamento por

similaridades, e descrição. Também é apresentada outra taxonomia, oriunda da obra de Agrawal e Shafer (1996), onde tais tarefas são relacionadas à uma de três classes, sendo elas a classificação (compreendendo as tarefas de classificação, regressão e agrupamento por similaridades, por exemplo), associação e sequências (semelhante à predição).

Este autor ainda expõe que diferentes técnicas podem ser utilizadas nestas atividades de mineração de dados, tendo por exemplo a construção de árvores de decisão, utilização de redes neurais, raciocínio baseado em memória e algoritmos de análise genética.

Para finalidade deste estudo, foram aplicadas e, portanto, mais criteriosamente estudadas as funcionalidades de mineração de dados de associação, classificação e análise de agrupamentos (c*lustering*).

3.1.1 Associação

Han, Kamber e Pei (2009) definem a mineração de dados pela funcionalidade de associação como a descoberta de conjuntos de elementos frequentes, sendo que estes satisfazem um nível mínimo de suporte ou percentagem relevante das tuplas provenientes de um banco de dados, a partir dos quais são geradas regras de associação.

Segundo Date (2004), o conjunto de dados utilizado na mineração de dados por associação, em sua totalidade, é denominado população. As regras de associação identificadas pelos algoritmos executados em cada tarefa possuem duas propriedades básicas, sendo elas os indicadores de suporte e de confiança.

O indicador de suporte é equivalente à percentagem de instâncias pertencentes à população que satisfazem um determinado requisito. O indicador de confiança, por sua vez, considerando que a natureza da mineração de regras de associação trata, justamente, da relação entre, no mínimo, dois elementos, revela a percentagem de regras suportadas em que a associação entre os elementos é, efetivamente, satisfeita.

De acordo com Han, Kamber e Pei (2009), regras de associação que satisfazem um determinado valor mínimo para os indicadores de suporte e confiança são consideradas fortes. Além disto, estes autores consideram que a mineração de dados pela busca de regras de associação pode ser dimensionada à busca de conjuntos de elementos frequentes, sendo dividida em duas etapas, conforme abaixo.

1. A descoberta de todos os conjuntos de elementos frequentes, onde cada conjunto de elementos deve ocorrer, no mínimo, tão frequentemente quanto o valor definido para suportar as regras de associação; e

2. A geração de regras de associação fortes a partir dos conjuntos frequentes de dados descobertos previamente, onde, por definição, estas regras devem satisfazer, além do indicador mínimo de suporte, também o indicador mínimo de confiança.

Camargo (2002) expõe a descoberta de regras de associação como um processo de agrupamento por afinidade, este, consistido na determinação de elementos que ocorrem simultaneamente em um conjunto de transações, além de que as associações ocorrem de maneira que a presença de determinado(s) elemento(s) em uma determinada transação implica(m) na presença de outro(s) elemento(s) na mesma transação.

Este autor ainda adiciona que, conceitualmente, a diferença entre os indicadores de suporte e confiança está no fato de que enquanto a confiança é uma medida de força da regra, o suporte tem relevância puramente estatística. Além disto, uma motivação para limitação do nível de suporte vem do fato de que, geralmente, há interesse em regras com valor de suporte superior à um determinado nível por razões provenientes do domínio das regras de negócio em estudo. Portanto, caso o indicador de suporte apresentado por uma determinada regra não for alto o suficiente, isto pode significar que a regra em questão pode ser desconsiderada por irrelevância. Entende-se que a prática da eliminação de regras consideradas irrelevantes é comum, visando a otimização do tempo de execução do processo de mineração de dados.

De acordo com Sumathi e Esakkirajan (2006), a descoberta de regras de associação interessantes entre grandes quantidades de dados provenientes de um domínio de negócio pode ajudar no processo de decisões inerentes à projeto de catálogos, formação de combinações para comercialização de produtos ou simples análise de circunstâncias e consequências.

3.1.2 Classificação

Camargo (2002) trata da funcionalidade de classificação em mineração de dados como o exame das características de um novo objeto e sua alocação em uma classe de um conjunto de classes previamente definidas, com base nestas características. Os objetos a serem classificados podem ser representados por tuplas em uma projeção de banco de dados. As características de um objeto são representadas pelos campos de uma tupla e a ação de classificação consiste na atualização de cada registro pelo preenchimento de um determinado campo com o indicador da classe a qual este objeto pertence.

Ganti, Gehrke e Ramakrishnan (1999) explicam que esta funcionalidade de mineração de dados utiliza dados referentes ao histórico de elementos do domínio do negócio em estudo para predizer a classificação de novos elementos. Para esta proposta, é requerida a utilização de um conjunto de dados, chamado de conjunto de dados de treinamento, onde estão contidos diversos atributos, sendo que, destes, um determinado atributo é chamado de atributo dependente, e os outros de atributos preditores. O objetivo, a partir da execução deste treinamento, é a criação de modelos que, baseando-se nos atributos preditores como entrada, sejam capazes de gerar um valor para o atributo dependente como saída ao analisar novos elementos.

Em acordo com esta abordagem, Camargo (2002) versa sobre a tarefa de classificação como a utilização de um conjunto de classes bem definidas e um conjunto de dados para treinamento de exemplos pré-classificados. Também exemplifica técnicas bastante eficazes na atividade de classificação, sendo elas a utilização de árvores de decisão, redes neurais, assim como raciocínio baseado em memória e análise de ligação, em alguns casos.

Ganti, Gehrke e Ramakrishnan (1999) acrescentam que caso o atributo determinado como dependente na atividade de classificação seja de natureza numérica, a tarefa é tratada como um problema de regressão. A tarefa é tratada como um problema de classificação em casos onde o atributo dependente é de natureza nominal ou categórica, e, então, os valores definidos como domínio deste atributo são chamados de rótulos das classes.

Em trabalho posterior, Ramakrishnan e Gehrke (2000) reafirmam esta mesma abordagem e ainda explicam que todos os atributos dos elementos, não somente o atributo dependente, são distinguidos em dois tipos, numéricos ou categóricos. Para atributos numéricos, podem ser executadas operações computacionais para processamento numérico, como a computação de valores para a média entre valores, enquanto, para atributos categóricos, seus domínios são definidos em um conjunto de valores para cada um deles. No que trata das regras geradas pela atividade, são geradas regras de classificação quando o atributo dependente é de natureza categórica, e são geradas regras de regressão quando o atributo dependente é de natureza numérica.

Abaixo, o quadro 1 demonstra um exemplo de expressão formal para uma regra de classificação fictícia, onde é determinado alto risco de segurança em trânsito de automóveis, caso o motorista tenha idade entre 16 e 25 anos e o tipo de carro seja esportivo ou caminhão.

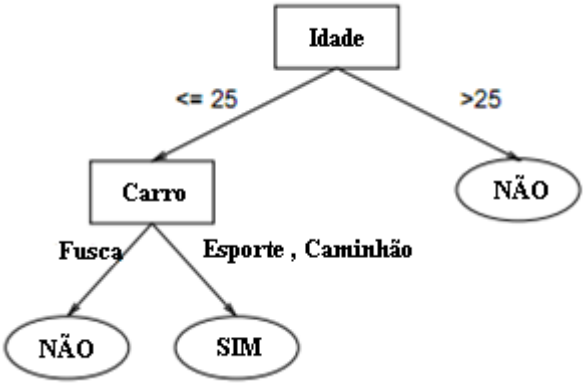
Quadro 1 - Exemplo de regra de classificação.

```
(16 <= idade <= 25) ^ (tipo_de_carro ∈ { esporte, caminhão} ) alto_risco = sim
```

Fonte: Adaptado de Ramakrishnan e Gehrke (2000).

A figura 2 demonstra a mesma regra, mas em notação gráfica, em formato de árvore de decisão, para fins de exemplificação.

Figura 2 - Exemplo de árvore de decisão.



Fonte: Adaptado de Ramakrishnan e Gehrke (2000).

Conforme Ramakrishnan e Gehrke (2000), assim como em regras de associação, níveis de suporte e confiança podem ser definidos para regras de classificação e regressão. Além disto, é também expresso que as regras de classificação e regressão diferem das regras de associação por considerarem valores contínuos (numéricos) e valores categóricos e, não apenas um campo cujo domínio é estritamente definido.

Witten, Frank e Hall (2011) explicam que dois artefatos podem ser gerados como saída em análises de natureza preditiva. São eles a matriz de confusão e o valor pra estatística de *kappa*.

A matriz de confusão, segundo estes autores, é utilizada para mensurar o desempenho de um algoritmo de classificação. Na forma de uma matriz bidimensional, ela demonstra a quantidade de instâncias de cada classe que foram classificadas em cada classe. A estatística de *kappa*, por sua vez, é uma métrica que, basicamente, informa a

probabilidade de as entidades em estudo estarem sendo classificadas ao acaso. Dada uma escala, um valor baixo significa alta probabilidade de isto estar acontecendo, e um valor alto, o inverso.

De acordo com Han, Kamber e Pei (2009), tarefas de classificação (e/ou também de regressão) podem necessitar de uma análise de relevância prévia, sobre os atributos disponibilizados para identificar aqueles que são significantemente relevantes ao processo de classificação (e/ou regressão). Estes devem ser selecionados para a realização da tarefa, enquanto outros atributos, considerados irrelevantes, podem ser simplesmente descartados e excluídos de consideração.

3.1.3 Clustering

Fayyad, Shapiro e Smyth (1996) definem a funcionalidade de *clustering* como uma tarefa descritiva comum cujo objetivo é a identificação de um conjunto finito de categorias (ou *clusters*) para a descrição dos dados. Estas categorias podem ser mutuamente exclusivas ou constituir uma representação mais complexa, como definições hierárquicas ou, ainda, com a presença de intersecções entre os grupos descritos após a análise.

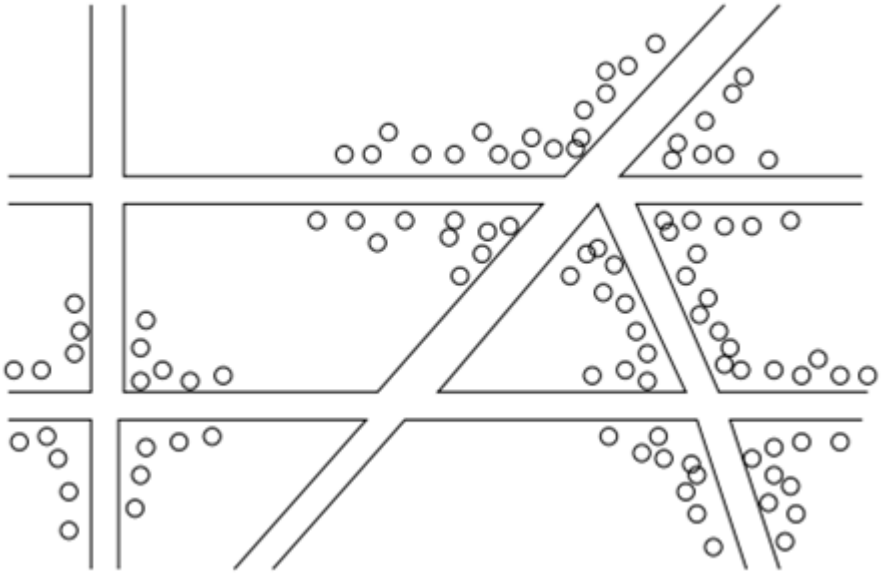
Segundo Harrington (2012), *clustering* é um tipo de funcionalidade de mineração de dados de natureza não supervisionada que, automaticamente, cria agrupamentos de elementos semelhantes, portanto, um tipo de classificação automática. Este autor entende que praticamente qualquer conjunto de dados pode ser analisado com esta funcionalidade da mineração de dados e que o nível de qualidade dos *clusters* identificados como resultado é proporcionalmente associada à similaridade entre os elementos contidos em cada *cluster*.

Os algoritmos utilizados em *clustering* utilizam os dados disponibilizados para agrupar os elementos (registros, no caso de se tratar de uma projeção de banco de dados) similares e retornar dados sobre estas semelhanças. Ocorre, portanto, a disposição de elementos similares em determinado agrupamento e elementos dissimilares em outro determinado agrupamento, além de que o entendimento de similaridade depende de métricas que podem ser definidas para este propósito.

Rajaraman, Leskovec e Ullman (2013) dispõem um exemplo bastante simples e conhecido da aplicação da funcionalidade de *clustering* que foi para a solução de um problema ocorrido no ano de 1854 onde, no tratamento de uma epidemia de cólera onde a

localização geográfica dos casos foi disposta sobre um mapa da cidade de Londres, conforme a figura abaixo.

Figura 3 - Plotagem de casos de cólera por *clustering*.



Fonte: Adaptado de Rajaraman, Leskovec e Ullman (2013).

O processo, realizado manualmente, devido às restrições tecnológicas da época da ocorrência dos fatos, mostrou que os casos da doença estavam agrupados em proximidade à intersecções de rodovias. Estas intersecções coincidiam com as localizações de poços artesianos que haviam sido contaminados e, portanto, pessoas que viviam perto destes poços sofreram a doença, enquanto pessoas que vivam perto de poços não contaminados ou não próximos às mesmas localizações não sofreram a doença. É estimado que, sem a habilidade de agrupar estes dados (utilização da funcionalidade de *clustering*), mesmo que da forma manual e simples, como ocorreu, a causa da epidemia de cólera não teria sido descoberta.

Rajaraman, Leskovec e Ullman (2013) entendem que as funcionalidades de *clustering* e de classificação são diferenciadas pelo fato de que na classificação o usuário tem conhecimento do objetivo da atividade, ou seja, este sabe o que está procurando em cada classe. Estes autores entendem que, teoricamente, a atividade de *clustering* produz o mesmo resultado que a classificação, mas sem a informação prévia de um conjunto de classes, rótulos de classes ou, no caso, agrupamentos ou *clusters*.

3.1.4 Abordagens Algorítmicas

Witten, Frank e Hall (2011) enfatizam a utilização do algoritmo *Apriori* para o tratamento de grandes conjuntos de dados em atividades de associação, o qual permite a geração de regras de associação que suprem limites mínimos de indicadores de suporte e confiança, além da utilização de uma metodologia de geração e teste de regras para padrões de conjuntos de elementos frequentes.

De acordo com Agrawal e Srikant (1994), o algoritmo *Apriori* reduz, iterativamente, o indicador de suporte mínimo até que encontre a quantidade requerida de regras com o indicador mínimo de confiança enquanto busca por regras de associação. Han, Kamber e Pei (2009) explicam que este algoritmo utiliza conhecimento prévio sobre as propriedades dos conjuntos de elementos frequentes nestas buscas.

Segundo Camargo (2002), o *Apriori* é o algoritmo mais utilizado para a descoberta de regras de associação e executa diversas leituras sobre o conjunto de dados em análise. Conforme Witten, Frank e Hall (2011), diversos melhoramentos foram sugeridos ao algoritmo, para redução da quantidade destas leituras, mas isto ainda pode causar grande custo de poder computacional pela natureza combinatória na geração de regras de associação do mesmo. Conforme Liu, Hsu e Ma (1998), este algoritmo pode, também, ser adaptado para identificar associações originárias de atividades de classificação.

No que se trata de atividades de classificação em mineração de dados, Witten, Frank e Hall (2011), inicialmente, versam sobre o algoritmo 1R para modelagem estatística, sendo que este utiliza um único atributo como base à cada tomada de decisão, escolhendo o que se aplica melhor à cada nova situação. No caso de construção de árvores de decisão em processos indutivos de classificação, é sugerida a utilização do algoritmo C4.5, o qual é considerado uma evolução do algoritmo ID3, tendo incorporado diversos melhoramentos à este, incluindo métodos para tratamento de atributos numéricos, valores ausentes e dados inconsistentes e a capacidade de geração de regras a partir de árvores. Estes autores entendem que, atualmente, o algoritmo C4.5 é o mais utilizado em atividades de classificação, em geral, apesar da existência de alternativas também eficazes no mesmo propósito, como o algoritmo CART (*Classification and Regression Trees*).

Conforme Han, Kamber e Pei (2009), o algoritmo *k-means* utiliza a metodologia de particionamento do conjunto de dados em um determinado número de agrupamentos, sendo esta a abordagem mais simples e fundamental em *clustering*. O número de *clusters* ou agrupamentos é assumido como conhecimento prévio e é utilizado como parâmetro no ponto inicial no particionamento de conjuntos de dados em evidência.

Witten, Frank e Hall (2011) concordam com esta abordagem e consideram o algoritmo *k-means* um clássico em tarefas de *clustering* em mineração de dados, onde é importante notar que a podem ser executados testes sobre a quantidade de agrupamentos, até que a atividade atinja seu objetivo da forma mais vantajosa e relevante à um domínio de negócio. Variações foram desenvolvidas com o propósito de obter melhoras de desempenho, conforme o algoritmo *k-means++*, por exemplo, onde o centro dos agrupamentos é determinado de maneira mais cuidadosa, impactando positivamente na velocidade de processamento e precisão dos resultados.

Para efeitos de estudo neste trabalho, faz-se notar que os algoritmos *Apriori*, 1R, C4.5, CART e *k-means* são utilizados para fins de experimentações e análise no conjunto de dados apresentado para pesquisa.

O capítulo 3 se propôs a mostrar a etapa de mineração de dados no processo de descoberta de conhecimento como uma variedade de abordagens e possibilidades, explicitando conceitos inerentes e apresentando as funcionalidades de associação, classificação e *clustering*. Também foram apresentados breves princípios sobre abordagens algorítmicas utilizadas em cada uma destas funcionalidades, conforme proposto dentro deste estudo.

O capítulo 4 se destina à apresentação da ferramenta WEKA, utilizada para prática de pré-processamento de dados e aplicação de funcionalidades de mineração de dados, conforme proposto neste estudo, através dos algoritmos disponibilizados pela mesma.

4 A FERRAMENTA WEKA

O projeto *Waikato Environment for Knowledge Analysis* (WEKA), desenvolvido pela Universidade de Waikato na Nova Zelândia[1], é uma coleção de algoritmos de aprendizado de máquina para realização tarefas de mineração de dados. Estes algoritmos podem ser executados diretamente sobre conjuntos de dados ou utilizados dentro de projetos em linguagem de programação Java, a mesma a qual este sistema é escrito e provê uma *Application Programming Interface* (API). Também é exposto que o WEKA contém ferramentas para pré-processamento de dados, tarefas de classificação, regressão, análise de agrupamentos (*clustering*) e associação, e, ainda, a visualização de resultados, além de poder ser utilizado no desenvolvimento de novas abordagens ao aprendizado de máquina, sendo que todo pacote de *software* é *open source* (de código-fonte aberto) e disponibilizado (licenciado) sob a *GNU General Public License*.

Segundo o manual da ferramenta WEKA (BOUCKAERT et al., 2013), em sua versão 3.6.9, a interface inicial deste sistema, o "WEKA *Graphic User Interface Chooser*", provê um ponto inicial para a inicialização das principais aplicações e ferramentas de suporte com interface gráfica do WEKA. Os quatro botões apresentados nesta interface, "*Explorer*", "*Experimenter*", "*KnowledgeFlow*" e "*Simple CLI*" são utilizados para acessar cada um dos ambientes disponibilizados por estas quatro aplicações, brevemente descritas abaixo, de acordo com suas aplicações propostas.

A figura 4 demonstra a interface inicial do sistema WEKA.

[1] *Software* e documentação inerente disponível para download em http://www.cs.waikato.ac.nz/ml/weka/.

Figura 4 - Interface inicial da ferramenta WEKA

Fonte: Adaptado de Frank et al. (2009).

4.1 Ambiente Explorer

O ambiente *Explorer* é utilizado para explorar os dados com a ferramenta WEKA e, segundo Witten, Frank e Hall (2011), a maneira mais simples de utilizar a ferramenta por interface gráfica. Este ambiente dispõe acesso à suas facilidades através da seleção de menus e preenchimento de parâmetros em formulários. Além disto, a interface gráfica do ambiente *Explorer* provê meios para a leitura e importação de conjuntos de dados, tanto no formato ARFF (sugerido para utilização na ferramenta WEKA), como em arquivos de planilhas de dados em formatos compatíveis.

Funcionalidades básicas de pré-processamento de conjunto de dados são disponibilizadas juntamente com a capacidade de efetuar tarefas de mineração de dados baseadas em funcionalidades e exploração dos diversos algoritmos disponibilizados pela WEKA. Também são disponibilizadas interfaces para a apresentação dos resultados oriundos da atividade de mineração de dados.

Apesar destas facilidades, estes autores ainda entendem que usuários da ferramenta devem ter conhecimento das atividades desempenhadas para compreender, plenamente, a significação dos resultados obtidos.

4.2 Ambiente Experimenter

O ambiente *Experimenter* é, conforme Bouckaert et al. (2013), utilizado para efetuar experimentos e realizar a condução de testes de natureza estatística entre abordagens de aprendizado de máquina.

Segundo Witten, Frank e Hall (2011), este ambiente foi projetado para resolver um problema básico das atividades de classificação e regressão em mineração de dados, que é o da determinação de métodos e valores de parâmetros que se aplicam melhor em cada estudo de caso. Estes autores entendem que não há maneira dedutiva de fazer esta determinação e, com esta motivação, foi desenvolvido o ambiente *Experimenter*, onde são disponibilizadas funcionalidades para automatizar o processo de comparação entre variedades de técnicas de aprendizado de máquina. Desta maneira, atividades de classificação e/ou regressão podem ocorrer com filtros e configurações de parâmetros diferentes com o propósito de coletar informações estatísticas sobre desempenho em tais atividades. Além disto, ainda vale citar que existe a possibilidade de utilizar este ambiente sobre uma plataforma computacional distribuída para balanceamento de carga em casos de experimentos estatísticos de larga escala.

4.3 Ambiente KnowledgeFlow

O ambiente *KnowledgeFlow* suporta, de acordo com Bouckaert et al. (2013), essencialmente, as mesmas funcionalidades do ambiente *Explorer*, porém, com uma facilidade de clicar-e-arrastar em sua interface gráfica. Funcionalmente, uma vantagem seria o suporte à aprendizado incremental.

De acordo com Witten, Frank e Hall (2011), a interface deste ambiente provê meios para o projeto de configurações para o processo de dados em fluxo contínuo (*streaming*) e esta é a natureza da diferença entre este ambiente e o ambiente *Explorer*. No caso deste, todo conjunto de dados é armazenado na memória principal do sistema, enquanto, no ambiente *KnowledgeFlow*, o aprendizado incremental, conforme mencionado anteriormente, pode ser realizado de forma efetiva pelo carregamento incremental de conjuntos de dados.

Em sua interface gráfica, este ambiente expõe a funcionalidade de arrastar componentes que podem representar algoritmos de aprendizado ou fontes de dados, de maneira que componham a configuração desejada pelo usuário. Desta maneira, permite a especificação de fluxo contínuo de dados (*streaming*) pela conexão entre componentes de fontes de dados, ferramentas de pré-processamento, algoritmos de aprendizado, métodos

de avaliação de conhecimento e componentes responsáveis pela visualização dos resultados, permitindo que dados sejam, de fato, carregados e processados incrementalmente.

4.4 Ambiente Simple CLI

De acordo com Bouckaert et al. (2013), o ambiente *Simple CLI* dispõe uma interface em linha de comando para permitir a execução de comandos da ferramenta WEKA por sistemas operacionais que não provêem ambientes de linha de comando próprios.

Witten, Frank e Hall (2011) explicam que este ambiente é utilizado para a entrada de comandos em sua forma bruta, como chamada a API do WEKA, o que dá ao usuário acessibilidade total aos recursos deste sistema.

Para efeitos neste estudo, o ambiente *Explorer* é utilizado para a realização das tarefas de mineração de dados com as funcionalidades de associação, classificação e análise de agrupamentos (c*lustering*), conforme exposto anteriormente, portanto, este ambiente é estudado e exposto mais criteriosamente no desenvolvimento do presente trabalho.

4.5 O Formato de Arquivo ARFF

De acordo com a Universidade de Waikato (2013), um arquivo *Attribute-Relation File Format* (ARFF), ou formato de arquivo de atributos e relação, consiste num arquivo de texto pleno que descreve uma lista de instâncias que compartilham um conjunto de atributos.

Arquivos do tipo ARFF devem conter duas seções distintas, sendo elas a seção do cabeçalho, com informações inerentes, e a seção de dados, onde estão estes, propriamente ditos. Podem haver linhas com a função de comentários na constituição deste tipo de arquivo.

A seção de cabeçalho de um arquivo ARFF contém a declaração do nome da relação e a declaração dos atributos do conjunto de dados com seus respectivos tipos. Os tipos de dados suportados pelo WEKA são numéricos (números inteiros e reais), categóricos (nominais ou strings), e temporais (datas e horários).

Um exemplo de cabeçalho de um arquivo ARFF é demonstrado no quadro 2, abaixo.

Quadro 2 - Exemplo de cabeçalho de arquivo ARFF.

```
@relation veiculos

@attribute tipo { "CARRO", "CAMINHAO", "MOTO" }
@attribute cor { "PRETO", "BRANCO", "AZUL" }
@attribute marca { "VW", "BMW", "FIAT", "HONDA" }
@attribute ano integer
@attribute peso integer
@attribute rodas integer
```

Fonte: Dados primários.

A seção de dados de um arquivo ARFF contém, além da declaração do início desta seção, a informação dos dados a serem analisados que, por sua vez, representam as instâncias a ser analisadas. Cada instância deve ter seus dados informados em uma única linha de texto no arquivo, sendo que o fim da linha denota o fim dos dados referentes a cada instância.

É importante notar que valores ausentes ou nulos podem ser representados pelo caractere correspondente ao sinal de interrogação ("?"). Valores nominais e *strings* têm sensitividade entre caracteres alfanuméricos maiúsculos ou minúsculos, além de que é necessária sua qualificação com aspas duplas, no caso da presença de espaços em sua composição. Valores de datas devem seguir o padrão de *strings* conforme especificado na declaração do respectivo atributo no cabeçalho. Também pode ser utilizada a interpretação de dados esparsos, onde dados nulos não são informados explicitamente, porém, este não é o caso no presente estudo.

Um exemplo de seção de dados de um arquivo ARFF é demonstrada abaixo, no quadro 3.

Quadro 3 - Exemplo de seção de dados de arquivo ARFF.

```
@data
"CARRO", "AZUL", "VW", 1978, 500, 4
"CARRO", "PRETO", "FIAT", ?, 550, 4
"MOTO", "PRETO", "HONDA", 2010, 120, 2
```

```
"CAMINHAO", "PRETO", "BMW", 2010, 3000, 8
"CARRO", "PRETO", "HONDA", 2010, 600, 4
"CARRO", "PRETO", "BMW", 2010, 800, 4
"CAMINHAO", "BRANCO", "BMW", ?, 2800, 6
```

Fonte: Dados primários.

Conforme pode ser verificado nos exemplos apresentados acima (quadros 2 e 3), existem notações específicas para a determinação do nome da relação, os atributos e sua tipificação no conjunto de dados e o indicador de início da seção de disposição do conjunto de dados, além da notação representativa de comentários em linha do arquivo, sendo elas, respectivamente, *@relation*, *@attribute*, *@data* e o símbolo de percentagem ("%") e, ainda, a quebra de linha, que também apresenta semântica inerente à interpretação dos dados contidos nos arquivos ARFF.

Segundo Witten, Frank e Hall (2011), a maioria dos *Integrated Development Environments* (IDE), ou ambientes integrados de desenvolvimento, permitem algum tipo de exportação de dados em formato de planilhas ou *Comma-Separated Values* (CSV), ou valores separados por vírgulas, para a representação de projeções do banco de dados e a ferramenta WEKA suporta a importação de conjunto de dados neste formato.

Apesar disto, o formato ARFF foi mantido para utilização no presente trabalho por questões do estudo da ferramenta WEKA e suas funcionalidades inerentes, assim como a padronização das práticas, conforme especificado.

4.6 Implementações Algorítmicas

Conforme exposto anteriormente, foram elencados algoritmos para o desenvolvimento do presente trabalho, em acordo com as atividades de mineração de dados propostas e, de acordo com Frank et al. (2009), o sistema WEKA provê implementações para estes algoritmos, conforme a tabela 1, abaixo.

Tabela 1 - Implementações de algoritmos na ferramenta WEKA

Categoria	Algoritmo	Implementação
Associação	Apriori	Apriori
Classificação	1R	OneR
Classificação	C4.5	J48
Classificação	CART	SimpleCart
Clustering	k-means	SimpleKMeans

Fonte: Frank et al. (2009).

Este capítulo se propôs a versar sobre os quatro ambientes disponibilizados na ferramenta WEKA, citando suas peculiaridades e aplicações, além de apresentar detalhes sobre o formato de arquivo ARFF, proposto para utilização na entrada de conjunto de dados no ambiente *Explorer*. O capítulo 5 mostra a utilização da WEKA na prática, em cada um dos cinco experimentos propostos no presente estudo.

5 OVINOCULTURA E MINERAÇÃO DE DADOS

Este capítulo versa sobre funcionalidades de mineração de dados aplicadas à um conjunto de dados oriundo do domínio da ovinocultura e os resultados obtidos nos experimentos conduzidos dentro das propostas deste estudo.

A seção 5.1, inicialmente, provê embasamento sobre o domínio de negócio da ovinocultura, conforme exposto no Regulamento do Serviço de Registro Genealógico de Ovinos no Brasil (ARCO, 2012). A seção 5.2 expõe detalhes sobre o conjunto de dados utilizado e cinco experimentos executados estão descritos na seção 5.3, detalhadamente nas respectivas subseções 5.3.1, 5.3.2, 5.3.3, 5.3.4 e 5.3.5.

5.1 Domínio de Negócio da Ovinocultura

O serviço de registros inerentes à criação de ovinos no Brasil é regulamentado e executado pela Associação Brasileira de Criadores de Ovinos - Assistência aos Rebanhos de Criadores de Ovinos (ARCO), em conformidade com delegação do Ministério de Agricultura, Pecuária e Abastecimento (MAPA) deste País. Este regulamento expressa, entre outros aspectos administrativos cujo detalhamento está fora do escopo deste trabalho, sobre os documentos de formulários de dados inerentes ao processo de regularização de rebanhos ovinos e as informações contidas neles para fins de registro.

A seguir, é provida uma descrição breve do ciclo de vida do processo de registro de ovinos nesta associação, bem como sobre os documentos, formulários, informações ou dados inerentes a cada fase deste processo.

O registro de ovinos, assumindo plena regularidade dos respectivos criadores perante o MAPA e a ARCO, começa pela notificação das atividades de cobertura em seus rebanhos. Este documento reflete informações de acasalamentos e, portanto, dados dos machos que foram colocados em confinamento para acasalamento com determinado grupo de fêmeas e o período em que este confinamento ocorreu. Este período, de acordo com o regulamento citado, não deve exceder o intervalo de 90 dias. No caso de não ocorrer monta natural, único caso passível de notificação de cobertura, deve haver a informação de relatórios de inseminação artificial e/ou fertilização *in-vitro*, além de casos de transferência de embriões posterior à monta natural ou não.

Caso estes dados não sejam informados em conformidade com o regulamento, pode gerar irregularidades referentes à desinformação do tipo de monta ou método de concepção realizado por seus progenitores, quando da avaliação de regularidade dos animais informados em notificações de nascimento.

A notificação de nascimento, como o próprio nome indica, provê informações sobre o nascimento de animais de determinado rebanho. Os dados contidos no formulário entregue pelos criadores deve conter indicação do rebanho, números de tatuagens, sexo dos animais, nomes dos animais, datas de nascimentos, indicação dos progenitores e tipo de monta ou método de concepção ao qual a notificação é referente.

Concomitantemente à notificação de nascimento, é realizada e informada a inspeção ao pé da mãe, procedimento realizado por técnicos zootecnistas credenciados e autorizados pela ARCO, onde são informadas as tatuagens e o sexo dos animais nascidos em cada rebanho. Caso este documento esteja em discordância com a notificação de nascimento, pode gerar irregularidades em função de discrepâncias entre os dados informados pelos criadores e o que foi, efetivamente, verificado pelos técnicos.

A ficha de confirmação é outro documento cuja responsabilidade é atribuída aos técnicos zootecnistas credenciados pela ARCO. Essencialmente, trata-se de um relatório técnico onde são indicados animais que atingiram maturidade suficiente para ter sua genealogia e padrão racial reconhecidos, o que, legalmente, só deve ocorrer após oito meses de vida do animal e até seus três anos de vida.

Apesar de a ficha de confirmação ser um documento de análise técnica, administrativamente, ela somente surte efeito para cada animal, individualmente, à aqueles listados e que não possuem irregularidades de acordo com o regulamento, estando em estado de aptidão na data da confirmação.

Abaixo, é disposta uma tabela com os códigos de irregularidades possíveis, juntamente com suas descrições, sendo que a atribuição de qualquer destas irregularidades impede que qualquer animal seja considerado apto.

Tabela 2 - Irregularidades e Descrições.

Código	Descrição da Irregularidade
1	Mãe não confirmada
2	Pai não confirmado
3	Irmãos gêmeos com datas de nascimento diferentes
4	Mãe tem produto com outro pai, nesta data
5	Mãe do produto não é de propriedade do criador
6	Criador já possui produto com a mesma tatuagem
7	Intervalo entre partos inválido

8	Intervalo entre cobertura e nascimento inválido
9	Cobertura não comunicada
10	O pai não é de propriedade do criador
11	Sem ficha de Inspeção ao Pé da Mãe
12	Criador não possui estoque de embriões
13	Criador não possui estoque de sêmen
14	Relatório de Transferência de Embriões não comunicado ou incompleto
15	Relatório de Inseminação Artificial não comunicado ou incompleto
16	Teste de parentesco não realizado
17	Teste de parentesco não qualifica

Fonte: ARCO, 2012.

A relevância deste processo de confirmação está, segundo o regulamento da ARCO (ARCO, 2012), na certificação de que os animais são, efetivamente, regularizados perante as entidades inerentes, ou seja, passaram por todas etapas regulares do processo de criação de ovinos no Brasil. Além disto, a certificação racial torna os proprietários destes animais aptos à usufruir plenamente dos atributos de seus rebanhos, o que influencia diretamente em viabilidade comercial, por exemplo, devido ao fato de que transferências de propriedade podem ser efetuadas de maneira regular somente se os animais possuírem seus certificados de confirmação.

Conforme exposto anteriormente, funcionalidades de mineração de dados foram aplicadas neste estudo com a execução de algoritmos de associação, classificação e *clustering* sobre um conjunto de dados oriundos do domínio de negócio da criação de ovinos (seção 5.3).

5.2 O Conjunto de Dados

O conjunto de dados disponibilizado para este estudo está disposto conforme um esquema relacional, com a tecnologia *SQLite*, em duas entidades, ou tabelas, sendo elas OVINOS e IRREGULARIDADES.

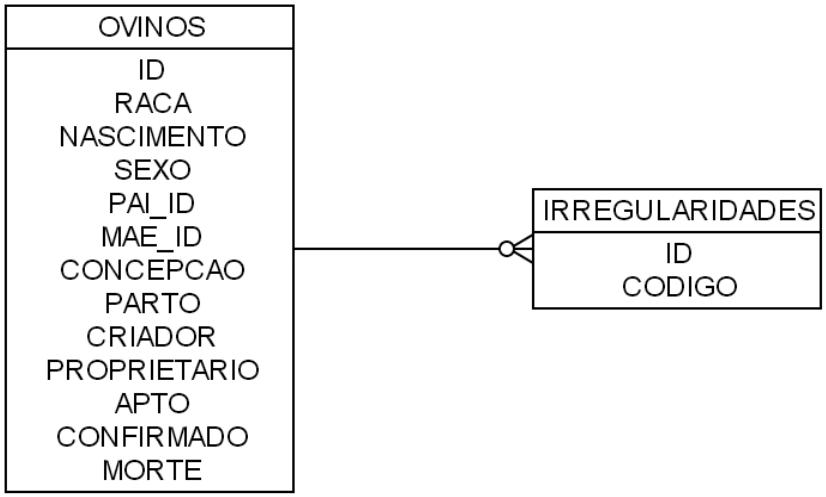
A primeira, a entidade OVINOS, é constituída por 13 atributos, considerando que cada tupla representa dados sobre cada animal, sendo eles, portanto, data de nascimento, sexo, raça, identificação do pai, identificação da mãe, seu tipo de concepção, tipo de parto, identificação do criador, identificação do proprietário, indicador de aptidão, indicador de

confirmação e data de morte, se morto, além de seu número de identificação própria, para fins de utilização como chave-primária no esquema relacional apresentado.

A segunda entidade, IRREGULARIDADES, contém o número de identificação de cada animal e o código das irregularidades atribuídas à cada animal, se não em estado de aptidão. O primeiro atributo citado na composição desta entidade serve como chave-estrangeira no relacionamento com a entidade OVINOS, onde existe um atributo correspondente como chave-primária, conforme exposto anteriormente. Isto é suficiente para a manutenção da integridade referencial no esquema relacional disposto neste estudo.

A figura 5, abaixo, representa um diagrama Entidade-Relacionamento (ER) demonstrando a relação entre as duas entidades e seus atributos.

Figura 5 - Diagrama ER representando o conjunto de dados utilizado neste estudo.



Fonte: Dados primários.

Para fins de geração do arquivo ARFF utilizado como entrada no processo de mineração de dados na ferramenta WEKA, foram calculados atributos adicionais, considerados relevantes para atividades de mineração de dados e, então adicionados ao arquivo final utilizado neste trabalho. Outros atributos contidos nas entidades do esquema relacional apresentado anteriormente, que foram considerados irrelevantes para as mesmas atividades, foram descartados.

Anteriormente à geração do arquivo ARFF, com base nos dados apresentados, foram calculados e acrescentados à entidade OVINOS a idade do animal em meses, a idade do animal em meses quando de sua morte, se morto, um indicador da ocorrência de

transferência de propriedade e a quantidade de irregularidades apresentadas pelos animais considerados inaptos. Foram, também, descartados dados referentes a data de nascimento, data de morte, tatuagem do animal, identificação do criador e identificação do proprietário. A entidade IRREGULARIDADES foi mantida conforme seu formato disposto anteriormente.

O produto final do processo de geração do arquivo ARFF consolidou dados de aproximadamente 50.000 instâncias da relação ovinos para análise.

Para fins de transformação de dados, conforme disposto acima, foi utilizada a linguagem de programação *PHP: Hypertext Preprocessor* (PHP) para automatização da recuperação dos dados e geração do arquivo ARFF utilizado no presente trabalho. O quadro 4 apresenta o código-fonte do programa escrito, em sua integridade, onde também estão explícitas sentenças escritas em *Structured Query Language* (SQL).

Quadro 4 - Código fonte de programa em linguagem PHP para geração de arquivo ARFF.

```
<?php

      $db = new pdo('sqlite:ovinos.sqlite');
      $db->setAttribute(PDO::ATTR_ERRMODE, PDO::ERRMODE_EXCEPTION);
      $sql = "
            SELECT
                  RACA,
                  SEXO,
                  CASE WHEN MORTE > DATE('1899-12-30') THEN 'N' ELSE 'S'
END VIVO,
                  CASE WHEN NASCIMENTO > DATE('1899-12-30') THEN
ROUND((JULIANDAY('NOW')-JULIANDAY(NASCIMENTO))/30) ELSE '?' END
MESES_VIVO,
                  CASE WHEN MORTE > DATE('1899-12-30') THEN
ROUND((JULIANDAY(MORTE)-JULIANDAY(NASCIMENTO))/30) ELSE '?' END
MORTO_MESES_VIVO,
                  CASE TIPO_MONTA WHEN 'Inseminação Artificial' THEN 'IA'
WHEN 'Monta Natural' THEN 'MN' ELSE TIPO_MONTA END CONCEPCAO,
                  CASE PARTO WHEN 'Simples' THEN '1' WHEN 'Duplo' THEN '2'
WHEN 'Triplo' THEN '3' WHEN 'Quadruplo' THEN '4' END PARTO,
                  CASE APTO WHEN 1 THEN 'S' ELSE 'N' END APTO,
                  CASE CONFIRMADO WHEN 1 THEN 'S' ELSE 'N' END CONFIRMADO,
                  CASE WHEN PROPRIETARIO <> CRIADOR THEN 'S' ELSE 'N' END
```

```
TRANSFERIDO,
                  (SELECT COUNT(*) FROM IRREGULARIDADES WHERE FBB =
OVINOS.FBB) IRREGULARIDADES
            FROM OVINOS
      ";
      $result = $db->query($sql);
      $file = fopen('ovinos_santa_ines_vivos_.arff', 'w');
      fwrite( $file, '@relation ovinos' . "\r\n");
      fwrite( $file, "\r\n");
      fwrite( $file, '@attribute raca { "BERGAMACIA BRASILEIRA", "CARIRI",
"CORRIEDALE", "CRIOULA", "DOHNE MERINO", "DORPER", "EAST FRIESIAN",
"HAMPSHIRE DOWN", "IDEAL", "ILE DE FRANCE", "KARAKUL", "LACAUNE", "MERINO
AUSTRALIANO", "MORADA NOVA", "POLL DORSET", "RABO LARGO", "ROMNEY MARSH",
"SANTA INES", "SOMALIS BRASILEIRA", "SUFFOLK", "TEXEL", "WHITE DORPER" } '
. "\r\n");
      fwrite( $file, '@attribute sexo { M, F } '                . "\r\n");
      fwrite( $file, '@attribute vivo { S, N } '                . "\r\n");
      fwrite( $file, '@attribute idade_vivo integer '           . "\r\n");
      fwrite( $file, '@attribute idate_morto integer '          . "\r\n");
      fwrite( $file, '@attribute concepcao { MN, IA, TE, FIV } '
      . "\r\n");
      fwrite( $file, '@attribute parto { 1, 2, 3, 4 } '         . "\r\n");
      fwrite( $file, '@attribute apto { S, N } '                . "\r\n");
      fwrite( $file, '@attribute confirmado { S, N } '          . "\r\n");
      fwrite( $file, '@attribute transferido { S, N } '         . "\r\n");
      fwrite( $file, '@attribute irregularidades integer '      . "\r\n");
      fwrite( $file, "\r\n");
      fwrite( $file, "@data" . "\r\n");
      while ( $tuple = $result->fetch() ) {
            fwrite(     $file,
                  (string) '"'.$tuple['RACA'].'"'         . ", " .
                  (string) $tuple['SEXO']                 . ", " .
                  (string) $tuple['VIVO']                 . ", " .
                  (string) $tuple['MESES_VIVO']           . ", " .
                  (string) $tuple['MORTO_MESES_VIVO']     . ", " .
                  (string) $tuple['CONCEPCAO']            . ", " .
                  (string) $tuple['PARTO']                . ", " .
                  (string) $tuple['APTO']                 . ", " .
                  (string) $tuple['CONFIRMADO']           . ", " .
                  (string) $tuple['TRANSFERIDO']          . ", " .
                  (int)    $tuple['IRREGULARIDADES']      . ", " .
```

```
                    "\r\n" );
        }
        fclose($file);

?>
```

Fonte: Dados primários.

Outras etapas de pré-processamento de dados foram aplicadas conforme a necessidade quando da execução de cada experimento disposto neste estudo na própria ferramenta WEKA.

5.3 Experimentos de Mineração de Dados Sobre o Conjunto de Dados

Esta seção trata, especificamente, de cada experimento realizado nos propósitos do presente trabalho.

5.3.1 Experimento 1

Sobre o conjunto de dados disponibilizado, foi executado o algoritmo *Apriori* com a finalidade de determinar associações frequentes entre atributos das instâncias de ovinos. Este experimento inicial não utiliza nenhuma parametrização fora do padrão estipulado pela ferramenta WEKA e não elimina quaisquer atributos do conjunto de dados a ser analisado. O limite de regras a serem apresentadas neste experimento foi determinado como 100.

Como resultados, foram geradas, na grande maioria, regras determinando relações entre quesitos que fazem parte das regras do domínio do negócio, como, por exemplo, a associação entre a ausência de irregularidades e o estado de aptidão.

Por outro lado, a associação entre animais do sexo feminino e a ausência de transferências em seu histórico, com indicador de confiança de 0,97 pode ser explicado pelo valor que fêmeas, mesmo que não consideradas provenientes de uma árvore genética de alta fertilidade, podem agregar à criação de ovinos quando da atividade de transferência de embriões.

Outra regra, à princípio parte do domínio de negócio e, portanto, algo já esperado, é a associação entre animais não confirmados e sem registro de transferências em seu histórico. O fato relevante na descoberta desta regra foi o indicador de confiança, onde foi apresentado 0,99, revelando que há casos onde animais tiveram transferências de

propriedades registradas mesmo sem terem registradas suas respectivas confirmações de padrão racial antes. Conforme exposto anteriormente, no capítulo 2, este tipo de inconsistência pode ser causado por falhas no próprio produto de *software* de gestão do negócio.

O quadro abaixo mostra um demonstrativo com 10 das regras de associação reveladas por este experimento, conforme geradas pela ferramenta WEKA e analisadas conforme descrito acima.

Quadro 5 - Demonstrativo de regras de associação geradas pelo algoritmo *Apriori*.

```
irregularidades=0 39596 ==> apto=S 39596    conf:(1)
confirmado=N 39368 ==> transferido=N 38922    conf:(0.99)
sexo=F concepcao=MN 24435 ==> vivo=S 24110    conf:(0.99)
sexo=F 28826 ==> vivo=S 28352    conf:(0.98)
concepcao=MN 41485 ==> vivo=S 40627    conf:(0.98)
concepcao=MN apto=S 33923 ==> vivo=S 33400    conf:(0.98)
parto=1 31326 ==> transferido=N 30724    conf:(0.98)
apto=S 39596 ==> vivo=S 38800    conf:(0.98)
sexo=F 28826 ==> transferido=N 28054    conf:(0.97)
sexo=F apto=S 23503 ==> transferido=N 22734    conf:(0.97)
```

Fonte: Dados primários.

5.3.2 Experimento 2

Sobre o conjunto de dados, em sua totalidade de instâncias de animais das raças Santa Inês, *Dorper* e *Texel*, foram aplicados algoritmos de classificação para determinar suas capacidades de classificar as instâncias de acordo com as raças de cada um dos animais informados, sendo este atributo utilizado como rótulo das classes utilizando os outros atributos disponíveis.

Estas raças foram elencadas para este experimento por conterem o maior plantel no País, no que consta no conjunto de dados em estudo. Nenhum atributo foi removido do conjunto de dados para fins desta atividade.

Em relação à análise das raças de maior plantel, conforme citado acima, com a utilização do algoritmo C4.5, foi construído um modelo de bastante complexidade, cuja árvore de decisão gerou 533 nós. O atributo referente ao tipo de concepção dos animais

foi determinado como raiz deste modelo, que atingiu 62% de correção na classificação. Porém, o indicador de estatística de *kappa* apresentado foi de 0,41, indicando uma possível relevância estatística ao resultado do processo. A matriz de confusão apresentada como resultado também indica que, predominantemente, animais das raças *Dorper* e *Texel* tiveram maior sucesso na distinção sobre animais da raça Santa Inês, sendo que esta teve maior taxa de erros do que acertos na classificação, conforme a tabela 3.

Tabela 3 - Matriz de confusão gerada com algoritmo C4.5.

Dorper	Santa Inês	Texel	Classificação
9069	2581	2302	Dorper
3528	4866	1885	Santa Inês
1416	697	5917	Texel

Fonte: Dados primários.

O algoritmo CART gerou um modelo de maior complexidade em relação ao algoritmo C4.5. Neste caso, a geração da árvore de decisão criou 1581 nós, mas também utilizando o atributo referente à concepção dos ovinos como raiz do modelo.

Os indicadores de correção de classificação e estatística de *kappa* descobertos foram bastante semelhantes aos gerados pelo algoritmo C4.5, variando brevemente apenas nas casas decimais, de maneira positiva. A matriz de confusão gerada também foi semelhante, mostrando uma pequena melhora na classificação dos ovinos da raça Santa Inês (Tabela 4).

Tabela 4 - Matriz de confusão gerada com o algoritmo CART.

Dorper	Santa Inês	Texel	Classificação
9011	2671	2270	Dorper
3388	4889	2002	Santa Inês
1390	673	5967	Texel

Fonte: Dados primários.

O algoritmo 1R, por sua vez, criou um modelo de baixa complexidade baseando-se apenas em um único atributo do conjunto de dados, ou seja, de utilidade estatística, apenas. O atributo utilizado na modelagem foi referente à idade dos animais em meses, conforme a tabela 5.

Tabela 5 - Modelo gerado com o algoritmo 1R.

Idade em Meses	Raça
< 10,5	Dorper
< 12,5	Texel
< 18,5	Dorper
< 22,5	Santa Inês
< 26,5	Dorper
< 29,5	Santa Inês
< 31,5	Texel
< 32,5	Dorper
< 34,5	Santa Inês
< 35,5	Dorper
< 36,5	Texel
< 37,5	Dorper
< 47,5	Santa Inês
< 49,5	Dorper
< 57,0	Santa Inês
< 61,5	Dorper
< 69,5	Santa Inês
< 72,5	Dorper
>= 72,5	Santa Inês
Desconhecido	Dorper

Fonte: Dados primários.

O indicador de correção na classificação atingiu 55,1%, enquanto o indicador de estatística de *kappa* apresentou valor de 0,3.

A matriz de confusão disposta na tabela 6 demonstra superior taxa de acertos nas raças *Dorper* e *Texel*, assim como nas análises anteriores, porém, a classificação dos animais da raça Santa Inês foi bastante prejudicada, tendo, proporcionalmente, 3 erros a cada acerto.

Tabela 6 - Matriz de confusão gerada com o algoritmo 1R.

Dorper	Santa Inês	Texel	Classificação
9641	1583	2728	Dorper
5795	2549	1935	Santa Inês
1975	440	5615	Texel

Fonte: Dados primários.

5.3.3 Experimento 3

Em respeito ao conjunto de dados, neste caso, o número de instâncias foi reduzido à apenas os animais certificados nas raças *Dorper* e *White Dorper*.

No passado recente, estas duas raças eram consideradas uma única e, portanto, registradas, analisadas e julgadas como uma só. O motivo é que, morfologicamente, a diferença entre os animais destas duas raças é que os animais da raça *Dorper* possuem a característica de pigmentação preta na cabeça, da linha do pescoço para cima, enquanto o restante de seu corpo é da cor branca. Os animais da raça *White Dorper*, por sua vez, não possuem essa característica de pigmentação preta em local nenhum de seu corpo, apresentando coloração branca por todo este, incluindo a cabeça.

Os testes realizados neste experimento utilizaram o algoritmo de classificação C4.5 para determinar a raça de cada instância dos animais, sendo este o atributo dependente, utilizando os outros atributos, preditores, para buscar padrões que distinguem os grupos de instâncias de ovinos das raças *Dorper* e *White Dorper*.

Utilizando a implementação J48 (do algoritmo C4.5) na ferramenta WEKA, esta foi capaz de gerar um modelo e uma árvore de decisão, porém, de maneira pouco aproveitável.

Ao utilizar 16087 instâncias de ovinos divididos entre as raças *Dorper* e *White Dorper*, no experimento, o resultado foi de 87% de acerto na determinação do padrão racial, porém, o indicador de estatística de *kappa* revelou valor 0,1 e uma matriz de confusão foi gerada conforme a tabela 7.

Tabela 7 - Matriz de confusão gerada com o algoritmo C4.5.

Dorper	White Dorper	Classificação
13836	116	Dorper
1906	229	White Dorper

Fonte: Dados primários.

O problema, neste caso, foi que as instâncias dos animais não estavam proporcionalmente divididos entre as duas raças, tendo uma razão de 1 animal *White Dorper* para cada 6,5 animais *Dorper*. Portanto, entende-se que o processamento resultante da execução do algoritmo utilizou estatística simples para construção de uma árvore de decisão, onde foi determinado que é mais provável que um animal seja da raça *Dorper*, simplesmente. Consequentemente, todas instâncias foram avaliadas como tal, chegando ao resultado de 87% de acerto, mesmo sem ter envolvido os outros atributos do conjunto de dados na avaliação, conforme fica explícito na figura 6.

Figura 6 - Árvore de decisão gerada no experimento 3 pelo algoritmo C4.5.

Fonte: Dados primários.

Para a resolução deste conflito, Witten, Frank e Hall (2011) sugerem uma técnica de amostragem chamada de amostragem sem substituição, onde, basicamente, são gerados números aleatórios entre 1 e o número total de instâncias da população em estudo (neste caso, 16087) com o objetivo de selecionar instâncias para determinada amostra. Esta técnica ainda provê tratamento para que números já gerados no tratamento não sejam repetidos, não permitindo, portanto, que instâncias sejam repetidas no conjunto de amostragem.

Desta forma, o conjunto de dados foi reduzido para 4196 instâncias, sendo mantidas 2061 instâncias de animais correspondentes à raça *Dorper* e 2135 instâncias de animais correspondentes à raça *White Dorper*.

Ao repetir o experimento, o algoritmo apresentou apenas 62% de acerto na classificação, apesar de que foi aumentado o indicador de estatística de *kappa* para 0,25 e uma matriz de confusão conforme a tabela 8.

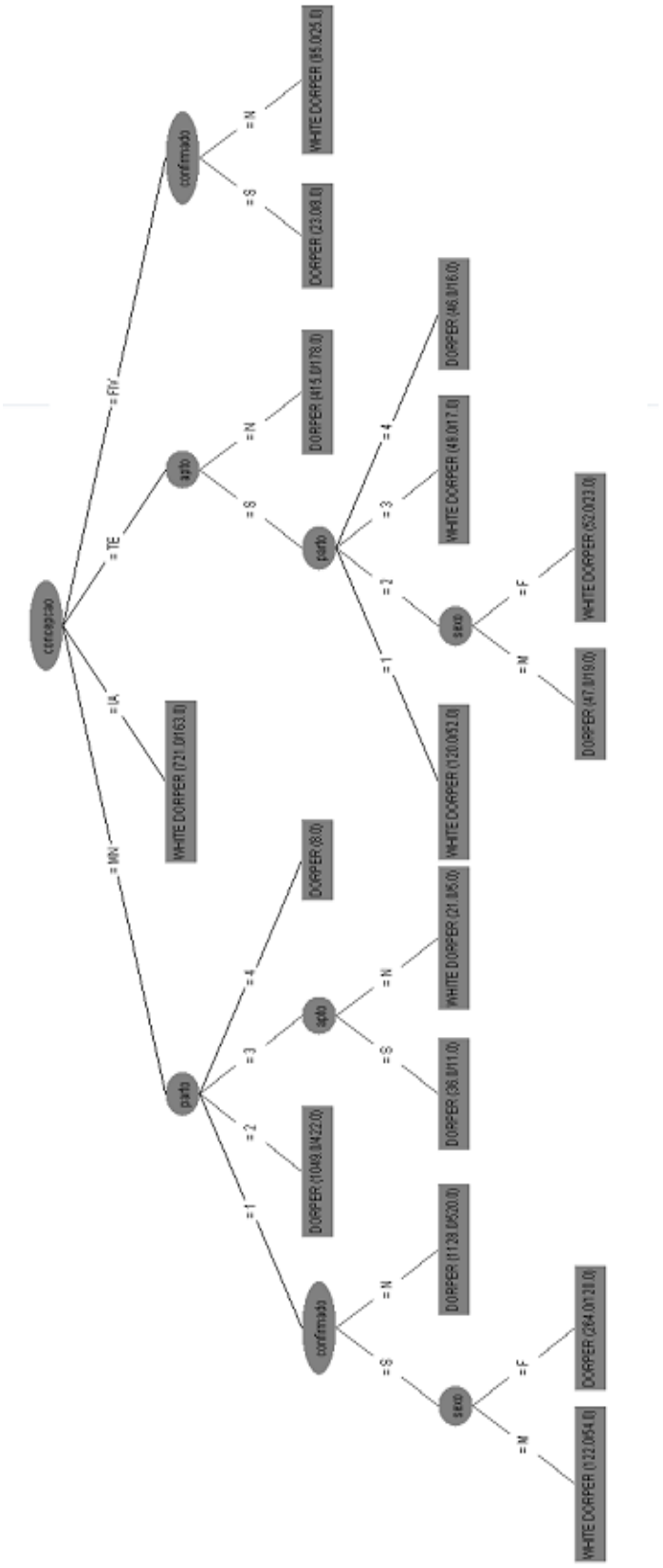
Tabela 8 - Matriz de confusão gerada no experimento 3 pelo algoritmo C4.5.

Dorper	White Dorper	Classificação
1365	696	Dorper
882	1253	White Dorper

Fonte: Dados primários.

A árvore de decisão gerada também foi influenciada pelo balanceamento proporcional realizado e foi disposta conforme a figura 7.

Figura 7 - Árvore de decisão gerada no experimento 3 com o algoritmo C4.5.



Fonte: Dados primários.

De acordo com a figura acima, dadas restrições de correção do modelo, conforme apresentado, pode-se interpretar uma dependência inicial sobre o tipo de concepção do animal, onde a geração embriões por inseminação artificial tende a ser mais comum à criação da raça *White Dorper*. Além disto, nota-se a ocorrência de partos quádruplos indicarem sua ocorrência na raça *Dorper*, apenas, quando de determinação baseada neste quesito.

5.3.4 Experimento 4

Neste experimento, foi conduzida uma atividade de associação baseada apenas nos dados de irregularidades de animais constantes no conjunto de dados em estudo.

Para fins de pré-processamento dos dados, quando da confecção do arquivo ARFF utilizado como entrada na ferramenta WEKA, foi criada e utilizada uma projeção de banco de dados específica, onde foram recuperados dados referentes à raça de cada um dos animais irregulares e apresentados seus indicativos de irregularidades em cada uma das possibilidades, sendo 17 no total, conforme a tabela 2, apresentado anteriormente no capítulo 5.

Novamente, foi aplicado o algoritmo *Apriori*, porém, sem critério de confiança para eliminação de regras irrelevantes. A única restrição ao resultado desta atividade foi a do número de regras a serem apresentadas, no caso, 10.

O algoritmo executou 18 iterações pelo conjunto de 7955 instâncias de animais irregulares e retornou duas regras de conjuntos de elementos frequentes, sendo

- Animais da raça *Dorper* associados à ausência de testes de parentesco (irregularidade de código 16) em 1119 casos; e
- Animais da raça Santa Inês associados à ausência de inspeção ao pé da mãe (irregularidade de código 11) em 1526 casos de irregularidades.

A raça Dorper foi associada à irregularidade 16 com 71% de confiança, enquanto a raça Santa Inês foi associada à irregularidade 11 com 60% de confiança.

Além disto, a raça *Texel* foi revelada como uma terceira problemática, no que se trata de ocorrências de irregularidades, porém, em menor proporção, não sendo associada à um conjunto de irregularidades em específico.

5.3.5 Experimento 5

O algoritmo *k-means* foi utilizado para fins de análise de agrupamentos (*clustering*) sobre o conjunto de dados em sua totalidade. Nesta atividade, foi eliminado, do conjunto de dados, o atributo determinando das raças de animais, utilizado como rótulo de classes em atividades de classificação, por irrelevância ao experimento.

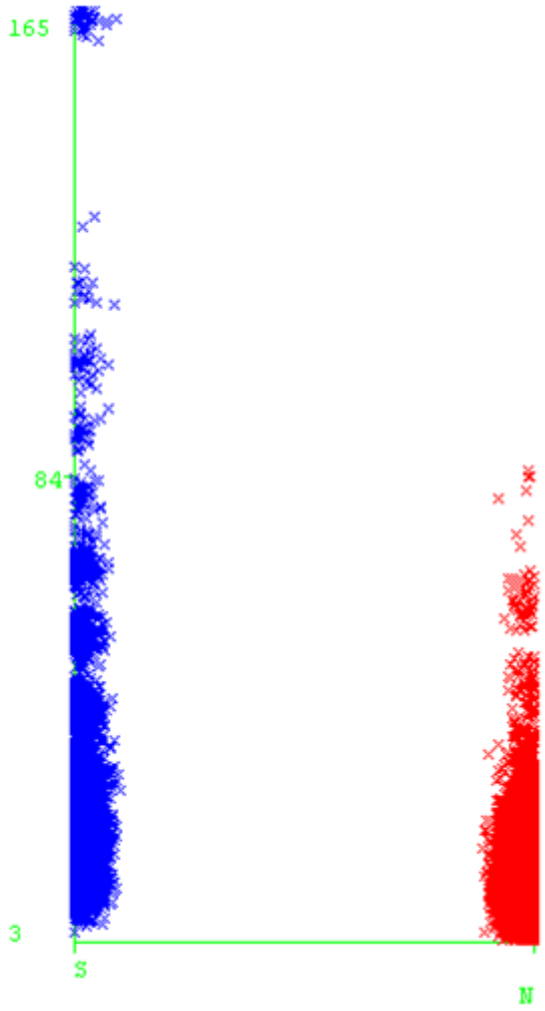
Neste caso, especificamente, a intenção da análise foi de agrupar instâncias de acordo com semelhanças nos valores de seus atributos, sem objetivo específico. Este experimento foi conduzido de forma exploratória e iterativa, elevando a quantidade de agrupamentos possíveis a cada iteração.

Através da interface da ferramenta WEKA, onde é permitido o cruzamento de dados em um ambiente gráfico em plano bi-dimensional, pôde ser observado um comportamento inerente às regras de negócio do domínio.

A figura 8 apresenta a plotagem de dois eixos, x (na horizontal) e y (na vertical), onde o primeiro representa a indicação de confirmação de animais, enquanto o segundo representa o atributo da idade dos animais em meses.

A plotagem das instâncias de animais no gráfico, conforme estes dois atributos, revela alta concentração de animais não confirmados a partir de seus três meses de idade. É aparente um hiato na atividade de confirmação desde três meses de idade dos animais até 8 meses de idade dos animais, quando estes atingem maturidade suficiente para terem seus certificados de padrão racial.

Figura 8 - Gráfico gerado no experimento 5 pelo algoritmo *k-means*.



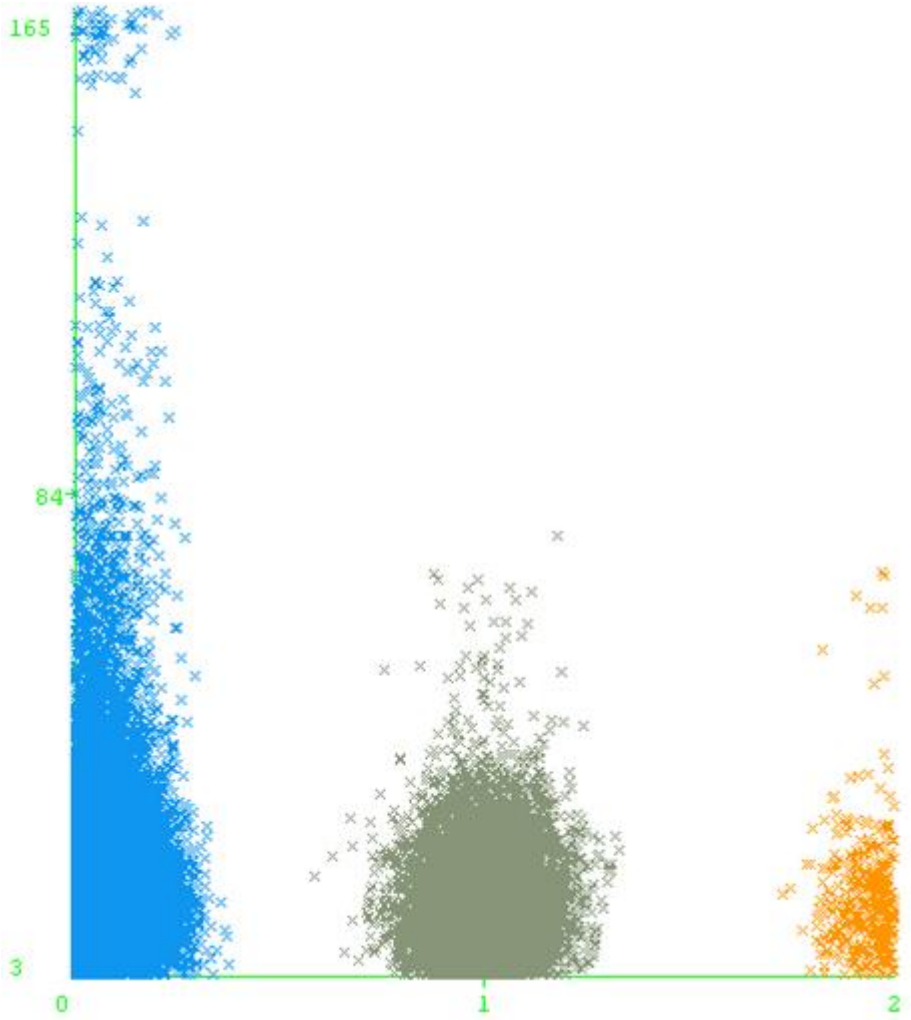
Fonte: Dados primários.

A figura 9 também apresenta dois eixos, x (na horizontal) e y (na vertical), sendo o eixo x representante da quantidade de irregularidades atribuídas aos animais, enquanto o eixo y representa a idade dos animais em meses.

Este gráfico cria suporte aos argumentos apresentados acima pela demonstração da atividade de correção de irregularidades a partir dos oito meses de idade dos animais, considerando que eles necessitam estar em estado de aptos, ou seja, ausência de irregularidades em seus registros, para poderem ser confirmados.

Em visualização à figura 9, nota-se a diminuição na quantidade de irregularidades a partir dos oito meses de idade dos animais, o que é requisito para a atividade de confirmação.

Figura 9 - Gráfico gerado no experimento 5 pelo algoritmo *k-means*.



Fonte: Dados primários.

O capítulo 5 se propôs à introduzir o domínio de negócio da ovinocultura no Brasil, de modo à apresentar seu potencial à atividade de mineração de dados. Neste contexto, a ferramenta WEKA foi utilizada sobre um conjunto de dados oriundo do domínio da criação de ovinos em cinco experimentos distintos.

Os experimentos 1 e 4 resultaram, respectivamente, na obtenção de regras de associação que revelaram possíveis atividades fora de padrão do sistema de gestão de negócio e alta ocorrência de determinadas irregularidades em determinadas raças ovinas.

Os experimentos 2 e 3 foram capazes, através de atividades de classificação, de demonstrar como diferentes abordagens algorítmicas podem influenciar na geração de árvores de decisão, além de evidenciar como desproporção quantitativa entre entidades de

diferentes classes pode levar à resultados irrelevantes neste tipo de atividade. Também foi indicada uma abordagem à solução deste problema, que foi utilizada com sucesso.

O experimento 5, utilizando-se da funcionalidade de *clustering* sobre o conjunto de dados, foi capaz de apresentar, visualmente, padrões de atividades em criação de ovinos em artefatos gráficos.

O capítulo 6, finalmente, apresenta as considerações finais e conclusões do presente estudo.

6 CONCLUSÃO

Diante dos resultados apresentados nos experimentos executados no presente trabalho, puderam-se observar diferentes abordagens à descoberta de conhecimento com mineração de dados, dada a preparação correta do conjunto de dados utilizado.

Ficaram evidentes as dificuldades causadas pela inconsistência de dados em sua fonte original, o que foi tratado com sucesso na geração do conjunto final de dados utilizados em cada um dos experimentos.

Contudo, assim como foram descobertas regras de associação já de conhecimento no domínio, ou irrelevantes, também foram descobertos indicativos de atividade errônea no sistema de *software* oriundo do domínio do negócio exposto.

Foi demonstrado, também, como algoritmos de classificação podem tratar a interpretação dos dados de maneira puramente estatística, podendo causar confusão ou má interpretação pelos usuários de ferramentas de mineração de dados. Foram identificados casos onde um único atributo foi utilizado para determinar classes, e também como probabilidade estatística pode ser adotada como abordagem algorítmica em caso de grande desproporção entre entidades de cada classe.

Por outro lado, foram descobertos fatores ou atributos determinantes quando da geração efetiva de árvores de decisão no domínio proposto, como o tipo de concepção dos animais e, também, o número de irmãos de mesmo parto.

Foram associadas irregularidades frequentes em duas raças ovinas, o que pode implicar na criação de abordagens administrativas para o tratamento destas, o que implica diretamente no domínio do negócio e melhoramento da ovinocultura.

A geração de agrupamentos foi capaz de apresentar, visualmente, as tendências para correção de irregularidades, de acordo com a idade dos ovinos e, também de acordo com a necessidade de obtenção do certificado de pureza racial.

Considerou-se, portanto, que o conjunto de dados em estudo foi tratado com sucesso nas fases de pré-processamento de dados e foi possibilitada a utilização da ferramenta WEKA através de suas funcionalidades de mineração de dados após o estudo das utilidades disponibilizadas pela própria ferramenta. Finalmente, foram obtidos resultados que serviram como base para o desenvolvimento de conhecimento com aplicabilidade relevante em ovinocultura.

Trabalhos futuros, considerando os resultados obtidos nestes experimentos, podem levar à informação de interessados, além da continuação de projetos de pesquisa na

mesma área, onde existem e sejam disponibilizadas quantidades de dados que possam ser estudadas com o fim de minerar conhecimento.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. **Fast Algorithms for Mining Association Rules in Large Databases**. 20th International Conference on Very Large Data Bases, 1994.

ASSOCIAÇÃO BRASILEIRA DE CRIADORES DE OVINOS – ASSISTÊNCIA AOS REBANHOS DE CRIADORES DE OVINOS. **Regulamento do Serviço de Registro Genealógico de Ovinos no Brasil**. Diário Oficial da União, 2012.

BOUCKAERT, R. et al. **WEKA Manual for Version 3.6.9**. The University of Waikato, 2013.

CAMARGO, S. **Mineração de Regras de Associação no Problema de Cesta de Compras Aplicada ao Comércio Varejista de Confecção**. Universidade Federal do Rio Grande do Sul, 2002.

DATE, C. J. **An Introduction to Database Systems**. 8ª Edição. Pearson Education, 2004.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. Ai Magazines, Vol. 17, 1996.

GANTI, V.; GEHRKE J.; RAMAKRISHNAN, R. **Mining Very Large Datasets**. Institute of Electrical and Electronic Engineers Computer Society, 1999.

HALL, M. et al. The WEKA **Data Mining Software**: An Update. SIGKDD Explorations, Vol, 11, 2009.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining**: Concepts and Techniques. 3ª Edição. Elsevier, 2009.

HARRINGTON, P. **Machine Learning in Action**. Manning, 2012.

LIU, B.; HSU, W.; MA, Y. **Integrating Classification and Association Rule Mining**. 4th International Conference on Knowledge Discovery and Data Mining, 1998.

RAJARAMAN, A.; LESKOVEC, J.; ULLMAN, J. **Mining of Massive Datasets**. 2013.

RAMAKRISHNAN, R.; GEHRKE, J. **Database Management Systems**. 2ª Edição. McGraw-Hill, 2000.

SUMATHI, S.; ESAKKIRAJAN, S. **Fundamentals of Relational Database Systems**. Springer, 2007.

WITTEN, I.; FRANK, E.; HALL, M. **Data Mining Practical Machine Learning Tools and Techniques**. 3ª Edição. Elsevier, 2011.