

FARN - FACULDADE NATALENSE PARA O DESENVOLVIMENTO DO RN

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Aluno : Rafael França da Rocha

**DATA MINING APLICADO EM BANCO DE DADOS DE EMPRESA DE
PRODUTOS DE LIMPEZA**

NATAL/RN
12/01/2011

NATAL/RN
12/01/2011

RAFAEL FRANÇA DA ROCHA

LIGA DE ENSINO DO RIO GRANDE DO NORTE
FACULDADE NATALENSE PARA O DESENVOLVIMENTO DO RIO GRANDE
DO NORTE
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

**DATA MINING APLICADO EM BANCO DE DADOS DE EMPRESA DE
PRODUTOS DE LIMPEZA**

Monografia apresentada como exigência
do Curso de Sistemas de Informação da
Faculdade Natalense para o
Desenvolvimento do Rio Grande do
Norte para obtenção do título de
Bacharel.

Orientador: Ytalo Rosendo do Amaral

NATAL/RN
12/01/2011

DATA MINING APLICADO EM BANCO DE DADOS DE EMPRESA DE
PRODUTOS DE LIMPEZA

Catálogo na Publicação – Biblioteca da FARN
Setor de Processos Técnicos

Rocha, Rafael França da

Data mining aplicado em banco de dados de empresa de produtos de
Limpeza/ Rafael França de Costa - Natal: FARN, 2010.
33f.

Orientador: Ytalo Rosendo do Amaral (Mestre)

Monografia (Bacharel em Sistema de Informação) – Faculdade
Natalense para o Desenvolvimento do Rio Grande do Norte.

1. Sistema de Informação – Monografia. 2. Software – Mineração
de dados – Monografia. I. Título.

RN/FARN/BC

CDU 004.65

BIBLIOTECA UNI-RN 004.65 R672d



2011

ex.1

M002080

(RB=

37266)

Rocha, Rafael França da

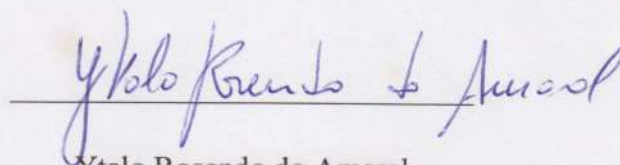
Data mining aplicado em banco de
dados de empresa de produtos de

LIGA DE ENSINO DO RIO GRANDE DO NORTE
FACULDADE NATALENSE PARA O DESENVOLVIMENTO DO RIO GRANDE
DO NORTE
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

RAFAEL FRANÇA DA ROCHA

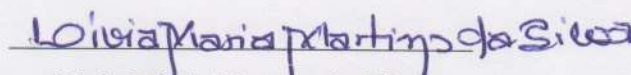
DATA MINING APLICADO EM BANCO DE DADOS DE EMPRESA DE
PRODUTOS DE LIMPEZA

Projeto de Monografia apresentada e aprovada em ____ de ____ de ____, pela banca
examinadora composta pelos seguintes membros:

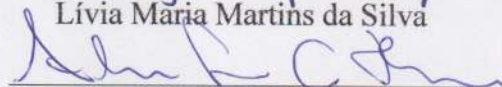


Ytalo Rosendo do Amaral

Orientador



Livia Maria Martins da Silva



Alexandre Luiz G. Damasceno

Natal/RN, 12 de 01 de 2011.

Resumo

A mineração de dados é uma área que está crescendo bastante devido à incapacidade do ser humano de interpretar tamanha quantidade de dados, muita informação e conhecimento, possivelmente úteis, podem estar sendo desperdiçados. Em consequência disso, a necessidade de se desenvolver novas ferramentas e técnicas também crescendo e se mostrando cada vez mais indispensável. Aqui será apresentado o processo da mineração de dados, cujo objetivo é a obtenção de conhecimento útil, para a utilização em um processo de tomada de decisão. A mineração dos dados será feita através do programa weka, que será mostrado um tutorial de utilização do mesmo.

Abstract

Data mining is an area that is growing dramatically due to the inability of humans to interpret so much data, lots of information and knowledge, possibly useful, may be being wasted. As a result, the necessity to develop new tools and techniques also growing and showing more and more indispensable. Here we will present the process of data mining, whose objective is to obtain useful knowledge for use in a process of decision making.

1.2.1. Geral	9
1.2.2. Específicas	9
1.3 Justificativa	13
2 REFERENCIAL TEÓRICO	11
2.1. Mineração de dados	11
2.1.1 Definindo dados, informação e conhecimento	12
2.1.2 O Processo de Mineração de Dados	18
2.2. Knowledge Discovery in Data Bases	17
3 METODOLOGIA	23
3.1. Levantamento bibliográfico	23
3.2. Tipo de Pesquisa	23
3.3. Métodos Utilizados	23
4 ANÁLISE DE RESULTADOS	23
5 CONCLUSÃO	32
6 REFERÊNCIAS BIBLIOGRÁFICAS	33

SUMÁRIO

1 INTRODUÇÃO.....	8
1.2 Objetivos.....	9
1.2.1. Geral.....	9
1.2.2. Específicos.....	9
1.3 Justificativa.....	10
2 REFERENCIAL TEÓRICO.....	11
2.1. Mineração de dados	11
2.1.1 Definindo dados, informação e conhecimento	12
2.1.2 O Processo de Mineração de Dados	14
2.2. Knowledge Discovery in Data Bases.....	17
3 METODOLOGIA.....	23
3.1. Levantamento bibliográfico	23
3.2. Tipo de Pesquisa	23
3.3. Métodos Utilizados	23
4 ANÁLISE DE RESULTADOS.....	23
5 CONCLUSÃO.....	32
6 REFERÊNCIAS BIBLIOGRÁFICAS.....	33

1.2. OBJETIVOS

1.2.1. Objetivo Geral

Monitorar a utilização do data mining através da extração de dados de uma fábrica de produção de lâmpada, buscando demonstrar quais as vantagens de aproveitar-se dessa ferramenta.

1. INTRODUÇÃO

- Obter dados para a partir deles fazer uso da mineração de dados.

Segundo Fayyad (1998) apud Santos et al., o ser humano não está preparado para interpretar grandes volumes de dados e/ou espaços multidimensionais. Entre as diversas alternativas de ferramentas e técnicas é possível destacar o Processo de Descoberta de Conhecimento – Knowledge Discovery in Data Bases (Conhecimento em Banco de Dados), o qual compreende uma das suas etapas denominada: Data Mining.

A mineração de dados surgiu para ser aplicada em pesquisas científicas como também para lucratividade de algumas empresas. Nela são incorporadas várias técnicas de diversas áreas como a inteligência artificial, banco de dados e estatística.

Data mining tem atraído uma grande quantidade de pesquisas e indústrias. É uma ferramenta que nos ajuda a extrair informações úteis de uma grande quantidade de dados que crescem rapidamente, e esses dados estão sendo bastante valorizados pelas empresas, pois podem ser encontradas várias informações contidas em meio a esses dados que seriam de grande importância.

1.2. OBJETIVOS

1.2.1. Objetivo Geral

Mostrar a utilização do data mining através da extração de dados de uma fábrica de produtos de limpeza, buscando demonstrar quais as vantagens de aproveitar-se dessa ferramenta.

1.2.2. Objetivo Específico

- Obter dados para a partir deles fazer uso da mineração de dados;
- Utilizar ferramentas de mineração de dados.
- Exibir qual o resultado alcançado depois da aplicação do data mining.

1.3. JUSTIFICATIVA

A necessidade de tomar decisões é constante em toda e qualquer empresa, seja ela grande ou pequena, porém isso exige um grande conhecimento e informações sobre determinadas áreas, para isso são necessárias ferramentas que sejam capazes de obter essas informações. Essas informações são conseguidas pela inteligência artificial através da análise dos dados, em busca de padrões que possam prever comportamentos futuros permitindo a empresa tomar uma nova decisão baseada em conhecimentos contidos em seus próprios bancos de dados.

O foco central de Mineração de Dados é o de como transformar dados armazenados em conhecimento, expresso em termos de formalismos de representação, tal como regras e relações entre dados. Existe conhecimento que pode ser extraído diretamente de dados para o uso de qualquer técnica, entretanto, existe também muito conhecimento que está de certa forma "enterrado" na Base de Dados, na forma de relações existentes entre itens de dados que, para ser extraído, é necessário o desenvolvimento de técnicas especiais. Assim na próxima seção são definidos os conceitos básicos dados, informação e conhecimento para em seguida abordar o processo de Mineração de Dados como um todo (REZENDE, 2005).

2. REFERENCIAL TEÓRICO

2.1. Mineração de dados

A mineração de dados é uma etapa na descoberta do conhecimento em bancos de dados que consiste no processo de analisar grandes volumes de dados sob diferentes perspectivas, a fim de descobrir informações úteis que normalmente não estão sendo visíveis. Para isso são utilizadas técnicas que envolvem métodos matemáticos, algoritmos e heurísticas para descobrir padrões e regularidades entre os dados pesquisados (BRUSSO APUD GONÇALVES, 1999).

Às várias tarefas desenvolvidas em data mining têm como objetivo primário a predição e a descrição. A predição usa atributos para prognosticar o desconhecido ou os valores futuros de outras variáveis de interesse. A descrição contempla o que foi descoberto nos dados sob o ponto de vista da interpretação humana (FAYYAD apud SANTOS et al., 2006)

O objetivo do processo de Mineração de Dados é a extração do conhecimento implícito por meio da descoberta de padrões e da criação de modelos de maneira automática a partir dos dados (REZENDE, 2005).

A mineração de dados compreende apenas uma das fases do processo Conhecimento em Banco de Dados, referente à aplicação de algoritmos para extração de padrões em um processo de prospecção de conhecimento em bases de dados que envolvem a automação da identificação e do reconhecimento de padrões. Esse processo é um conjunto de procedimentos contínuos que compartilham o conhecimento descoberto a partir de bases de dados (SANTOS, 2006).

O foco central de Mineração de Dados é o de como transformar dados armazenados em conhecimento, expresso em termos de formalismos de representação, tal como regras e relações entre dados. Existe conhecimento que pode ser extraído diretamente de dados sem o uso de qualquer técnica, entretanto, existe também muito conhecimento que está de certa forma “embutido” na Base de Dados, na forma de relações existentes entre itens de dados que, para ser extraído, é necessário o desenvolvimento de técnicas especiais. Assim na próxima seção são definidos os conceitos básicos dados, informação e conhecimento para em seguida abordar o processo de Mineração de Dados como um todo (REZENDE, 2005).

Todo o processo de mineração de dados é orientado em função de seu domínio de aplicação e dos repositórios de dados inerentes aos mesmos. Para usar os dados é necessário que estes estejam estruturados de forma a serem consultados e analisados adequadamente (REZENDE, 2005).

Algoritmos de aprendizado de máquina são amplamente utilizados na tarefa de Mineração de Dados. Estes algoritmos são baseados na construção de árvores de decisão, utilizando-se como matéria prima os dados de treinamento. Através do percurso destas árvores, da raiz até um nó folha é possível inferir classes a dados baseados em determinados valores para cada atributo (SANTOS, 2008).

Outra abordagem para realização da Mineração de Dados utiliza algoritmos de aprendizado de máquina baseados na Computação Evolutiva – os Algoritmos Genéticos. Estes algoritmos realizam buscas globais e permitem uma melhor interação entre atributos que os algoritmos baseados na estratégia gulosa que são geralmente utilizados para a tarefa de Mineração de Dados e por isso são extremamente úteis na procura de padrões relevantes em dados apud Santos(2008).

No final dos anos 90, aprendeu-se como tirar proveito das melhores abordagens da análise estatística clássica, das redes neurais, das árvores de decisão, da análise de mercado e de outras técnicas poderosas, de uma maneira muito mais convincente e eficaz. Adicionalmente, a chegada de sistemas sérios de data warehouse tem sido um ingrediente necessário que fez a mineração de dados algo real e tangível (SANCHES,2003).

2.1.1. Definindo dados, informação e conhecimento

Os conceitos de dados, informação e conhecimento estão interligados. Na Figura 1.1 é mostrada uma representação gráfica do relacionamento entre dados, informação e conhecimento, em função da capacidade de entendimento e da independência de contexto que cada um destes conceitos implica (Kock Jr., McQueen, & Baker; Kock Jr., McQueen, & Corner apud REZENDE,2005).

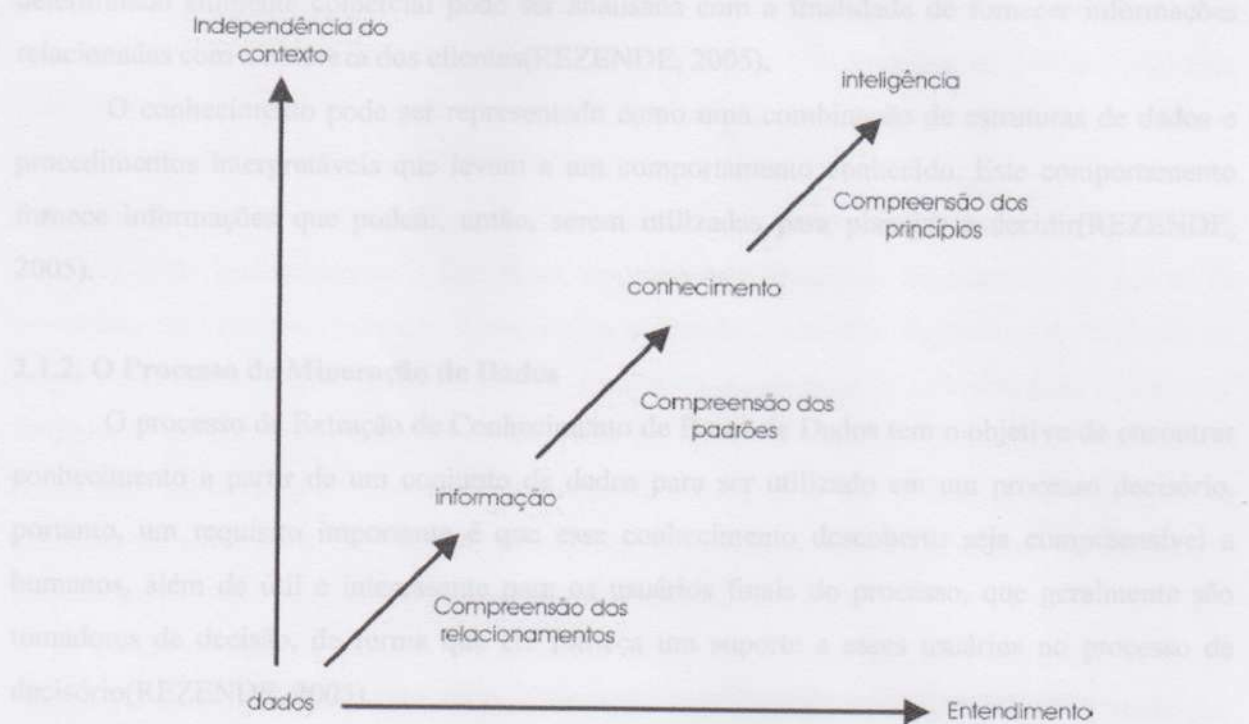


Figura 1.1 Dado, informação e conhecimento. REZENDE (2005)

O dado é um elemento puro, quantificável sobre um determinado evento. Dados são fatos, números, texto ou qualquer mídia que possa ser processada pelo computador. Hoje em dia, as organizações estão acumulando vastas e crescentes quantidades de dados em diferentes formatos e em diferentes tipos de repositórios. Ressalta-se que o dado, por si só, não oferece embasamento para o entendimento da situação (REZENDE, 2005).

A informação é o dado analisado e contextualizado. Envolve a interpretação de um conjunto de dados, ou seja, a informação é constituída por padrões, associações ou relações que todos aqueles dados acumulados podem proporcionar. Por exemplo, a análise de pontos de equilíbrio no mercado pode fornecer informação acerca de quais produtos estão sendo vendidos e a frequência de tais operações (REZENDE, 2005).

A informação pode gerar conhecimento que ajude na análise de padrões históricos para conseguir uma previsão dos fatos futuros (pelo menos no contexto das variáveis que estão sendo envolvidas na análise). Por exemplo, a informação dos dados sumarizados nas vendas de um

determinado ambiente comercial pode ser analisada com a finalidade de fornecer informações relacionadas com a natureza dos clientes(REZENDE, 2005).

O conhecimento pode ser representado como uma combinação de estruturas de dados e procedimentos interpretáveis que levam a um comportamento conhecido. Este comportamento fornece informações que podem, então, serem utilizadas para planejar e decidir(REZENDE, 2005).

2.1.2. O Processo de Mineração de Dados

O processo de Extração de Conhecimento de Bases de Dados tem o objetivo de encontrar conhecimento a partir de um conjunto de dados para ser utilizado em um processo decisório, portanto, um requisito importante é que esse conhecimento descoberto seja compreensível a humanos, além de útil e interessante para os usuários finais do processo, que geralmente são tomadores de decisão, de forma que ele forneça um suporte a esses usuários no processo de decisório(REZENDE, 2005).

Assim, o processo de Extração de Conhecimento de Bases de Dados, ou Mineração de Dados, inicia-se com o entendimento do domínio da aplicação, considerando aspectos como os objetivos dessa aplicação e as fontes de dados (Base de Dados da qual se pretende extrair conhecimento). Em seguida, é realizada uma seleção de dados a partir dessas fontes, de acordo com os objetivos do processo. Os conjuntos de dados resultantes dessa seleção são, então, pré-processados, ou seja, recebem um tratamento para poderem ser submetidos aos métodos e ferramentas na etapa de extração de padrões(REZENDE,2005).

A etapa de extração de padrões tem o objetivo de encontrar modelos (conhecimento) a partir de dados. Normalmente, como resultado dessa etapa, tem-se preditores que, para um dado problema, fornecem uma decisão. Após a etapa de extração de padrões, vem a de pós-processamento, na qual o conhecimento é avaliado quanto a sua qualidade e/ou utilidade para que, em caso positivo, seja utilizado para apoio a algum processo de tomada de decisão (REZENDE, 2005).

É importante notar que, por ser um processo eminentemente iterativo, as etapas da Mineração de Dados não são estanques, ou seja, a correlação entre as técnicas e métodos

utilizados nas várias etapas é considerável, a ponto da ocorrência de pequenas mudanças sem uma delas afetar substancialmente o sucesso de todo o processo. Portanto, os resultados de uma determinada etapa podem acarretar mudanças a quaisquer das etapas posteriores ou, ainda, o recomeço de todo o processo (REZENDE, 2005).

Normalmente, os dados disponíveis para análise não estão em um formato adequado para a Extração de Conhecimento. Além disso, em razão de limitações de memória ou tempo de processamento, muitas vezes não é possível a aplicação direta dos algoritmos de extração de padrões aos dados. Dessa maneira, torna-se necessária a aplicação de métodos para tratamento, limpeza e redução do volume de dados antes de iniciar a etapa de Extração de Padrões. É importante salientar que a execução das transformações deve ser guiada pelos objetivos do processo de extração a fim de que o conjunto de dados gerado apresente as características necessárias para que os objetivos sejam cumpridos (REZENDE, 2005).

O processo, descrito em Fayyad apud Pitoni(2002), consiste em uma seqüência de cinco etapas, partindo dos dados existentes e chegando à descoberta do conhecimento extraído dos mesmos. Essas etapas são exemplificadas a baixo e ilustradas na figura 1.2

- Seleção dos dados: esta etapa consiste em escolher qual o conjunto de dados que será submetido ao processo. Estes dados normalmente ficam armazenados em um armazém de dados.
- Pré-processamento: nesta etapa, os dados podem sofrer uma qualificação, a fim de corrigir erros e inconsistências que poderão existir.
- Transformação: nesta etapa, os dados são convertidos em um formato adequado para serem acessados pelos algoritmos de mineração. É nela que também se realiza uma possível redução no número de variáveis, sumarizando os dados que serão submetidos à mineração.
- Mineração: é a etapa mais importante do processo. É nela que o algoritmo escolhido é aplicado sobre os dados a fim de se descobrir padrões interessantes. É fundamental para que esta etapa obtenha resultados de qualidade a correta aplicação dos passos anteriores.

Interpretação dos dados: nesta última etapa do processo, os resultados obtidos na mineração são interpretados e avaliados, selecionando-se o conhecimento considerado útil.

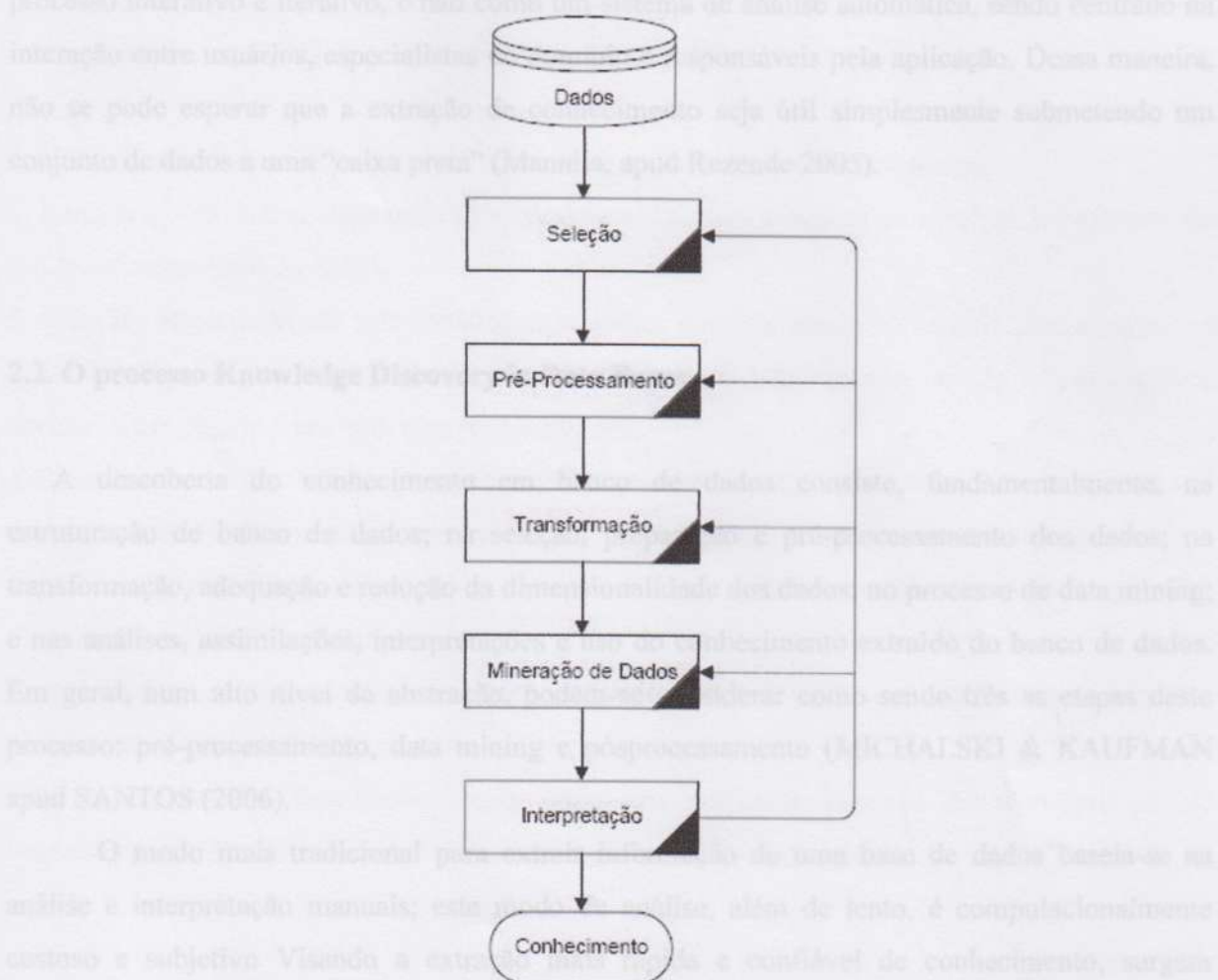


Figura 1.2 O processo de descoberta de conhecimento. Pitoni(2002).

O processo de Extração de Conhecimento de Bases de Dados tem o objetivo de encontrar conhecimento a partir de um conjunto de dados para ser utilizado em um processo decisório. Portanto, um requisito importante é que esse conhecimento descoberto seja compreensível a humanos, além de útil e interessante para os usuários finais do processo, que geralmente são em campos tais como bancos de dados, aprendizado de máquinas, reconhecimento de padrões, estatísticas, inteligência artificial, aquisição de conhecimento para sistemas especialistas,

tomadores de decisão, de forma que ele forneça um suporte a esses usuários no processo de decisório Rezende(2005).

A extração de conhecimento a partir de grande quantidade de dados é vista como um processo interativo e iterativo, e não como um sistema de análise automática, sendo centrado na interação entre usuários, especialistas do domínio e responsáveis pela aplicação. Dessa maneira, não se pode esperar que a extração de conhecimento seja útil simplesmente submetendo um conjunto de dados a uma “caixa preta” (Mannila, apud Rezende 2005).

2.2. O processo Knowledge Discovery in Data Bases

A descoberta do conhecimento em banco de dados consiste, fundamentalmente, na estruturação de banco de dados; na seleção, preparação e pré-processamento dos dados; na transformação, adequação e redução da dimensionalidade dos dados; no processo de data mining; e nas análises, assimilações, interpretações e uso do conhecimento extraído do banco de dados. Em geral, num alto nível de abstração, podem-se considerar como sendo três as etapas deste processo: pré-processamento, data mining e pósprocessamento (MICHALSKI & KAUFMAN apud SANTOS (2006).

O modo mais tradicional para extrair informação de uma base de dados baseia-se na análise e interpretação manuais; este modo de análise, além de lento, é computacionalmente custoso e subjetivo. Visando a extração mais rápida e confiável de conhecimento, surgem métodos automáticos denominados Knowledge Data Discovery, ou simplesmente KDD. O KDD aplica métodos interdisciplinares – especialmente métodos estatísticos e de aprendizado de máquina para extrair conhecimento de alto nível a partir de bases de dados reais e tem a Mineração de Dados como principal tarefa(SANTOS, 2008)

O conhecimento em banco de dados evoluiu, e continua evoluindo, da interseção de pesquisa em campos tais como bancos de dados, aprendizado de máquinas, reconhecimento de padrões, estatísticas, inteligência artificial, aquisição de conhecimento para sistemas especialistas,

visualização de dados, descoberta de máquina, descoberta científica, recuperação de informação, e computação de altodesempenho. Os sistemas de software Conhecimento em Banco de Dados incorporam teorias, algoritmos, e métodos de todos estes campos SANCHES(2003).

Segundo AMO(2004) a descoberta de conhecimento em banco de dados é um processo amplo consistindo das seguintes etapas e ilustrado na figura 1.3:

1. Limpeza dos dados: etapa onde são eliminados ruídos e dados inconsistentes.
2. Integração dos dados: etapa onde diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados.
3. Seleção: etapa onde são selecionados os atributos que interessam ao usuário. Por exemplo, o usuário pode decidir que informações como endereço e telefone não são de relevantes para decidir se um cliente é um bom comprador ou não.
4. Transformação dos dados: etapa onde os dados são transformados num formato apropriado para aplicação de algoritmos de mineração (por exemplo, através de operações de agregação).
5. Mineração: etapa essencial do processo consistindo na aplicação de técnicas inteligentes a fim de se extrair os padrões de interesse.
6. Avaliação ou Pós-processamento: etapa onde são identificados os padrões interessantes de acordo com algum critério do usuário.
7. Visualização dos Resultados: etapa onde são utilizadas técnicas de representação de conhecimento a fim de apresentar ao usuário o conhecimento minerado.

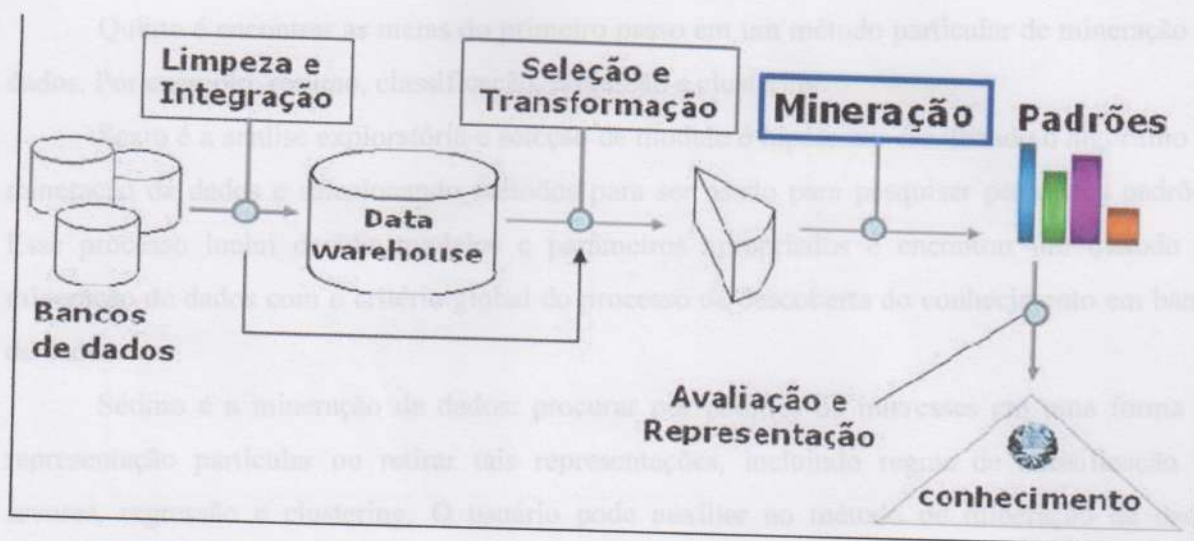


Figura 1.3 As etapas do processo de Knowledge Discovery in Data Bases. Amo(2004).

De acordo com Fayyad, Piatetsky-Shapiro, and Smyth (1996), devem ser seguidos alguns passos do processo serão descritos abaixo e na figura 1.4:

Primeiro é desenvolvido uma compreensão do domínio da aplicação e identificar os objetivos do processo de descoberta do conhecimento em banco de dados do ponto de vista do usuário.

Segundo é selecionar um dado alvo: selecionando os dados ou focando em subconjuntos de variáveis ou exemplos de dados, que a descoberta será realizada.

Terceiro é a limpeza dos dados e pré-processamento. Operação básica incluindo remover “ruídos” se for apropriada, coletar as informações necessárias, decidindo estratégias para lidar com campos de dados perdidos e contar com seqüências de informações.

Quarto é a redução dos dados e a projeção: encontrando características úteis para representar os dados dependendo da meta. Com dimensionalidade, redução ou métodos de transformação, o numero de variáveis efetivas desconsideradas podem ser reduzidas ou representações não variantes podem ser encontradas.

Quinto é encontrar as metas do primeiro passo em um método particular de mineração de dados. Por exemplo, resumo, classificação, regressão e clustering.

Sexto é a análise exploratória e seleção de modelo e hipóteses: escolhendo o algoritmo da mineração de dados e selecionando métodos para ser usado para pesquisar por dados padrões. Esse processo inclui decidir modelos e parâmetros apropriados e encontrar um método de mineração de dados com o critério global do processo de descoberta do conhecimento em banco de dados.

Sétimo é a mineração de dados: procurar por padrões de interesses em uma forma de representação particular ou retirar tais representações, incluindo regras de classificação ou árvores, regressão e clustering. O usuário pode auxiliar no método de mineração de dados realizando corretamente os passos do procedimento.

Oitavo é interpretar padrões encontrados, possivelmente retornando para qualquer um dos passos anteriores para melhor iteração. Esse passo também envolve visualização de padrões extraídos e visualizações de dados recebidos dos modelos extraídos.

Nono é agir na descoberta de conhecimento: usando o conhecimento diretamente, incorporando em outro sistema para ações futuras, ou simplesmente documentando isto e reportando para as partes interessadas. Esse processo também inclui checar e resolver potenciais conflitos com conhecimento prévio.

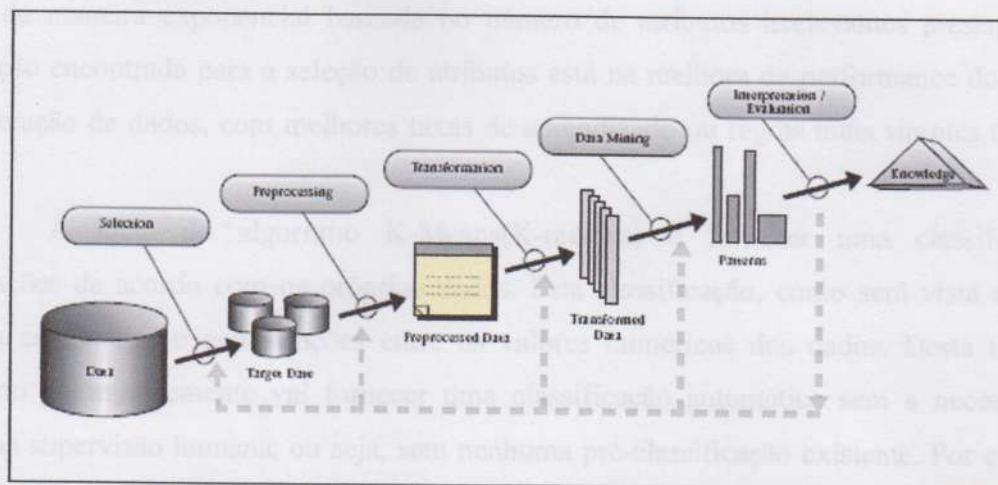


Figura 1.4 Os passos do processo de descoberta. Fayyad, Piatetsky-Shapiro, and Smyth (1996)

É geralmente desejável a intervenção de um analista humano nas etapas citadas anteriormente, a fim de que seu conhecimento sobre as informações relevantes que devem ser extraídas pelo processo de descoberta do conhecimento em banco de dados seja utilizado (FREITAS, 2001)apud Santos.

Após o refinamento dos dados a serem utilizados, a discretização pode ser utilizada para transformar um atributo com valores contínuos em um atributo categórico (categorical) ou nominal, formando intervalos discretos (FREITAS, 2001) apud Santos.

A etapa de seleção de atributos encontra atributos que possuem relevância para o objetivo que deve ser satisfeito pelo processo de descoberta do conhecimento em banco de dados, encontrando um subconjunto de atributos relevantes (atributos capazes de distinguir exemplos pertencentes a classes diferentes) entre os atributos originais do registro para ser utilizado no processo de mineração de dados. A motivação para a realização desta etapa é baseada nos seguintes fatores: algoritmos de mineração de dados que utilizam métodos de indução tem seu tempo de execução aumentado caso haja muitos atributos presentes a presença de atributos irrelevantes ou redundantes pode de alguma maneira levar o algoritmo a uma classificação equivocada dos dados, levando o processo a produzir dados pouco relevantes ou não confiáveis experimentos comprovam que o número de exemplos para garantir certa taxa de classificação

UNI-RN

cresce de maneira exponencial baseada no número de atributos irrelevantes presentes. Outra motivação encontrada para a seleção de atributos está na melhora da performance do algoritmo de mineração de dados, com melhores taxas de aprendizado ou regras mais simples (SANTOS, 2008).

A idéia do algoritmo K-Means(K-médias) é fornecer uma classificação de informações de acordo com os próprios dados. Esta classificação, como será vista a seguir, é baseada em análise e comparações entre os valores numéricos dos dados. Desta maneira, o algoritmo automaticamente vai fornecer uma classificação automática sem a necessidade de nenhuma supervisão humana, ou seja, sem nenhuma pré-classificação existente. Por causa desta característica, o K-Means é considerado como um algoritmo de mineração de dados não supervisionado.(Fonseca,2010)

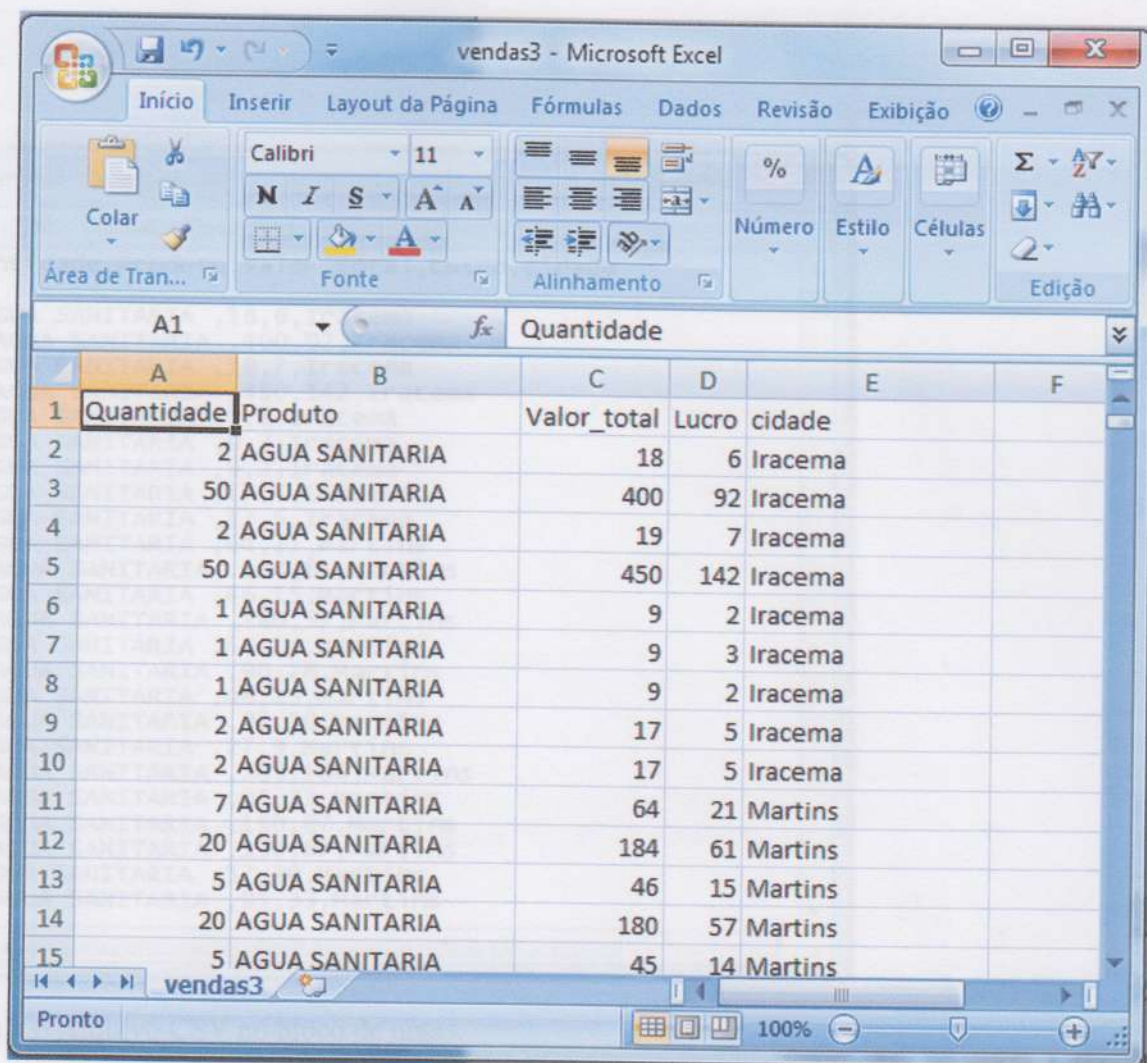
O algoritmo funciona da seguinte maneira, ele vai analisar todos os dados de uma tabela e criar classificações. Isto é, o algoritmo vai indicar uma classe (cluster) e vai dizer quais linhas pertencem a esta classe. O usuário deve fornecer ao algoritmo a quantidade de classes que ele deseja. Este número de classes que deve ser passada para o algoritmo é chamado de k e é daí que vem a primeira letra do algoritmo: K-Means.(Fonseca,2010)

O programa que será utilizado para fazer a mineração de dados será o weka, abaixo mostrarei um pequeno tutorial do funcionamento deste programa.

Primeiramente usamos programas de planilhas eletrônicas e banco de dados os quais permitem exportar os dados em um arquivo onde as vírgulas são os separadores.

Nesse caso foi usado a planilha do Microsoft Excel, para gerar um arquivo onde os dados são separados por vírgula devemos clicar menu arquivo selecionar a opção salvar como e selecionar a opção CSV(separado por vírgula).

O próximo passo é abrir esse arquivo com o bloco de notas, onde a primeira linha representa o cabeçalho das colunas, ou seja, o nome dos atributos sendo separados por vírgula, as linhas seguintes representam os registros onde cada atributo do registro está separado por vírgula.



Microsoft Excel window titled 'vendas3 - Microsoft Excel'. The ribbon shows 'Início', 'Inserir', 'Layout da Página', 'Fórmulas', 'Dados', 'Revisão', and 'Exibição'. The active cell is A1, containing the text 'Quantidade'.

	A	B	C	D	E	F
1	Quantidade	Produto	Valor_total	Lucro	cidade	
2	2	AGUA SANITARIA	18	6	Iracema	
3	50	AGUA SANITARIA	400	92	Iracema	
4	2	AGUA SANITARIA	19	7	Iracema	
5	50	AGUA SANITARIA	450	142	Iracema	
6	1	AGUA SANITARIA	9	2	Iracema	
7	1	AGUA SANITARIA	9	3	Iracema	
8	1	AGUA SANITARIA	9	2	Iracema	
9	2	AGUA SANITARIA	17	5	Iracema	
10	2	AGUA SANITARIA	17	5	Iracema	
11	7	AGUA SANITARIA	64	21	Martins	
12	20	AGUA SANITARIA	184	61	Martins	
13	5	AGUA SANITARIA	46	15	Martins	
14	20	AGUA SANITARIA	180	57	Martins	
15	5	AGUA SANITARIA	45	14	Martins	

Figura 1.5 Arquivo do Microsoft Excel.

Nesse caso foi usado a planilha do Microsoft Excel, para gerarmos um arquivo onde os dados são separados por vírgula devemos clicar menu arquivo selecionar a opção salvar como e selecionar a opção CSV(separado por virgula).

O próximo passo é abrir esse arquivo com o bloco de notas, onde a primeira linha representa o cabeçalho das colunas, ou seja, o nome dos atributos sendo separados por virgula, as linhas seguintes representam os registros onde cada atributo do registro está separado por virgula.

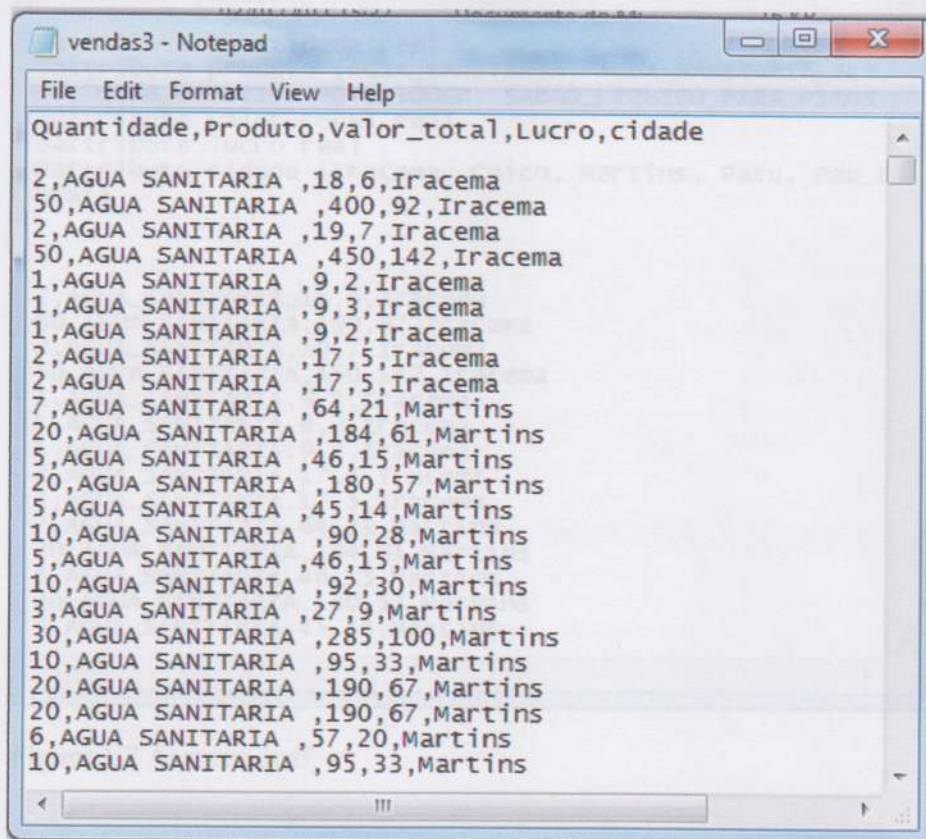
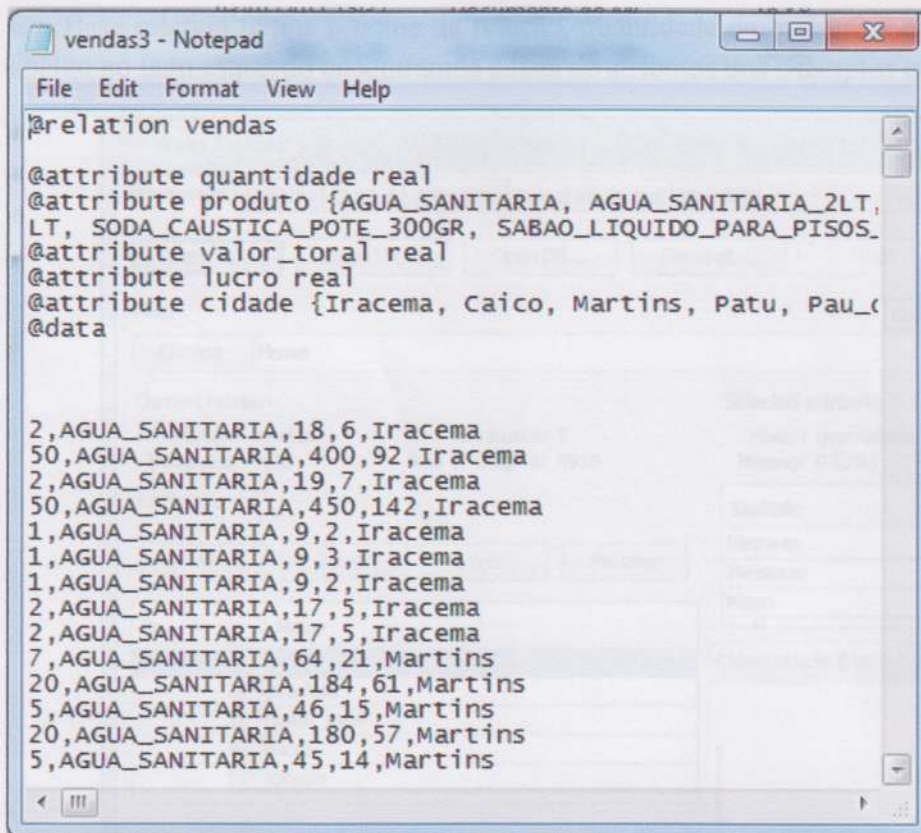


Figura 1.6 Arquivo CSV no bloco de notas.

O passo seguinte é gravar o arquivo, para isso selecionamos no menu Arquivo a opção Salvar como e em Nome do arquivo digite o nome do arquivo com a extensão arff (.arff).



```
@relation vendas

@attribute quantidade real
@attribute produto {AGUA_SANITARIA, AGUA_SANITARIA_2LT,
LT, SODA_CAUSTICA_POTE_300GR, SABAO_LIQUIDO_PARA_PISOS...}
@attribute valor_toral real
@attribute lucro real
@attribute cidade {Iracema, Caico, Martins, Patu, Pau...}
@data

2,AGUA_SANITARIA,18,6,Iracema
50,AGUA_SANITARIA,400,92,Iracema
2,AGUA_SANITARIA,19,7,Iracema
50,AGUA_SANITARIA,450,142,Iracema
1,AGUA_SANITARIA,9,2,Iracema
1,AGUA_SANITARIA,9,3,Iracema
1,AGUA_SANITARIA,9,2,Iracema
2,AGUA_SANITARIA,17,5,Iracema
2,AGUA_SANITARIA,17,5,Iracema
7,AGUA_SANITARIA,64,21,Martins
20,AGUA_SANITARIA,184,61,Martins
5,AGUA_SANITARIA,46,15,Martins
20,AGUA_SANITARIA,180,57,Martins
5,AGUA_SANITARIA,45,14,Martins
```

Figura 1.7 Arquivo .arff.

Depois de salvar com a extensão arff, devemos adicionar @relation nome_do_conjuto_de_dados, para cada atributo usa @attribute, e após colocar uma linha com @data e logo em seguida os dados em si, como na figura 1.7, Agora podemos inicializar a análise dos dados usando o algoritmo apropriado, assumindo que já tenhamos instalado o Weka.

Após a instalação, inicializamos o WEKA onde será mostrada uma tela inicial. Para executarmos um dos algoritmos do pacote WEKA pressionamos o botão *Explorer*, a janela *Weka Knowledge Explorer* será aberta, devemos então carregar os dados para serem analisados os quais podem ser originados de um arquivo (*Open file...*) de uma URL (*Open URL...*) ou ainda de um banco de dados (*Open DB...*)

No exemplo utilizaremos dados originados de um arquivo, para tanto pressionamos o botão *Open file...* localizamos o arquivo *clima.arff* criado anteriormente com mostra a Figura 1.8.

Em Base relation temos o nome da relação, quantidade de instancias e quantidade de atributos, abaixo no lado esquerdo encontramos a lista de atributos dos exemplos que carregamos.

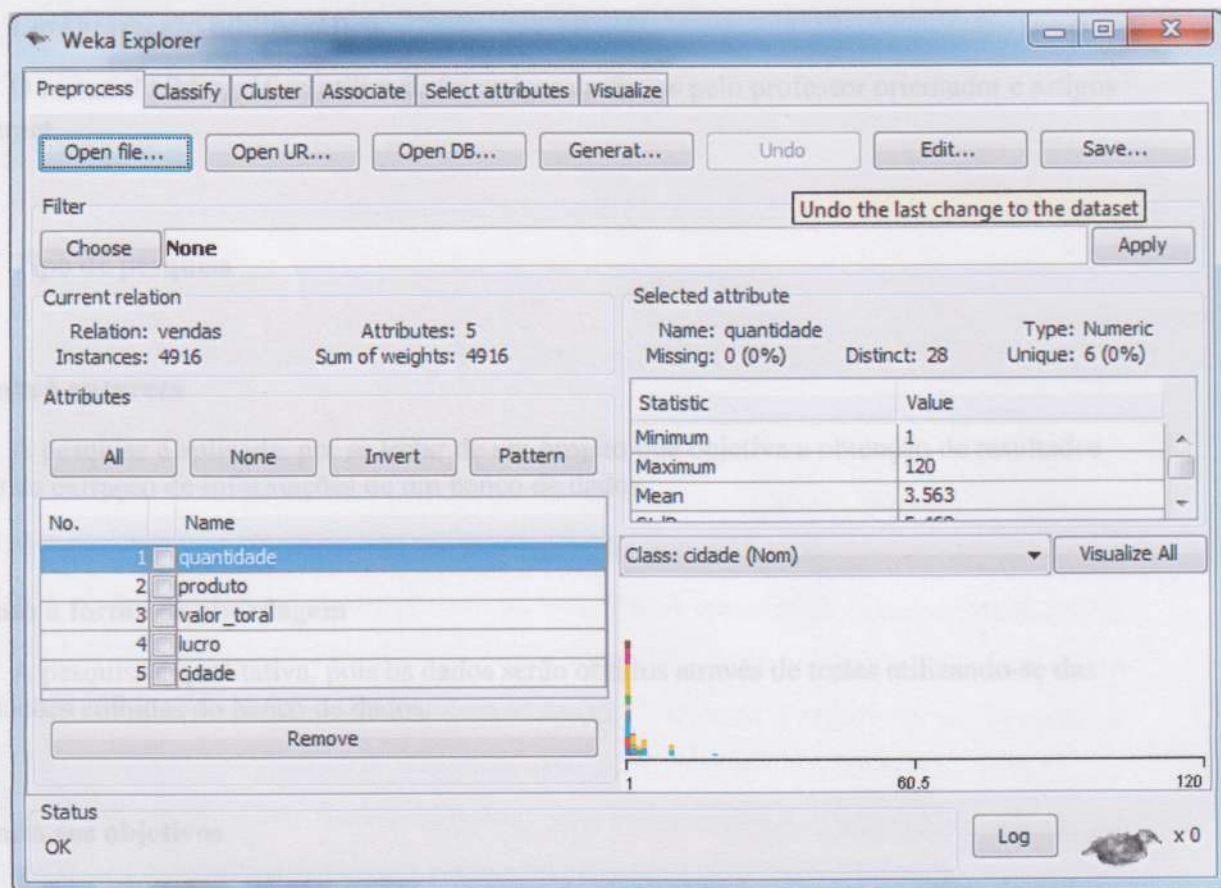


Figura 1.8 Abrindo arquivo no weka.

Na parte superior se encontramos as seguintes abas Preprocess onde se podemos abrir, editar e salvar a base, Classify conjunto de algoritmos que implementam os esquemas de aprendizagem que funcionam como classificadores, Cluster contem os algoritmos para geração de grupos, Associate conjunto de algoritmos para gerar regras de associação, Select attributes determina a relevância dos atributos, Visualise explora os dados.

3.3. Métodos utilizados

Estudo da teoria de mineração de dados, a partir das referências teóricas que foram estudadas nos artigos pesquisados, para maior conhecimento de suas capacidades.

3. METODOLOGIA

3.1. Levantamento bibliográfico

O material bibliográfico utilizado foi artigos enviados pelo professor orientador e artigos da Internet.

3.2. Tipo de pesquisa

- Quanto à natureza

A pesquisa é aplicada, por se tratar de um projeto que objetiva a obtenção de resultados através da extração de informações de um banco de dados.

- Quanto à forma de abordagem

A pesquisa é qualitativa, pois os dados serão obtidos através de testes utilizando-se das informações colhidas do banco de dados.

- Quanto aos objetivos

A pesquisa é descritiva, por envolver o uso de técnicas padronizadas de coleta de dados.

- Quanto aos procedimentos técnicos

O procedimento técnico será estudo de caso, pois será feito um estudo em cima de um banco de dados buscando mais detalhes sobre o mesmo.

3.3. Métodos utilizados

Estudo da teoria da mineração de dados, a partir das referências teóricas que foram estudadas nos artigos pesquisados, para maior conhecimento de suas capacidades.

4. Análise de Resultados

Análise dos dados obtidos utilizando ferramentas do Data mining mostrando todas as informações conseguidas através de sua extração, a fim de que o projeto apresente o resultado buscado.

Nesse projeto irei trabalhar com a parte de cluster, mais precisamente com o algoritmo SimpleKMeans.

Para gerar as classes e classificar as ocorrências, o algoritmo faz uma comparação entre cada valor de cada linha por meio da distância. Geralmente utiliza-se a distância euclidiana para calcular o quão 'longe' uma ocorrência está da outra. A maneira de calcular esta distância vai depender da quantidade de atributos da tabela fornecida. Após o cálculo das distâncias o algoritmo calcula centróides para cada uma das classes. Conforme o algoritmo vai iterando, o valor de cada centróide é refinado pela média dos valores de cada atributo de cada ocorrência que pertence a este centróide. Com isso, o algoritmo gera k centróides e coloca as ocorrências da tabela de acordo com sua distância dos centróides. Abaixo será mostrado o resultado que foi encontrado, como exemplo.

kMeans

=====

Number of iterations: 12

Within cluster sum of squared errors: 10.729291009046023

Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (4916)	Cluster#	
		0 (4618)	1 (298)
quantidade	3.5633	2.6416	17.8456
lucro	21.2864	16.4318	96.5168

Clustered Instances

0	4618 (94%)
1	298 (6%)

Figura 1.9 Resultado entre quantidade e lucro.

No total foram usados 5 atributos(quantidade, lucro, valor_total, produtos e cidades) dentre os quais apenas 2 foram escolhidos para a clusterização, aqui foi criado 2 centroids que foram nomeados como 0 e 1, na figura acima é mostrado a quantidade total de instancias que é 4916, a media total de quantidade 3,5633 e lucro 21,2864, no centroid 0 foram encontradas 4618 instancias que corresponde a 94% do total e o centroid 1 298 instancias correspondente a 6% do total, também mostra que o centroid 0 está centrado no ponto 2.6416 da quantidade e 16.4318 do lucro e o centroid 1 está na posição 17.8456 da quantidade e 96.5168 do lucro.

Com essas informações nos podemos afirmar que, apesar de vendas quantidades altas próximos ao valor de 17.8456 termos um bom lucro, nos obtemos mais lucro aonde a quantidade se aproxima de 2.6416.

```

kMeans
=====

Number of iterations: 3
Within cluster sum of squared errors: 7773.0
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute                               Full Data                               Cluster#
                                      (4916)                                0              1
                                      (4577)                                (4577)          (339)
=====
produto                               DESINFETANTE_LAVANDA_2LT DESINFETANTE_VIOLETA_2LT DESINFETANTE_LAVANDA_2LT
cidade                               Apodi                               Apodi           Limoeiro_do_norte

Clustered Instances

0      4577 ( 93%)
1      339 ( 7%)

```

Figura 1.10 Resultado entre produto e cidade.

Na figura 1.10 temos o exemplo feito com 2 atributos do tipo string, aqui ele nos mostra que o produto mais vendido é o desinfetante lavanda 2 litros e a cidade que mais comprou que foi Apodi. No centroid 0 está localizado a 93% do total de instancias e a cidade que mais aparece é Apodi e o produto que foi mais vendido lá é o desinfetante violeta 2 litros, já no centroid 1 a cidade que mais aparece é o limoeiro do norte que teve mais vendas de desinfetante lavanda 2 litros.

Apesar de ter sido mostrado exemplos com apenas 2 atributos por vez, podem ser usados mais atributos se for necessário.

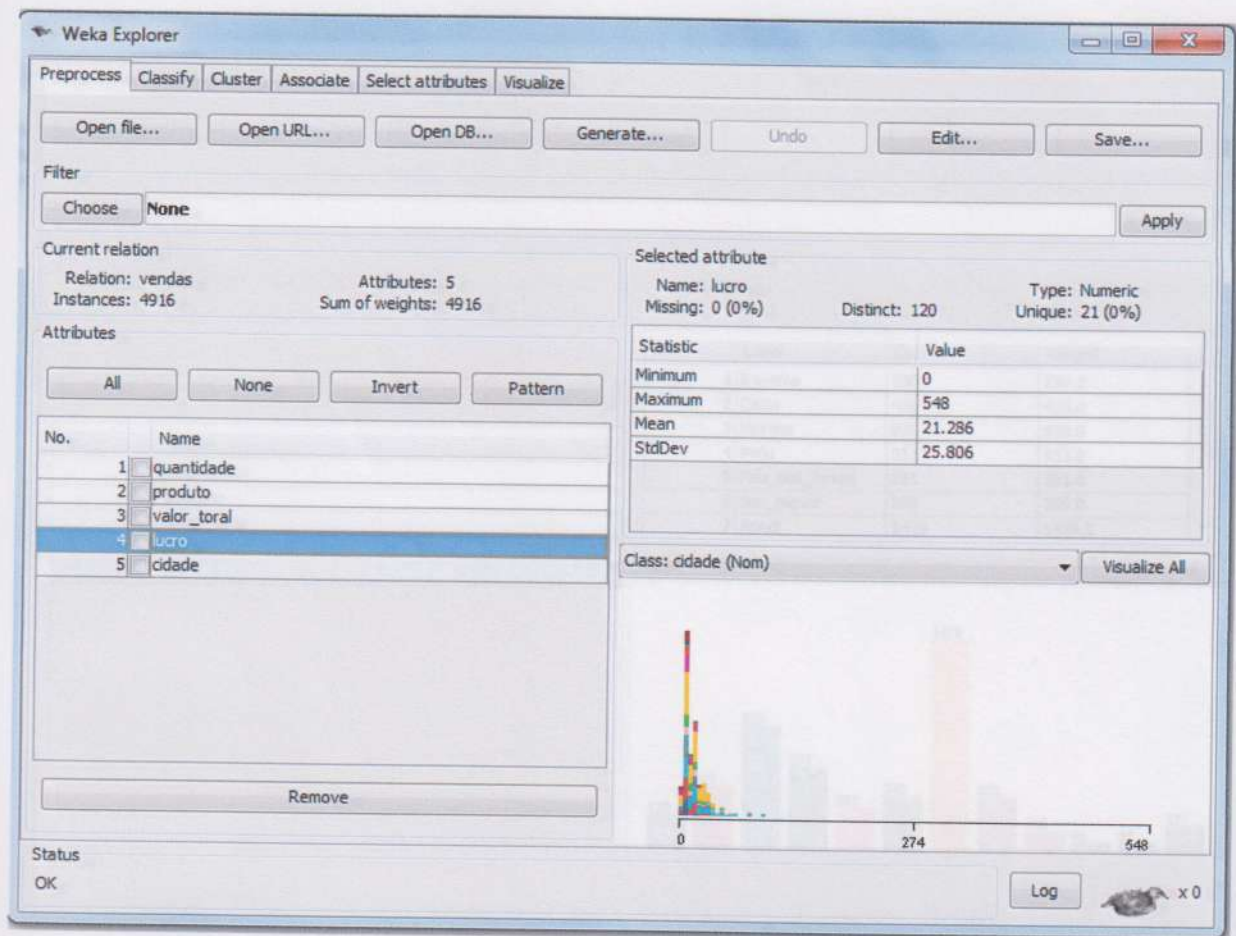


Figura 1.11 Pagina principal do weka.

Na figura 1.11 é mostrado a pagina inicial depois que é exportado o arquivo. Aqui nos já podemos ver alguns dos dados antes mesmo de ser feita a clusterização, acima pode ser visto o atributo lucro selecionado e a direita algumas informações sobre ele, tais como, valor mínimo que é 0, valor Maximo 548, média 21.286, desvio padrão 25.806, e abaixo desses resultados podemos ver um pequeno gráfico desses valores.

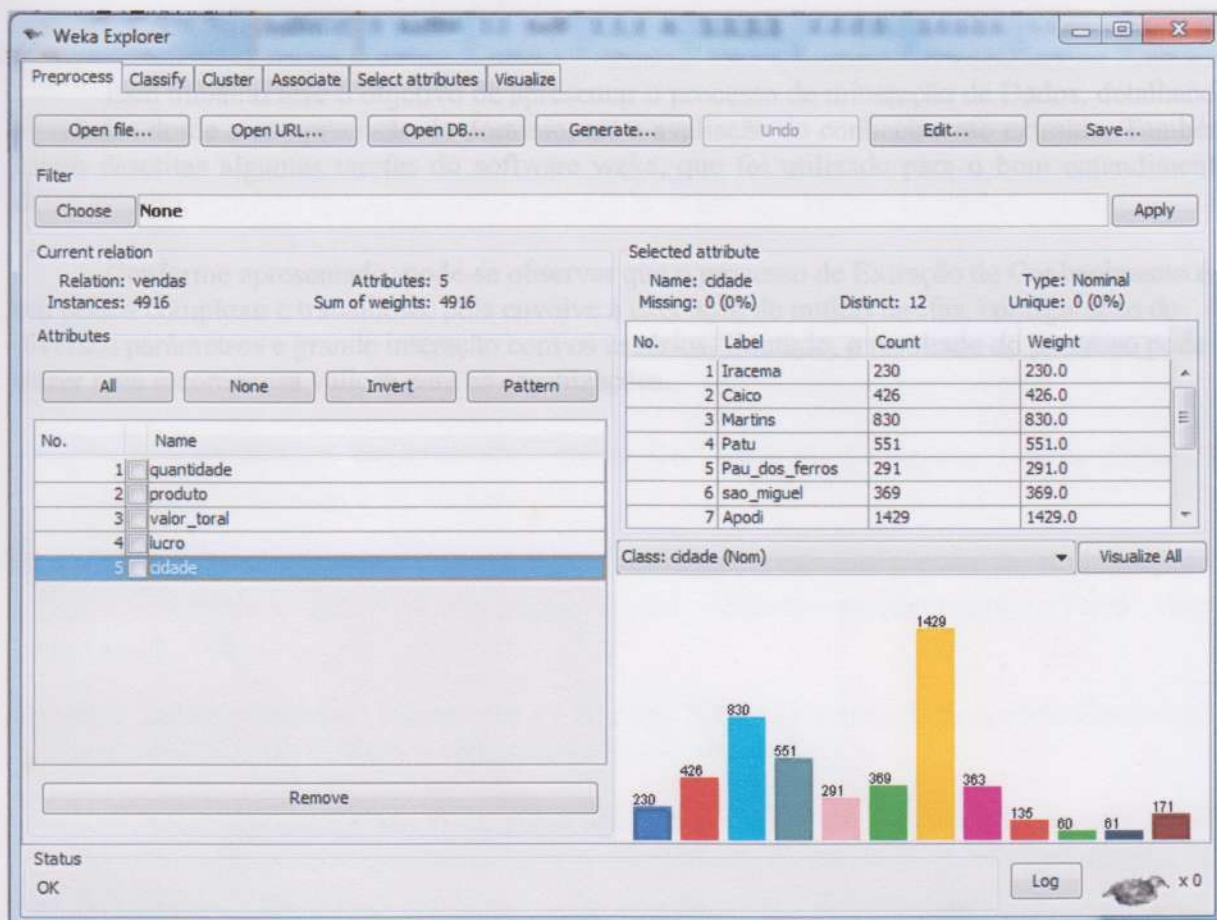


Figura 1.12 Gráfico weka.

Na figura 1.12 podemos ver um exemplo de um atributo em string. Aqui vemos que são mostrados os nomes de cada cidade, são 12 cidades diferentes, e na contagem podemos ver quantas vezes aparecem os nomes de cada uma nas instancias, no gráfico é possível ver detalhadamente.

5. Conclusão

Este trabalho teve o objetivo de apresentar o processo de mineração de Dados, detalhando atividades desde a compreensão do domínio, até a avaliação do conhecimento extraído. Também foram descritas algumas tarefas do software weka, que foi utilizado para o bom entendimento desse assunto.

Conforme apresentado, pode-se observar que o processo de Extração de Conhecimento é um pouco complexo e trabalhoso, pois envolve a execução de muitas tarefas, configuração de diversos parâmetros e grande interação com os usuários. Contudo, o resultado do processo pode trazer uma recompensa valiosa para as organizações.

Kirby, Richard. WEKA Explorer User Guide for Versão 3.3.4. Departamento de Ciência da Computação da universidade de Waikato 2002.

Pront, Rafael Moreira. Mineração de Regras de Associação nos Canais de Informação do Direto. Universidade Federal de Rio Grande do Sul - trabalho de diplomação, Porto Alegre, setembro de 2002.

Rezende, Solange Oliveira. Mineração de Dados. XXV Congresso da Sociedade Brasileira de Computação 2005. UNISINOS - São Leopoldo/RS, julho de 2005.

Sanches, Andre Rodrigo. Uma visão geral sobre Mineração de Dados. Universidade de São Paulo Instituto de matemática e estatística - Pós-Graduação em Ciência da Computação, São Paulo - Novembro de 2003.

Santos, Celso Bilyntkevycz dos et al. Data Mining aplicado na base de dados dos produtos mais vendidos de uma supermercado. XXVI ENLCEP - Fortaleza, CE, Outubro de 2006.

Santos, Ivairton Monteiro. Data Warehose como ferramenta de apoio em sistemas de monitoramento ambiental. Universidade Federal de Mato Grosso - Cuiabá-MT, Março de 2003.

Santos, Joilson Santos. Mineração de Dados Utilizando Algoritmos Genéticos. Universidade Federal de Bahia - Salvador - Bahia 2008.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Amo, Sandra de. **Técnicas de Mineração de Dados**. XXIV Congresso da Sociedade Brasileira de Computação. Jornada de Atualização em Informática. Salvador, Agosto 2004.

Fonseca, Felipe Cesar Stanzani; Beltrame, Walber Antônio Ramos. **Aplicações Práticas dos Algoritmos de Clusterização Kmeans e Bisecting K-means**. Universidade Federal do Espírito Santo (UFES) - Vitória – ES 2010

GONÇALVES, Lóren Pinto Ferreira. **Mineração de dados em supermercados: o caso do supermercado “tal”**. Universidade Federal do Rio Grande do Sul – Escola de Administração – Programa de Pós Graduação em Administração. Porto Alegre, dezembro de 1999.

Kirkby, Richard. **WEKA Explorer User Guide for Versio 3.3.4**. Departamento de Ciência da computação da universidade de Waikato 2002.

Pitoni, Rafael Moreira. **Mineração de Regras de Associação nos Canais de Informação do Direto**. Universidade federal do rio grande do sul – trabalho de diplomação . Porto Alegre, setembro de 2002.

Rezende, Solange Oliveira. **Mineração de Dados**. XXV Congresso da Sociedade Brasileira de Computação 2005. UNISINOS – São Leopoldo/RS, julho de 2005.

Sanches, Andre Rodrigo. **Uma visão geral sobre Mineração de Dados**. Universidade de São Paulo instituto de matemática e estatística – Pós-Graduação em Ciência da Computação. São Paulo – Novembro de 2003.

Santos, Celso Bilynkiewicz dos et al. **Data Mining aplicado na base de dados das categorias mais vendidas de um supermercado**. XXVI ENEGEP - Fortaleza, CE, Outubro de 2006.

Santos, Ivairton Monteiro. **Data Warehouse como ferramenta de auxílio em sistemas de monitoramento ambiental**. Universidade Federal de Mato Grosso - Cuiabá-MT
Março de 2003

Santos, Joilma Souza. **Mineração de Dados Utilizando Algoritmos Genéticos**. Universidade Federal da Bahia - Salvador – Bahia 2008