UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
ESPECIALIZAÇÃO EM BANCO DE DADOS

DANILO BONIFÁCIO TELES
FABRICIO NOGUEIRA DOS SANTOS
LEANDRO PEDROSA

A Mineração de Dados

Aplicada à Inteligência de Negócios

Goiânia
2014

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
ESPECIALIZAÇÃO EM BANCO DE DADOS

AUTORIZAÇÃO PARA PUBLICAÇÃO DE MONOGRAFIA EM FORMATO ELETRÔNICO

Na qualidade de titular dos direitos de autor, **AUTORIZO** o Instituto de Informática da Universidade Federal de Goiás – UFG a reproduzir, inclusive em outro formato ou mídia e através de armazenamento permanente ou temporário, bem como a publicar na rede mundial de computadores (*Internet*) e na biblioteca virtual da UFG, entendendo-se os termos "reproduzir" e "publicar" conforme definições dos incisos VI e I, respectivamente, do artigo 5º da Lei nº 9610/98 de 10/02/1998, a obra abaixo especificada, sem que me seja devido pagamento a título de direitos autorais, desde que a reprodução e/ou publicação tenham a finalidade exclusiva de uso por quem a consulta, e a título de divulgação da produção acadêmica gerada pela Universidade, a partir desta data.

Título: A Mineração de Dados – Aplicada à Inteligência de Negócios

Autor(a): Danilo Bonifácio Teles
Fabricio Nogueira dos Santos
Leandro Pedrosa

Goiânia, 17 de Abril de 2014.

Danilo Bonifácio Teles
Fabricio Nogueira dos Santos
Leandro Pedrosa
– Autor

Edmundo Spoto – Orientador

Leandro Luís Galdino de Oliveira – Co-Orientador

DANILO BONIFÁCIO TELES
FABRICIO NOGUEIRA DOS SANTOS
LEANDRO PEDROSA

A Mineração de Dados

Aplicada à Inteligência de Negócios

Monografia apresentada ao Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do Certificado de Especialização em Computação.

Área de concentração: Banco de Dados.

Orientador: Prof. Edmundo Spoto

Co-Orientador: Prof. Leandro Luís Galdino de Oliveira

Goiânia
2014

DANILO BONIFÁCIO TELES
FABRICIO NOGUEIRA DOS SANTOS
LEANDRO PEDROSA

A Mineração de Dados

Aplicada à Inteligência de Negócios

Monografia apresentada no Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás como requisito parcial para obtenção do Certificado de Especialização em Computação, aprovada em 17 de Abril de 2014, pela Banca Examinadora constituída pelos professores:

Prof. Edmundo Spoto
Instituto de Informática – UFG
Presidente da Banca

Prof. Leandro Luís Galdino de Oliveira
Universidade Federal de Goiás – UFG

Prof. Sergio T. Carvalho
Universidade Federal de Goiás – UFG

Profa. Daiane Dias da Silva
Universidade Federal de Goiás – UFG

Aos familiares e amigos que nos incentivaram e apoiaram nossas decisões.

Agradecimentos

Primeiramente aos nossos familiares pelo apoio e compreensão nos momentos difíceis. Ao Prof. Dr. Edmundo Sérgio Spoto pela orientação, encaminhamento dado a este trabalho. A todos os professores que contribuíram para nosso crescimento pessoal e profissional. Nossos sinceros agradecimentos.

A maneira como você coleta, gerencia e utiliza as informações determina se você vai vencer ou perder.

Bill Gates,

.

Resumo

Bonifácio Teles Danilo,Dos Santos Nogueira Fabricio, Pedrosa Leandro. **A Mineração de Dados**. Goiânia, 2014. 62p. Monografia de Especialização. Especialização em Banco de Dados, Instituto de Informática, Universidade Federal de Goiás.

Técnicas sobre a Inteligência de Negócio têm sido um importante tópico de estudo. A descoberta de conhecimento em banco de dados e o processo de mineraÃ§Ã£o de dados destaca-se por ser um conjunto de processos que une o uso do poder de processamento das máquinas atuais, com avançados algoritmos de inteligência artificial, projetados para analisar uma grande massa de dados com a finalidade de se descobrir informações que até então eram desconhecidas ou desconsideradas. O objetivo deste estudo foi identificar as diferentes técnicas de utilizaçã e manipulaão de grandes massas de dados para a geração de conhecimento com ganhos reais para indivíduos e corporações, com tempo e custo adequados. A metodologia empregada é um estudo bibliográfico, onde a busca por periódicos relevantes, livros e revistas científicas sobre o tema foi realizada. O principal resultado deste estudo é identificar as ferramentas que permitem a descoberta do conhecimento, após análise de grande volume de dados.

Palavras–chave

Inteligência de negócios. Descoberta de conhecimento em banco de dados. Mineração de dados.

Abstract

Bonifácio Teles Danilo,Dos Santos Nogueira Fabricio, Pedrosa Leandro. **T**. Goiânia, 2014. 62p. Monografia de Especialização. Especialização em Banco de Dados, Instituto de Informática, Universidade Federal de Goiás.

Techniques on Business Intelligence has been an important topic of study. knowledge discovery in database and the process of Data Mining stands out for being a tool that combines the use of the processing power of today's machines, with advanced artificial intelligence algorithms designed to analyze a large body of data for the purpose of discovering information that hitherto were unknown or disregarded. The aim of this study was to identify the different techniques to use and manipulate large amounts of data to generate knowledge with real gains for individuals and corporations , with adequate time and cost. The methodology used was a literature study , where the search for relevant books and journals on the subject journals was performed . The main result of this study was to identify the tools that enable knowledge discovery , after analyzing large volumes of data.

Keywords

Business Intelligence. knowledge discovery in database. Data Mining.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

A necessidade humana fez com que o homem se destacasse das outras espécies em nosso planeta por entenderem o meio em que vivem e desenvolverem formas de facilitar sua sobrevivência. Isso tem, ao longo do tempo, proporcionado o acúmulo de conhecimento e a geração de tecnologias que evoluem a passos largos, principalmente nos dois últimos séculos.

Essa evolução cada vez mais acelerada trouxe consigo o aumento da quantidade de dados que já não podem ser mais simplesmente transmitidos entre pessoas ou armazenados em meios que dificultem o seu acesso. Esta necessidade fez com que sistemas fossem projetados para estes fins. Assim, toda essa massa de dados não era trabalhada de forma a gerar o principal: O conhecimento. Tal conhecimento pode ser extraído por meio de técnicas de Inteligência de Negócio, as quais serão apresentadas ao longo deste documento, tendo como foco principal o processo de Descoberta em conhecimento em banco de dados.

A Mineração de Dados destaca-se como uma das mais interessantes e inovadoras formas de analisar e localizar padrões em uma grande massa de dados extraindo informações e gerando conhecimentos para os níveis estratégicos em diversas áreas. Ela permite que, através do uso de técnicas avançadas de inteligência artificial e poderosos recursos de hardware disponíveis, seja possível realizar uma profunda busca em grandes massas de dados por informações que, pela limitação humana, podem ser ignoradas ou até mesmo desconhecidas.

Em nossos estudos sobre banco de dados constatamos que em todo projeto de sistemas de informação o gerenciamento de dados é considerado como um dos ponto principais em sua elaboração e construção, no entanto, estes dados nem sempre recebem a devida atenção e em boa parte das corporações funcionam apenas como fonte para sistemas transacionais.

Logo, percebe-se que havendo um tratamento nestes dados de forma adequada, eles podem se tornar uma grande fonte de conhecimento, direcionando as corporações a se aperfeiçoarem em seus processos e sistemas, visando obter o máximo de ganho, seja ele tangível ou não.

Neste contexto, é possível exemplificar que através de uma simples análise da carteira de clientes de uma determina organização pode-se descobrir quais de seus atuais clientes possuem potencial para adquirir outras linhas de seus produtos e oferecê-los de forma a atender necessidades previamente manifestadas por estes clientes dentro de um perfil pré-estabelecido.

A Inteligência de Negócio e a descoberta do conhecimento em banco de dados devem entrar diretamente no processo da construção de uma solução para o negócio de médias a grandes corporações. O correto tratamento dos dados em sua forma bruta, sua contextualização e a extração do conhecimento auxiliados pelas técnicas aqui estudadas podem levar qualquer domínio de conhecimento a otimização de seus processos e às melhores tomadas de decisões.

CAPÍTULO 2

Inteligência de negócio

> Apresentar formas de se trabalhar os dados de forma a obter o melhor resultado dos mesmos, através de técnicas de mineração de dados é necessário estudar antes de tudo a aplicação mais ampla das informações. Ou seja, o processo de Inteligência de negócio baseia-se na transformação de dados em informações, depois em decisões e finalmente em ações. [34].

Logo, esse estudo foi iniciado pelo tópico Inteligência de negócio que abrange várias ferramentas com foco no uso de informações para apoio à tomada de decisões.

2.0.1 Histórico

O termo Inteligência de negócio foi criado pela empresa Gartner na década de 80, entretanto sua aplicação já existia há alguns séculos, por exemplo, a sociedade do Egito Antigo que observava as cheias do Rio Nilo e a utilizavam como referência para determinar ações que deveriam efetuadas.

> A sociedade do Oriente Médio antigo utilizava os princípios básicos de Inteligência de negôcio quando cruzavam informações obtidas junto à natureza em beneficio de suas aldeias. Analisar o comportamento das marés, os períodos chuvosos e de seca, a posição dos astros, entre outras, eram formas de obter informações que seriam utilizadas para tomar decisões importantes, permitindo a melhoria de vida de suas respectivas comunidades[10].

Como exemplo é possível citar a sociedade do Egito antigo que observava as cheias do rio Nilo, e a utilizavam como referência para determinar ações que deveriam efetuadas.

Segundo o Departamento de Hidráulica e Saneamento da Universidade Federal da Bahia esta observação dos ciclos que ocorriam com o rio e o mapeamento de suas variações, juntamente com seus períodos de ocorrência, proporcionou se definir as melhores

épocas para plantio e até mesmo a criação de um "nilômetro" que era uma escala acompanhada exclusivamente pelos sacerdotes para a leitura do nível do rio Nilo, as variações do nível do rio nesta escala determinavam inclusive qual taxa de imposto a ser cobrada.

Figura 2.1: *Variações do nível das águas do rio Nilo. fonte:[30]*

2.1 Dado, informação, conhecimento e decisão

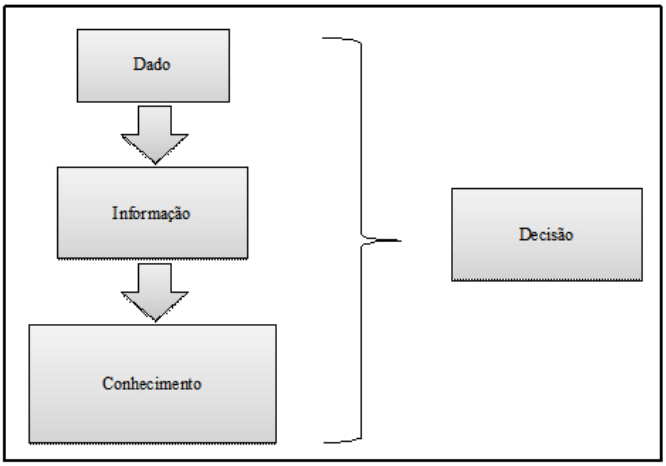

Figura 2.2: *Representação da relação entre dado, informação, conhecimento e decisão. fonte:[10]*

A Inteligência de negócio aborda como será a tomada de decisão e o embasamento que será utilizado para tal, isso é feito baseado em informações que são extraídas de dados existentes na corporação. Assim, faz-se necessário definir o que é dado, informação, conhecimento e o foco principal que é a decisão.

2.1.1 Dado

Dado é o valor bruto sem contexto; seria o que se armazena propriamente no banco de dados.

Analogia: Se dissermos para um grupo de pessoas: Imaginem o dado como sendo apenas o número 2, algumas pessoas desse grupo podem perguntar, mas 2 o quê? Não há nada mais para se complementar é só um dado sem um contexto.

2.1.2 Informação

Informação é a contextualização de um valor bruto, ou seja, é o resultado do tratamento do dado, comparado, classificado ou relacionado a outros dados.

Analogia: O 2 é apenas um número, quando colocamos: "A equipe de basquete da cidade fez apenas 2 pontos na partida". Com essa informação as pessoas já conseguem entender melhor sobre do que se trata o assunto, estamos acrescentando algo a mais nos dados, ele não é apenas um valor seja numérico, caracteres alfa numéricos ou datas e assim, pulamos do nível de dado para o nível de informação.

2.1.3 Conhecimento

Conhecimento é quando se aplica a informação recebida em algum contexto e através da contextualização dessa informação, obtemos o conhecimento e com o conhecimento gerado através das informações, temos o auxílio na tomada de decisões.

Analogia: Se a equipe de basquete da cidade fez apenas 2 pontos isso significa que eles não foram tão eficientes na partida, ou seja, nós tínhamos apenas o número 2, que se trata de um dado, depois obtivemos a informação de que a equipe de basquete fez apenas 2 pontos na partida, a informação foi gerada e agora, se vamos discutir o desempenho da equipe no jogo, essa informação passa a ser conhecimento, afinal de contas, aplicamos a informação dentro de um contexto. Dessa forma podemos tirar como conhecimento que a equipe não foi bem pois não jogaram em sua melhor forma física ou a equipe adversária foi totalmente superior a nossa equipe, podemos extrair também que a equipe fez apenas 1 cesta na partida que resultou em 2 pontos.

Estes três itens são a base para atingir o principal deles: A tomada de decisão. É justamente neste ponto que a correta análise dos dados pode ser o diferencial para se ter uma informação de qualidade, fidedigna e integra.

CAPÍTULO 3

Armazém de dados

Um armazém de dados pode ser definido como um conjunto de dados destinado a auxiliar em decisões de negócios. Seus dados não são voláteis, pois uma vez carregados não podem mais sofrer alterações.

Cada conjunto de dados, ao ser carregado em um Armazém de dados representa um determinado tempo que o identifica dentre os demais e fica associado a uma visão instantânea e sumarizada dos dados operacionais que corresponde ao momento de carga dos dados, para se realizar análise e tendências.

Para se definir uma forma mais objetiva é importante fazer uma comparação com o conceito tradicional de banco de dados, um banco de dados que, nada mais é do que uma coleção de dados operacionais armazenados e utilizados pelo sistema de aplicações de um domínio específico, podem ser chamados de dados "operacionais"ou "primitivos".

Os bancos de dados operacionais armazenam as informações necessárias para as operações diárias do domínio, são utilizados para registrar e executar operações pré-definidas, por isso seus dados podem sofrer constantes mudanças conforme as necessidades atuais. Por não ocorrer redundância nos dados e as informações históricas não ficarem armazenadas por muito tempo, este tipo de banco de dados não exige grande capacidade de armazenamento.

Com base nestes conceitos pode-se concluir que o Armazém de dados não é um fim, mas sim um meio que as empresas dispõem para analisar informações históricas, podendo utilizá-las para a melhoria dos processos atuais e futuros.

Resumindo as principais características de um Armazém de dados tem-se:

- **Orientado a assuntos**: por exemplo, vendas de produtos a diferentes tipos de clientes, atendimentos e diagnósticos de pacientes, rendimento de estudantes;
- **Integrado**: diferentes nomenclaturas, formatos e estruturas das fontes de dados precisam ser acomodadas em um único esquema para prover uma visão unificada e consistente da informação;
- **Séries temporais**: o histórico dos dados por um período de tempo superior ao usual em BDs transacionais permite analisar tendências e mudanças.

- **Não volátil**: Os dados de uma Armazém de dados não são modificados como em sistemas transacionais (exceto para correções), mas somente carregados e acessados para leituras, com atualizações apenas periódicas.

Observe na tabela 3.1 uma relação de funcionalidades de um Armazém de dados e um Banco de operacional e suas diferenças:

Tabela 3.1: *Comparação entre DW e Bancos de dados fonte: [10]*

Características	**Banco de dados Operacionais**	***Data Warehouse***
Objetivo	Operações diárias do negócio	Analisar o negócio
Uso	Operacional	Informativo
Tipo de processamento	OLTP	OLAP
Unidade de trabalho	Inclusão, alteração e exclusão	Carga e consulta
Números de usuários	Milhares	Centenas
Tipo de usuários	Operadores	Comunidade gerencial
Interação do usuário	Somente pré-definida	Pré-definida e ad-hoc
Volume	Megabytes – Gigabytes	Gigabytes – Terabytes
Histórico	60 a 90 dias	5 a 10 anos
Granularidade	Detalhados	Detalhados e resumidos
Redundância	Não ocorre	Ocorre
Estrutura	Estática	Variável
Manutenção desejada	Mínima	Constante
Acesso de registros	Dezenas	Milhares
Atualização	Contínua (Tempo real)	Periódica (*Em batch*)
Integridade	Transação	A cada atualização
Número de índices	Poucos/Simples	Muitos/Complexos
Intenção dos índices	Localizar um registro	Aperfeiçoar consultas

Para organizar os dados em um Armazém de dados, são necessários novos métodos de armazenamento, estruturação e novas tecnologias para a geração e recuperação dessas informações, essas tecnologias diferem dos padrões operacionais de sistemas de banco de dados em três maneiras:

- Disponibilizam visualizações informativas, pesquisando, reportando e modelando capacidades que vão além dos padrões de sistemas operacionais frequentemente oferecidos;
- Armazenam dados frequentemente em formato de cubo *On-line Analytical Processing* (OLAP)[1] multidimensional, permitindo rápida agregação de dados e detalhamento das análises (*drilldown*, *drill trought* etc.);

[1]O OLAP (On-line Analytical Processing) oferece uma alternativa diferente. Voltado para a tomada de decisões, proporciona uma visão dos dados orientados à análise, além de uma navegação rápida e flexível. O OLAP recebe dados do OLTP para que possa realizar as análises e essa carga de dados acontece conforme a necessidade da empresa e sendo um sistema de tomada de decisão não realiza transações (INSERT, UPDATE, DELETE) [25].

- Dispõem de habilidade para extrair, tratar e agregar dados de múltiplos sistemas operacionais em *Data Marts* ou *Data Warehouses* separados.

A ferramenta de extração dos dados é uma parte muito importante do projeto do Armazém de dados, mas apenas uma pequena parcela de um conjunto bastante complexo de soluções de *hardware* e *Software*, e somente depois de definido e projetado o escopo do projeto e depois de construído o repositório de dados, é que se deve chegar às ferramentas de *front-end* responsáveis pelo meio de campo entre as bases de dados e os usuários finais da área executiva.

3.1 *Mercado de dados*

O mercado de dados é um sub-conjunto do armazém de dados, ou seja, o agrupamento ou sumarização dos dados em assuntos específicos e relacionados. Numa visão comparativa dos dados, onde considera-se os requisitos, escopo, integração, tempo, agregação, análise e dados voláteis, percebe-se que a diferença está no escopo, pois enquanto o armazém de dados é feito para atender um domínio como um todo, o mercado de dados é criado para atender um sub-conjunto desse domínio. Observe que atender um sub-conjunto pode significar reunir dados de outros setores, já que, na prática, raramente um único setor possui ou gera toda informação que precisa.

3.2 Extração, Transformação e Carga

O processo de extração, transformação e carga, referênciado por sua sigla em inglês ETL, é utilizado para efetuar a carga dos dados em um Armazém de dados. A primeira parte desse processo é a extração de dados dos sistemas que podem partir de diferentes fontes, podem ser bases de dados relacionais em ambientes heterogêneos, arquivos de texto, planílhas eletrônicas, bases de dados não relacionais etc., ou seja, qualquer fonte de dados que atenda o domínio.

O estágio de transformação aplica uma série de regras ou funções aos dados extraídos para derivar os dados a serem carregados. Algumas fontes de dados necessitarão de pouca manipulação de dados. Em outros casos, podem ser necessários um ou mais de um dos seguintes tipos de transformação:

- Seleção de apenas determinadas colunas para carregar (ou a seleção de nenhuma coluna para não carregar);
- Tradução de valores codificados (se o sistema de origem armazena 1 para sexo masculino e 2 para feminino, mas o Data Warehouse armazena M para masculino e F para feminino, por exemplo), o que é conhecido como limpeza de dados;

- Codificação de valores de forma livre (mapeando “Masculino”,“1” e “Sr.” para M, por exemplo);
- Derivação de um novo valor calculado (montanteVendas iqual à quantidade mutiplicado pelo preço Unitário, por exemplo);
- Junção de dados provenientes de diversas fontes;
- Resumo de várias linhas de dados (total de vendas para cada loja e para cada região, por exemplo);
- Geração de valores de chaves substitutas (*surrogate keys*);
- Transposição ou rotação (transformando múltiplas colunas em múltiplas linhas ou vice-versa);
- Quebra de uma coluna em diversas colunas (como por exemplo, colocando uma lista separada por vírgulas e especificada como uma cadeia em uma coluna com valores individuais em diferentes colunas).

[23].

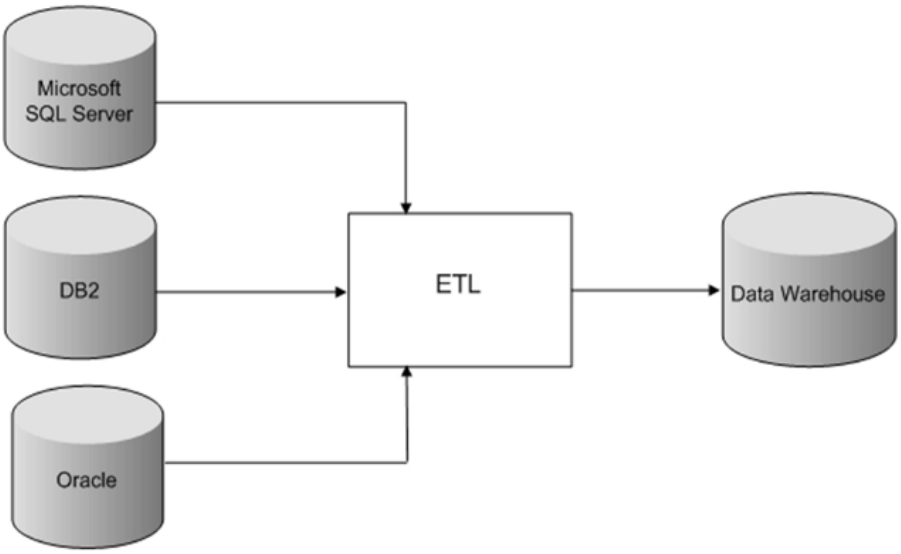

Figura 3.1: *Fluxo geral de ETL. fonte: [37]*

O processo de carga de um *Data Warehouse* (*DW*) varia amplamente, dependendo das necessidades da organização. As informações existentes de alguns *Data Warehouses* podem ser substituídas semanalmente, com dados cumulativos e atualizados. No entanto outros *DW* (ou até mesmo partes do mesmo *DW*) podem ter dados adicionados a cada hora. A temporização e o alcance de reposição ou acréscimo constituem opções de projeto estratégicas que dependem do tempo disponível e das necessidades de negócios. Sistemas mais complexos podem manter um histórico e uma pista de auditoria de todas as mudanças sofridas pelos dados.

3.3 *Data Warehousing*

Para se criar relações de dados de clientes, por exemplo, é necessário um grande número de dados, estes dados são obtidos de um grande banco de dados, um *Data Warehouse* como citado anteriormente nas características gerais de um *Data Warehouse*, entretanto agora serão abordados aspectos mais relevantes para a mineração de dados. Somente a partir destes dados será possível utilizar as técnicas avançadas de inteligência artificial e estatística, ou seja, o Mineração de dados.

> A cada dia dados estão sendo gerados. Quando se realiza uma compra por telefone, por exemplo, o número telefônico, a duração da chamada, o número do cartão de crédito, o endereço da entrega, o produto escolhido e outros dados como nível sociocultural, preferências e *hobbies*, podem ser facilmente adquiridos e armazenados em bancos de dados.
>
> A filosofia empresarial dirigida ao cliente considera cada item de informação sobre o cliente, cada interação em pontos de venda, cada chamada ao serviço de atendimento ao cliente e cada visita a uma página da *world wide web* (www) como uma oportunidade de obter dados e aprender sobre o cliente.
>
> Certamente que obter dados não significa aprender sobre o cliente! De fato, muitas empresas armazenam Gigabytes de dados ocupando espaço e sem aprender nada acerca dos seus clientes e produtos. Nestes casos, os dados são armazenados para fins operacionais, como controle de estoque e cobrança, e após seu uso são simplesmente descartados sem a consideração de que podem representar fonte de informação valiosa para a empresa. [13].

A automatização que existe hoje em todos os processos, juntamente com a ajuda dos recursos computacionais faz com que grandes quantidades de dados sejam geradas, e uma empresa para analisar seus dados, precisa encontrar uma forma de unificar e apresentar estes dados quando necessário de forma organizada.

> Para que a mineração de dados seja realizada, é necessário o acesso a uma massa de dados limpa, consistente e unificada em sua linguagem e lógica. Certamente que analistas vêm realizando Mineração de dados há muitos anos se utilizando de ferramentas simples e bancos dedados separados, porém a construção de um *Data Warehouse* em muito facilita o processo de mineração de dados e de decisão. [13].

3.3.1 A organização dos dados

Conforme já mencionado anteriormente os dados são o ponto mais importante das técnicas de Mineração de dados e do processo de *Data Warehousing*. Eles podem ser classificados de várias formas, entretanto segundo [13] o interessante para os propósitos da mineração é em termos de abstração.

> Quanto maior a abstração do dado, menor o seu volume disponível. Abstração tem relações fortes com hierarquia. Na camada mais inferior desta hierarquia, encontramos o dado operacional em grandes quantidades. Cada produto comprado, cada transação bancária, cada ligação telefônica é um dado operacional.
>
> Normalmente, a maior dificuldade do processo de *Data Warehousing*: é unificar os sistemas de aquisição destes dados operacionais que são utilizados para diferentes propósitos, com diferentes plataformas de hardware e diferentes *Softwares* de aquisição. Além disso, as mudanças nos dados operacionais são muito comuns devido à, por exemplo, introdução de novos produtos, expansão de usuários e introdução de novas tecnologias.
>
> O próximo nível na hierarquia da abstração é o dado resumido cuja função é fornecer uma fotografia que reúna de maneira condensada a informação sobre um determinado grupo volumoso de dados, como, por exemplo, os dados de um setor, produto ou filial da empresa em um certo instante ou período. Se os dados operacionais mudam, os dados resumidos não podem ser facilmente comparáveis. Uma das funções do *Data Warehouse* é permitir que os dados resumidos sejam mais estáveis.
>
> O nível de abstração imediatamente mais alto é o do modelo de dados que nos informa sobre que dados dispomos, como é sua relação com os outros dados e suas formas ou tipos. É com esta informação e com a matéria-prima dos dados operacionais que construímos os dados resumidos. Subindo mais um nível em sua hierarquia, encontramos a informação sobre o próprio dado, ou seja, o metadado.[13].

Ainda segundo [13] Metadado é um modelo lógico do dado com suas entidades, atributos e relações significativas para o analista que realiza a mineração de dados e na para o analista de sistema que desenvolve os programas operacionais da organização.

> Metadado é uma abstração do dado. É o dado de alto nível que descreve o dado de baixo nível. Metadado é o instrumento que transforma dado "cru"em conhecimento. Pode ser útil pensar em metadado como uma "pinça"com que se pode tratar o dado cru. Por exemplo, é o metadado, na forma de definição

de campo, que informa que uma dada cadeia de bits é um endereço de cliente, parte de uma imagem fotográfica ou parte do código de um programa de computador.[14].

Um sistema eficiente de metadados permite ao usuário final do *Data Warehouse* visualizar seu conteúdo entender com seus diferentes elementos se relacionam, identificar onde se encontra cada tipo de informa e adicionalmente ganhar confiança nas informações nele contidas. Finalmente o nível mais elevado de abstração presente em nossa hierarquia dos dados são as especificações das propriedades gerais e das restrições do sistema de dados.

3.3.2 As arquiteturas do *Data Warehouse*

Várias foram as arquiteturas propostas para um *Data Warehouse* ao longo de sua história, começando pelo *Middleware* que provê uma interface comum para uma rede de sistemas de aquisição de dados distribuída pela organização e fora dela. Normalmente, estes sistemas não possuem os atributos necessários para a Mineração de dados. Os sistemas de *Data Warehouse* setorizados que fornecem dados de interesse específico de um setor da empresa, são mais úteis, porém sua proliferação pode gerar problemas de comunicação e inconsistência nos dados. A direção mais promissora para o *Data Warehousing* parece apontar para os sistemas *multitired* capazes de reconhecer diferentes plataformas das fontes de dados, transportar os dados para um repositório central, limpá-los, descrevê-los em metadados e ainda permitir o acesso fácil através de ferramentas específicas (*Data Marts*) definidas em função das necessidades de cada usuário ou grupo.[13].

3.4 Data Ware Housing

3.4.1 Arquiteturas alternativas

Arquitetura Global

A arquitetura global traz a ideia de ter as informações para tomada de decisão de todos os departamentos da empresa. A arquitetura global pode ser fisicamente centralizada ou distribuída:

- Centralizada-instalado e utilizado em um único local
- Distribuída-pode estar espalhado em diversos locais físicos, como por exemplo, em diversas filiais

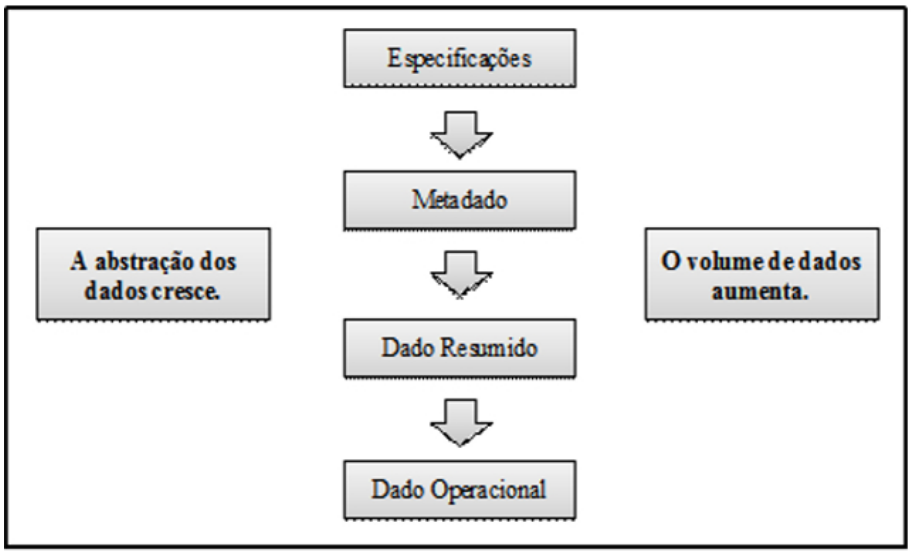

Figura 3.2: *Arquitetura do Data Warehouse geral. fonte: [13]*

[18].

Ainda segundo [18] esta arquitetura a implantação é demorada e complexa administração dos dados, aumentando a necessidade de mais profissionais para participarem do projeto, assim aumentado os valores a serem investidos.

Arquitetura de *Data Mart* Independente

> Uma arquitetura de *Data Mart* independente consiste em uma série de *Data Marts* independentes que são controlados por um grupo particular específico e são construídos especificamente para atender as suas necessidades. Pode existir, portanto, nenhuma conectividade entre esses *Data Marts*.[34]

Ou seja neste tipo de arquitetura os *Data Marts* são categorizados de forma que os dados existentes em cada um deles é referente somente ao interesse do setor a que ele se destina.

Este tipo de implementação tem um impacto mínimo no tempo gasto e na alocação de recursos, porém a integração mínima e a falta de uma visão mais global dos dados podem representar algum tipo de empecilho.

Segundo [18] este tipo de arquitetura traz uma implementação mais rápida e com menos recursos, pois o *Data Mart* é instalado para atender um departamento específico e contém somente os dados extraídos do sistema daquele departamento, não possui integração com demais *Data Marts* os custos envolvidos são menores. Como a ideia principal do *DW* é ter informações gerenciais, neste caso de arquitetura somente seriam tomadas decisões no âmbito departamental.

Arquitetura de *Data Marts* Integrados

Parecem com a Arquitetura Global, porém, os *Data Marts* são separados por departamento, gerando os seus próprios dados e conectados através de interfaces de integração que possibilitam a troca de informações entre eles. Normalmente possuem alguns cadastros duplicados, como, por exemplo, o cadastro de clientes que acaba sendo duplicados em todos os *Data Marts* dos departamentos.

No formato global isso não ocorreria, pois estaria tudo centralizado, mas isso pode ser contornado tratando de forma clara as rotinas de integração.

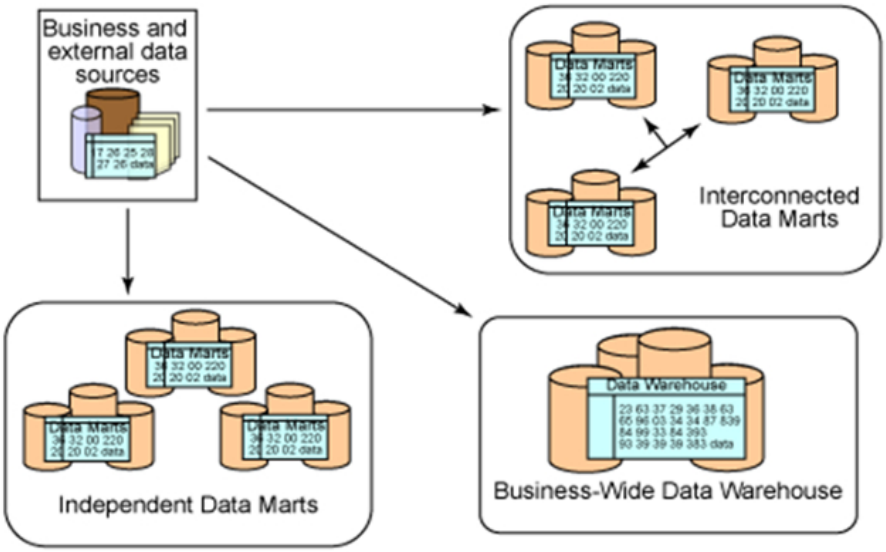

Figura 3.3: *Arquiteturas alternativas de um Data Warehouse. fonte: [33]*

3.5 Dados

Até o momento foi abordado o conceito geral de Mineração de dados, o conceito e as arquiteturas de um *Data Warehouse* que é o que provê os dados necessários para a mineração de dados, ou seja, a matéria prima que será utilizada. Entretanto estes dados possuem características que serão abordadas a partir de agora.

3.6 Os tipos de dados

Segundo [40] os conjuntos de dados diferem de diversas formas, como por exemplo, os atributos usados para descrever objetos de dados podem ser de diferentes tipos - quantitativos ou qualitativos - e podem ter características especiais.

> Alguns conjuntos de dados contêm séries de tempos ou objetos com relacionamentos explícitos entre si, e isso determina quais ferramentas e técnicas podem ser usadas para analisar os dados. Além disso, novas pesquisas em mineração de dados são muitas vezes guiadas pela necessidade de acomodar novas áreas de aplicações e seus novos tipos de dados. Os dados muitas vezes estão longe da perfeição e embora a maior parte das técnicas de mineração de dados pode tolerar algum nível de imperfeição nos dados, um foco na compreensão e melhora da qualidade dos dados geralmente melhora a qualidade das análises resultantes. [40].

Um conjunto de dados muitas vezes pode ser visto como uma coleção de objetos de dados, outros nomes para um objeto de dados são registros, ponteiros, vetores, padrões, eventos, casos, exemplos, observações ou entidades. Por sua vez, objetos de dados são descritos por um número de atributos que capturam as características básicas de um objeto, como a massa de um objeto físico ou o tempo no qual um evento tenha ocorrido. Conforme Tan, Steinbach e Kumar (2009) outros nomes para um atributo são variável, característica, campo, recurso ou dimensão.

> Exemplo: Muitas vezes, um conjunto de dados é um arquivo, no qual os objetos são registros (ou linhas) no arquivo e cada campo (ou coluna) corresponde a um atributo. Por exemplo, a seguir mostra um conjunto de dados que consiste de informações sobre alunos. Cada linha corresponde a um aluno e cada coluna é um atributo que descreve algum aspecto de um aluno, como a média (GPA) ou número de identificação (ID). (TAN; STEINBACH; KUMAR, 2009, p. 3). [40].

Tabela 3.2: *Dados de exemplo contendo informações de alunos. Fonte: [40]*

ID Aluno	**Ano**	*Medida GPA*	⋮
	⋮		
1034262	Terceiro	3,24	⋮
1052663	Segundo	3,51	⋮
1082246	Primeiro	3,62	⋮
⋮	⋮	⋮	⋮

Embora conjuntos de dados baseados em registros sejam comuns, tanto em arquivos horizontais ou quanto em sistemas de bancos de dados relacionais, há outros tipos importantes de conjuntos de dados e sistemas para armazenamento de dados. Neste sentido, será abordado alguns dos tipos de conjuntos de dados que são comumente encontrados na mineração dedados.

3.6.1 Atributos e medidas

O que é um atributo? Segundo [40] “Um atributo é uma propriedade ou característica de um objeto que pode variar, seja de um objeto para outro ou de tempo para outro”.

Por exemplo, a cor dos olhos varia de pessoa para pessoa, enquanto que a temperatura de um objeto varia com o tempo. Observe que a cor dos olhos é um atributo simbólico com um pequeno número de valores possíveis marrom, preto, azul, verde, castanho, etc., enquanto que a temperatura é um atributo numérico com um número potencialmente ilimitado de valores.

> No nível mais básico, os atributos não se relacionam com números ou símbolos. Entretanto, para discutir e analisar com maior precisão as características de objetos atribuímos números ou símbolos a eles. Para fazer isso de uma forma bem definida, precisamos de uma escala de medição. [40].

Ainda para os atributos os autores tem uma segunda definição:

> Uma escala de medição é uma regra (função) que associa um valor numérico ou simbólico a um atributo de um objeto. Formalmente, o processo de medição é a aplicação de uma escala de medida associada a um determinado atributo de um objeto específico.
>
> Embora isto possa parecer um pouco abstrato, nos dedicamos ao processo de medição todo o tempo. Por exemplo, subimos em uma balança de banheiro para determinar nosso peso, classificamos alguém como sendo do sexo masculino ou feminino ou contamos o número de cadeiras em uma sala para ver se haverá assentos para todas as pessoas que vierem à uma reunião. Em todos esses casos, o "valor físico"de um atributo de um objeto é mapeado para um valor numérico ou simbólico. [40].

Com esta fundamentação, torna-se possível discutir o tipo de um atributo, um conceito importante para determinar se uma técnica de análise de dados específica é consistente com um determinado tipo de atributo.

3.6.2 O tipo de um atributo

A determinação do tipo de um atributo é importante para que se possa determinar qual ação será mais adequada a ele e até mesmo se ele deve fazer parte de uma ação, conforme afirma Tan, Steinbach e Kumar (2009) a seguir:

> O tipo de atributo nos informa quais propriedades do mesmo são refletidas nos valores usados para medi-lo. Conhecer o tipo de um atributo é importante porque nos informa quais propriedades dos valores medidos são consistentes com as propriedades correspondentes do atributo e, portanto, evita que executemos ações tolas, como calcular a média das IDs dos funcionários. Observe que é comum se referir ao tipo de um atributo como o tipo de uma escala de medição. [40].

Exemplo: (Idade e Número de ID do Funcionário). Dois atributos que poderiam ser associados a um funcionário são sua ID e idade (em anos), ambos os atributos podem ser representados como números inteiros. Entretanto, embora seja razoável falar em média de idade de um funcionário, não faz sentido falar em média de ID do funcionário. De fato, o único aspecto dos funcionários que pretende-se capturar com o atributo ID é que eles são distintos e consequentemente, a única operação válida para IDs de funcionários é testar se são iguais.

Para o atributo idade, as propriedades dos números inteiros usadas para representar a idade são bem as propriedades do atributo. Mesmo assim, a correspondência (Relação de Inteiros e Idades) não é completa, já que, por exemplo, idades têm um máximo, enquanto que números inteiros não.

3.6.3 Tipos de conjunto de dados

Dados de registro: Grande parte do trabalho de mineração de dados supõe que o conjunto de dados seja uma coleção de registros (objetos de dados), cada um dos quais consistindo de um conjunto fixo de campos de dados (atributos). Dados em registros geralmente são armazenados em arquivos horizontais ou em bancos de dados relacionais. Bancos de dados relacionais são certamente mais do que uma coleção de registros, porém a mineração de dados muitas vezes não usa algumas das informações adicionais disponíveis em um banco de dados relacional, ao invés disso, o banco de dados serve como um lugar conveniente para encontrar registros.

Dados de transação: É um tipo especial de dados em registros, onde cada registro (transação) envolve um conjunto de itens. Considere uma mercearia, o conjunto de produtos comprados por um cliente durante uma ida à mercearia constitui uma transação, enquanto que os produtos individuais que foram comprados são os itens.

Este tipo de dado é chamado de dado de cesta de mercado porque os itens em cada registro são os produtos em uma "cesta de mercado"de uma pessoa. Dados de transação é uma coleção de conjuntos de itens, mas podem ser vistos como um conjunto de registros cujos campos são atributos assimétricos.

Com maior frequência, os atributos são binários, indicando se um item foi comprado ou não, porém, mais comumente, os atributos podem ser discretos ou contínuos, como o número de itens comprados ou a quantia gasta nesses itens.

Tabela 3.3: *Dados em registros. Fonte: [40]*

TID	Reembolso	Estado Civil	Renda tributável	Tomador de empréstimo em dívida
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

Tabela 3.4: *Dados de transação. Fonte: [40]*

TID	Itens
1	Pão, Refrigerante, Leite
2	Cerveja, Pão
3	Cerveja, Refrigerante, Fralda, Leite
4	Cerveja, Pão, Fralda, Leite
5	Refrigerante, Fralda, Leite

3.7 Qualidade dos dados

Aplicações de mineração de dados são muitas vezes aplicadas a dados que foram coletados para outro propósito, ou para aplicações futuras, porém não especificadas. Por este motivo, a mineração de dados geralmente não pode aproveitar os significativos benefícios de “abordar questões de qualidade na fonte".

> Em comparação, grande parte das estatísticas lida com o projeto de experimentos ou pesquisas que obtêm um nível pré-especificado de qualidade de dados. Devido ao fator de se evitar problemas de qualidade de dados geralmente não ser uma opção, a mineração de dados enfoca a detecção e correção de problemas de qualidade de dados e o uso de algoritmos que possam tolerar baixa qualidade de dados.
>
> O primeiro passo, a detecção e correção, é muitas vezes chamado de limpeza dos dados. A próxima seção discute aspectos específicos de qualidade

de dados. O foco são questões de medição e coleta de dados, embora alguns problemas relacionados a aplicações também sejam discutidos. [40].

3.8 Trabalhando os dados

Os dados são a principal fonte para a mineração de dados, entretanto estes dados podem ser de diferentes tipos e origens, levando com que ocorram distorções na mineração. Sendo assim para se trabalhar com os dados são necessários algumas etapas para o preparo dos mesmos para o processo de mineração a primeira etapa é a de pré-processamento.

> A etapa de pré-processamento, no processo de descoberta de conhecimento – *KDD* (*Knowledge Discovery in databases*) compreende a aplicação de várias técnicas para captação, organização, tratamento e a preparação dos dados. É uma etapa que possui fundamental relevância no processo de KDD. [17].

Compreende desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de mineração de dados que serão utilizados. [17].

Segundo [17] algumas das principais funções da etapa de pré-processamento dos dados:

- **Seleção de atributos** – A Seleção de Atributos é uma etapa da fase de pré-processamento do processo de Descoberta de Conhecimento em Banco de Dados. Como o próprio nome já diz, o objetivo é escolher um subconjunto de atributos (também conhecidos como variáveis) ou criar outros atributos que substituam um conjunto deles a fim de reduzir a dimensão do banco de dados. Com essa redução de dimensão, reduz-se a complexidade do banco de dados e assim o tempo de processamento para extrair dele algum conhecimento. Além disso, atributos desnecessários podem causar ruído no resultado final e isto pode ser evitado com a aplicação de técnicas de Seleção de Atributos. Saiba mais.
- **Limpeza dos dados** - abrange qualquer tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade (completude, veracidade e integridade) os fatos por eles representados. Informações ausentes, errôneas ou inconsistentes nas bases de dados devem ser corrigidas de forma a não comprometer a qualidade dos modelos de conhecimento a serem extraídos ao final do processo de KDD.
- **Discretização** - alguns algoritmos de mineração de dados, especialmente os algoritmos de classificação, requerem que os dados estejam na forma de atributos categorizados. Assim, muitas vezes é necessário transformar um atributo contínuo em categórico.

- **Binarização** - Algoritmos utilizados para descoberta de padrões de associação requerem que os dados estejam na forma de atributos binários. Assim, muitas vezes tanto os atributos contínuos quanto os discretos necessitam ser transformados em um ou mais atributos binários.
- **Construção de atributos** - essa operação consiste em gerar novos atributos a partir dos atributos existentes. A importância desse tipo de operação é justificada pois novos atributos, além de expressarem relacionamentos conhecidos entre atributos existentes, podem reduzir o conjunto de dados simplificando o processamento dos algoritmos de Mineração de Dados.
- **Transformação de variáveis** - se refere a uma transformação que seja aplicada a todos os valores de um atributo. Em outras palavras, para cada objeto, a transformação é aplicada ao valor do atributo para aquele objeto. Uma transformação que pode-se citar é a normalização dos dados, que consiste em ajustar a escala dos valores de cada atributo de forma que os valores fiquem em pequenos intervalos, tais como de -1 a 1 ou de 0 a 1. Tal ajuste se faz necessário para evitar que alguns atributos, por apresentarem uma escala de valores maior que outros, influenciem de forma tendenciosa determinados métodos de Mineração de Dados.

CAPÍTULO 4

Mineração de dados

4.1 Descoberta de conhecimento em banco de dados

A mineração de dados é um processo de busca em grandes massas de dados por padrões que possam ser utilizados para a geração de conhecimento, entretanto ele por si só não pode ser definido como responsável pelo processo de geração deste conhecimento. O processo como um todo é conhecido como Descoberta de conhecimento em banco de dados que é referênciada por sua sigla em inglês: *KDD*.

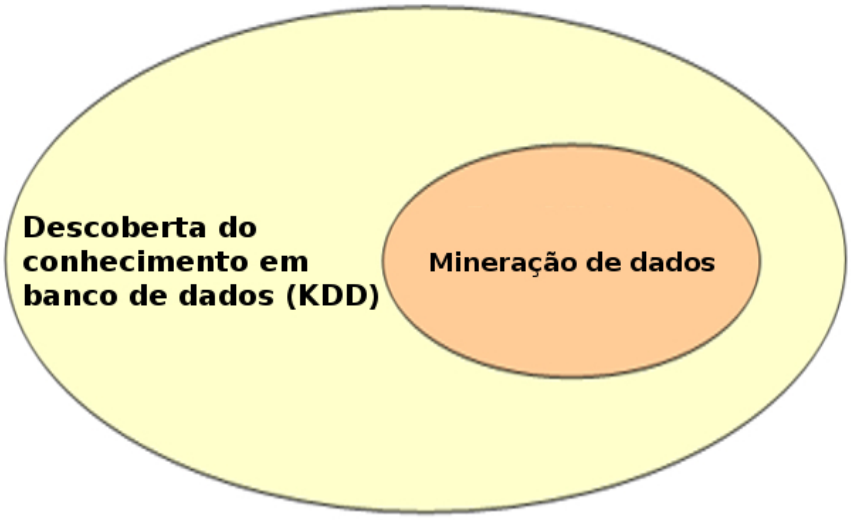

Figura 4.1: *Relação entre KDD e Mineração de dados. fonte: O autor*

A mineração de dados veio da impossibilidade do ser humano de analisar manualmente todas as informações geradas por uma grande massa de dados. Tem por objetivo, obter informações que não são muito óbvias, que não sejam possíveis de se analisar manualmente. Podemos associar a Mineração de Dados a eficiência e a vantagem competitiva. Qualquer instituição que consiga aplicar técnicas de Mineração de Dados irá ter uma grande vantagem competitiva em relação ao seu concorrente. Não envolve

somente corporações na concepção de negócio, mas também grandes redes de pesquisas ligadas a universidades, medicina e outras instituições governamentais.

O processo de descoberta de conhecimento em banco de dados poder ser aplicado a qualquer área que tenha um volume de dados que possa ser explorado, entretanto nem todas as tarefas de descoberta da informação podem ser consideradas mineração de dados, como por exemplo uma simples busca em bancos de dados por mecanismos rotineiros, é nada mais do que tarefas simples de recuperação de dados que são suportadas por sistemas de informação simples.

São alguns exemplos de mineração de dados:

O governo dos EUA se utiliza do mineração de dados há bastante tempo para identificar padrões de transferência de fundos internacionais que se parecem com lavagem de dinheiro do narcotráfico.

Vendas cruzadas podem ser realizadas com facilidade se um banco de dados com informações sobre o passado do cliente existir. Sabendo as necessidades e gostos do cliente, novos produtos podem ser oferecidos pela empresa mantendo a fidelidade do cliente que não precisa ir buscar o produto em outro local.

Devido à competição empresarial, clientes mudam de empresa com facilidade. A Mineração de dados pode ser usado para verificar porque os clientes trocam uma empresa por outra e oferecer serviços, vantagens e ofertas que evitem esta fuga de clientes. É mais fácil manter um cliente do que adquirir um novo. Com o processo de mineração de dados, pode-se localizar que oferta fazer a que cliente para mantê-lo na empresa ou mesmo localizar os clientes que podem sair da empresa sem representar prejuízo.

Na Medicina já é possível a criação e manutenção de grandes bancos de dados com informação sobre sintomas, resultados de exames, diagnósticos, tratamentos e curso das doenças para cada paciente. A mineração destes dados pode fornecer conhecimento novo como, por exemplo, a relação entre algumas doenças e certos perfis profissionais, sócio culturais, hábitos pessoais e local de moradia. Estas relações são utilizadas para melhor entendimento das doenças e seus tratamentos.

Outras áreas que por característica própria de seu estudo, como a Astronomia e a Geologia, geram e acumulam enormes quantidades de dados já estão utilizando intensamente o Mineração de dados para descobrir conhecimento novo que a olho nu não seriam facilmente percebidos.

4.2 Processo de Mineração de dados

A mineração de dados é o processo de descoberta automática de informações úteis em grandes depósitos de dados. As técnicas de mineração de dados são organizadas para agir sobre grandes bancos de dados com o intuito de descobrir padrões úteis

e recentes que poderiam de outra forma permanecer ignorados. Elas também fornecem capacidade de previsão do resultado de uma observação futura. Trata-se de um conjunto de técnicas reunidas da estatística e da inteligência artificial com o objetivo de descobrir conhecimento novo que por ventura esteja escondido em grades massas de dados armazenados em bancos de dados.

Apesar de algumas tarefas de descoberta da informação serem importantes e envolver o uso de algoritmos e estruturas de dados sofisticadas, a explicação é que estas tarefas se baseiam em técnicas tradicionais da tecnologia da informação e em recursos óbvios dos dados para criar estruturas de índice para organizar e recuperar de forma eficiente as informações. Contudo, a mineração de dados tem sido usada para melhorar sistemas de recuperação de informações, e segundo Kumar et al [40], a mineração de dados é uma parte integral da descoberta de conhecimento em bancos de dados, é o processo geral de conversão de dados brutos em informações úteis.

CAPÍTULO 5

Ferramentas de mineração de dados

5.1 *IBM SPSS Modeler*

O *IBM SPSS Modeler* é um *Software* desenvolvido pela IBM para ser utilizando como ferramenta de mineração de dados. Disponível em duas versões – Professional e *Premium* – que se distinguem pelo facto do *IBM SPSS Modeler Premium* incluir capacidades de análise de texto (*Text Analytics*), de análise de redes sociais (*Social Network Analytics*) e de análise de entidades (*Entity Analytics*) que é voltada para banco de dados relacionais.

Utiliza um conjunto alargado de técnicas preditivas para revelar as relações e as tendências presentes nos dados disponíveis em qualquer instituição, independentemente do volume atual e das perspectivas do seu crescimento futuro. A proposta deste *Software* é integrar-se com a infraestrutura informática que possui e dispõe de diversas formas de apresentação dos resultados, garantindo que o conhecimento preditivo está disponível para quem dele necessita e no momento apropriado. Pode mesmo incluir nas suas aplicações informáticas rotinas preditivas desenvolvidas com o *IBM SPSS Modeler*, automatizando processos e otimizando a decisão em toda a sua organização. Por exemplo, um modelo de classificação de clientes pode ser colocado na sua base de dados de clientes de modo a que quando ocorrer a introdução de um novo cliente este seja de imediato classificado consoante as suas caraterísticas.

Pode beneficiar do conhecimento preditivo obtido com o *IBM SPSS Modeler* qualquer que seja o seu setor de atividade. Por exemplo, pode adotar estratégias proativas, em vez de reagir aos acontecimentos, em domínios tão diferentes como:

- Gestão das relações com os seus clientes ou utilizadores dos seus serviços;
- Detecção e prevenção de fraudes;
- Controle de riscos;
- Administração de cuidados de saúde e investigação científica;
- Gestão de programas;
- Segurança pública e segurança interna.

A interface gráfica do *IBM SPSS Modeler* permite aos analistas concentrarem-se mais facilmente no problema a resolver sem ter que perder tempo com tarefas de programação. Á medida que trabalham são mapeados visualmente 'fluxos' interativos de procedimentos de Mineração de dados, o que lhes permite interagir com a informação em qualquer momento e desenvolver os modelos rapidamente e com confiança, permitindo assim que os analistas fiquem livres de tarefas técnicas não produtivas.

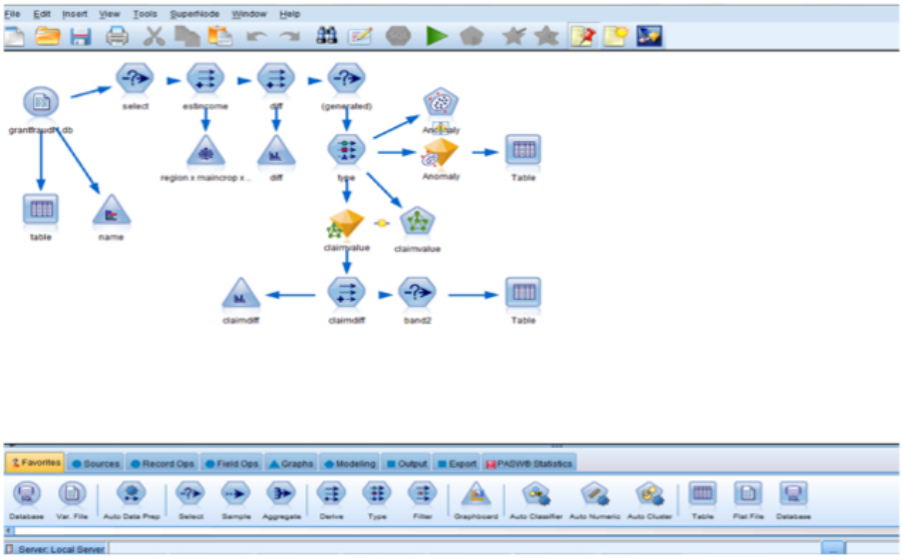

Figura 5.1: *Interface amigável e intuitiva SPSS Modeler. Fonte: [7]*

Graças ás poderosas técnicas automatizadas de modelização, o *IBM SPSS Modeler* pode rapidamente identificar as melhores técnicas analíticas e combinar várias estimativas para aumentar a precisão dos resultados e ao reunir numa única ferramenta todas as funções necessárias aos analistas, aumenta a sua eficiência e produtividade.

5.2 *SAP Sybase IQ*

Sybase IQ é o servidor analítico líder de mercado feito especificamente para negócios de inteligência crítica e análises. Como uma tradição em trabalhos analíticos em fraudes o *Sybase IQ* permite que as organizações seguradoras de todo o mundo detectem de forma mais precisa e previnam fraude possibilitando que conjuntos de dados massivos com grandes dimensões de dados e técnicas estatísticas.

O *Sybase IQ* permite que empresas seguradoras peguem declarações suspeitas cedo no processo de declaração construindo sistemas de detecção de fraude usando sistemas de análises estatísticas em quantidades massivas de dados até 100 vezes mais rápido a uma fração do custo, reduzindo o impacto final da fraude ou aumentando

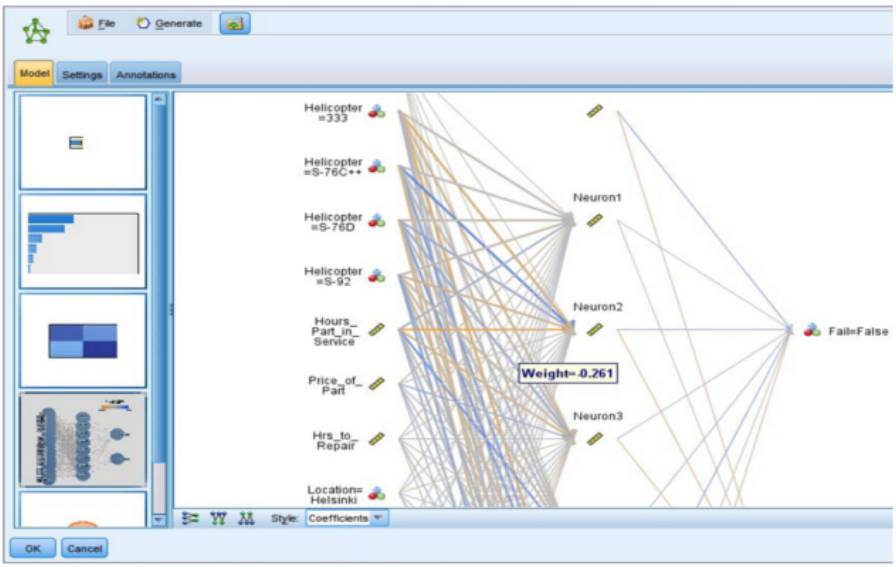

Figura 5.2: *Demonstração de uma rede neural criada através do SPSS Modeler.Fonte: [7]*

a precisão de detecção de e sistemas de prevenção, e aumentando a velocidade de investigação.

O *Sybase IQ* é parte integrante de um ecossistema de tecnologias que suportam a análise de fraude para empresas seguradoras, que inclui as seguintes ferramentas:

5.2.1 *Sybase IQ*

O primeiro banco de dados analítico permitindo armazenamento e recuperação de dados não estruturados como parte do mesmo repositório de transações ou dados analíticos, permite que empresas tenham a mesma compreensão de dados textuais encontrados na mídia social;

5.2.2 O *Sybase Complex Event Processing*

Oferece uma infraestrutura facilmente programável para desenvolvimento e implantação de alertas em tempo real que requerem alto rendimento e baixa latência;

5.2.3 *SAP Business Objects*

Combinado com *Sybase IQ* permite que clientes implantem de forma rápida e mais inovada, gastando menos em suas análises ambientais;

5.2.4 *Sybase Industry Warehouse Studio*

Permite que seguradoras implantem o sistema analítico de trabalho no menos prazo com o mínimo de risco.

5.2.5 Vantagens específicas do produto

- A arquitetura baseada em colunas acelera consultas complexas e reduz a necessidade de armazenamento;
- Projetado para lidar com números grandes de usuários simultâneos;
- Arquitetura aberta com escalabilidade flexível usando hardware comum;
- Funcionalidade de classe empresarial e registro avançado de registros de análises.

Figura 5.3: *Arquitetura SAP Sybase IQ. [38]*

5.2.6 Versões disponíveis do produto

5.2.7 *Sybase IQ Enterprise Edition*

O *Sybase IQ Enterprise Edition* foi projetado para permitir que muitos usuários executem análises e consultas extremamente rápidas, flexíveis e interativas, usando ferramentas de consulta prontas para uso. O *Sybase IQ* pode ser carregado usando vários métodos, incluindo o servidor Sybase *ETL*, a partir de arquivos simples, ou base de dados específicas. Possui uma opção designada como *Multiplex Grid* (*Grade multiplex*) que permite que os mecanismos do *Sybase IQ* sejam executados em vários nós de um cluster de disco compartilhado para carregar ou consultar uma única imagem do banco de dados. A opção *Multiplex Grid* extraordinárias escalabilidade de usuário, escalabilidade

de trabalhos de carregamento e alta disponibilidade. Para muitos aplicativos *DSS* e de *Data Warehouse*, o *Sybase IQ Enterprise Edition* pode aprimorar o tempo de resposta de consulta em até 100 vezes.

5.2.8 *Sybase IQ Small Business Edition*

O *Sybase IQ Small Business Edition* fornece a funcionalidade completa do *Sybase IQ*, exceto o *multiplex* ou quaisquer opções disponíveis para a *Enterprise Edition*. Ele oferece suporte a ambientes de relatório e de análise com até 250 gigabytes de dados armazenamentos, sendo executados em uma máquina com até quatro núcleos, acessado por até 25 conexões de usuário.

5.2.9 *Sybase IQ Single Application Server Edition*

O *Sybase IQ Single Application Server Edition* é um mecanismo analítico altamente otimizado, projetado especificamente para fornecer resultados mais rápidos nas soluções de relatório, análise e *Data Warehouse*. Ele é otimizado para aplicativos com esquemas simples e grandes quantidades de dados em um ou dois ambientes de máquina. Ele foi projetado para permitir que muitos usuários executem análises interativas e consultas *ad-roc* rápidas e flexíveis, usando ferramentas de consulta prontas para uso.

5.3 *Oracle Advanced Analytics*

O *Oracle Advanced Analytics* estende o banco de dados da *Oracle* em uma abrangente plataforma de análise avançada por meio de dois componentes principais: *Oracle R Enterprise* e *Oracle Data Mining*. Com o *Oracle Advanced Analytics*, os clientes têm uma plataforma abrangente para aplicativos de análise em tempo real que fornece informações sobre assuntos comerciais importantes como previsão de insatisfação, recomendação de produtos e alerta de fraude.

O *Oracle R Enterprise* amplia o banco de dados com a biblioteca de linguagens de programação R de funcionalidade estatística', além de levar as computações até o banco de dados, os usuários da ferramenta "R" podem usar habilidades e ferramentas para desenvolvimento "R" existentes, e os *scripts* agora também podem ser executados de maneira transparente e ser dimensionados em relação a dados armazenados em um banco de dados.

O *Oracle Data Mining* fornece algoritmos avançados de Data Mining executados como funções SQL nativas para criação do modelo de banco de dados incorporado e implantação do modelo.

Os usuários do *Oracle Data Miner* podem criar, avaliar, compartilhar e implantar metodologias de analíticos preditivas, ao mesmo tempo em que deixa os algoritmos de Mineração de dados específicos da Oracle acessíveis a R.

5.4 *Weka* (*Waikato Environment for Knowledge Analysis*)

A mineração de dados não é o domínio exclusivo das grandes empresas e do *Software* caro, há um *Software* que possui quase todos os recursos que as ferramentas proprietárias, este *Software* se chama *WEKA*. O *WEKA* é um produto da Universidade de Waikato (Nova Zelândia) e foi implementado pela primeira vez em sua forma moderna em 1997.

Ele usa a *GNU General Public License* (*GPL*). O Software foi escrito na linguagem *Java* e contém uma *GUI* para interagir com arquivos de dados e produzir resultados visuais (pense em tabelas e curvas). Ele também tem uma *API* geral, assim é possível incorporar o *WEKA*, como qualquer outra biblioteca, a seus próprios aplicativos para fazer coisas como tarefas de mineração de dados automatizadas no lado do servidor.[1].

O *WEKA* fornece quatro opções principais:*Explorer*, *Experimenter*, *Knowledge Flow* e o *Simple CLI*, conforme pode se observar na figura a seguir:

Figura 5.4: *Tela principal do Software Weka. Fonte: Screenshot do Software*

Na tela do *Explorer*, existem diversas abas, cada uma com sua finalidade. Quando a tela do *Explorer* é carregada somente a primeira aba está ativa, isso porque é necessário carregar um conjunto de dados antes de começar a explorar os dados. Segue abaixo, a descrição e a figura de cada aba presente na tela do *Explorer*.

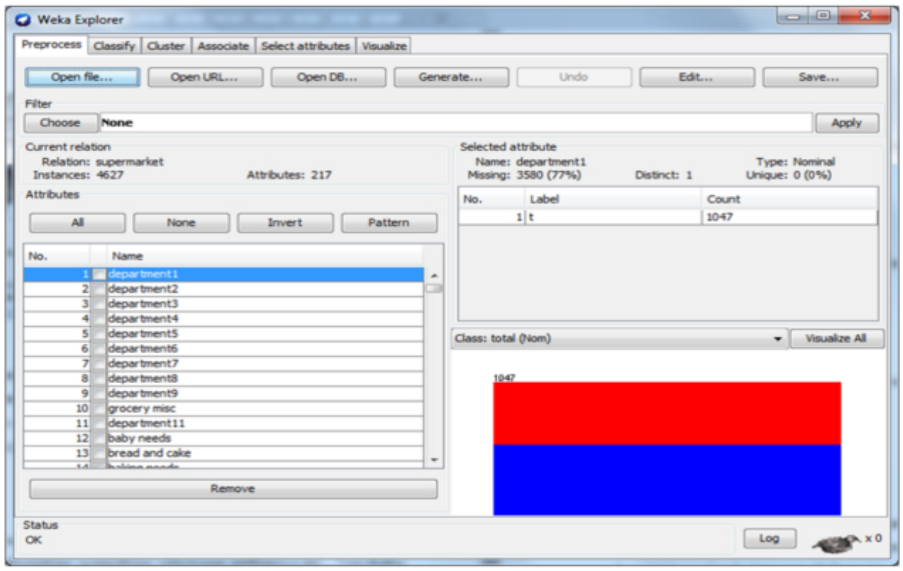

Figura 5.5: *Tela da opção Explorer do software Weka.Fonte: Screenshot do Software*

Abas e funções

- *Preprocess*: Aba onde se pode escolher o conjunto de dados que será usado no experimento e/ou pré-processar (discretizar, normalizar, selecionar atributos e etc...) os dados desse conjunto, é normalmente a primeira aba e funciona como a tela de entrada do modo *Explorer*.
- *Classify*:Aba onde você pode selecionar tanto o tipo de treinamento/teste quanto a técnica de aprendizagem de máquina que será usada no experimento. Nesta tela ao se clicar com o botão auxiliar sobre o resultado da operação é possível gerar gráficos demonstrativos dos mesmos.
- *Cluster*:Aba onde você pode selecionar técnicas de aprendizagem de máquina baseadas em *cluster* (agrupamento).
- *Associate*:Aba onde é possível aprender regras de associação para o conjunto de dados passado.

  Por questões de teste foi utilizada uma base de dados exemplo que tem diversos registros de compras de produtos em um supermercado, após o trabalho de associação do software foram encontrados alguns padrões nas compras realizadas, por associação entre os itens.

  Best rules found:

  1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 conf:(0.92)
  2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 conf:(0.92)
  3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t

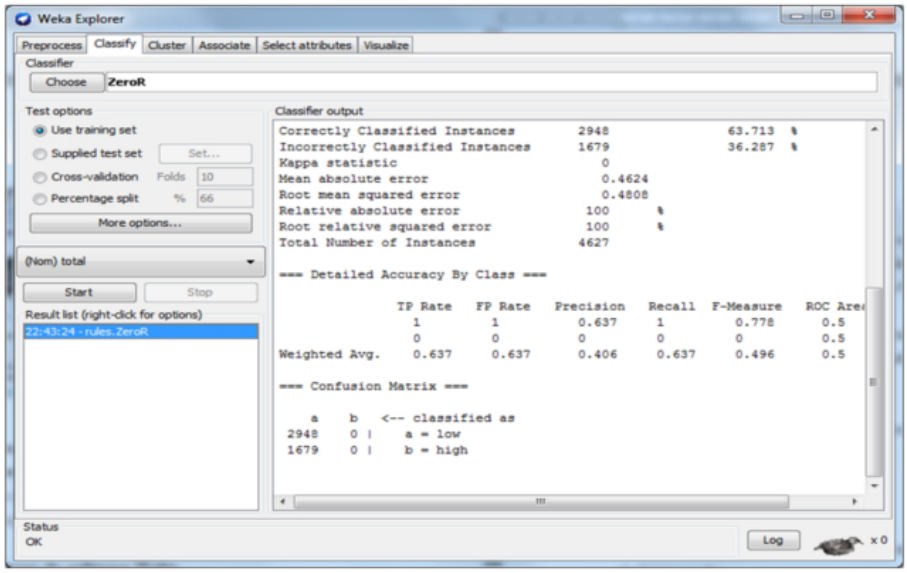

Figura 5.6: *Aba Classify do Weka Explorer. Fonte: Screenshot do Software*

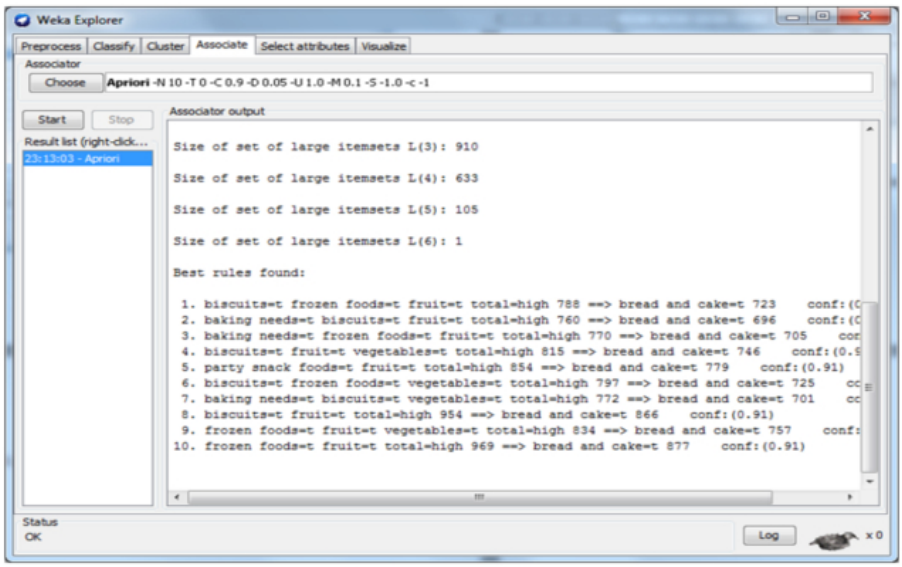

Figura 5.7: *Aba Associate do Weka Explorer. Fonte: Screenshot do Software*

705 conf:(0.92)

4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746 conf:(0.92)
5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779 conf:(0.91)
6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725 conf:(0.91)
7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701 conf:(0.91)

8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866 conf:(0.91)
9. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757 conf:(0.91)
10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877 conf:(0.91)

- *Select Attributes*: Aba onde é possível selecionar os atributos mais relevantes dentro do conjunto de dados passado. Por questões de teste foi utilizada uma base de dados exemplo que tem diversos registros de compras de produtos em um supermercado, após o trabalho de associação do software foram encontrados alguns padrões nas compras realizadas, por associação entre os itens.

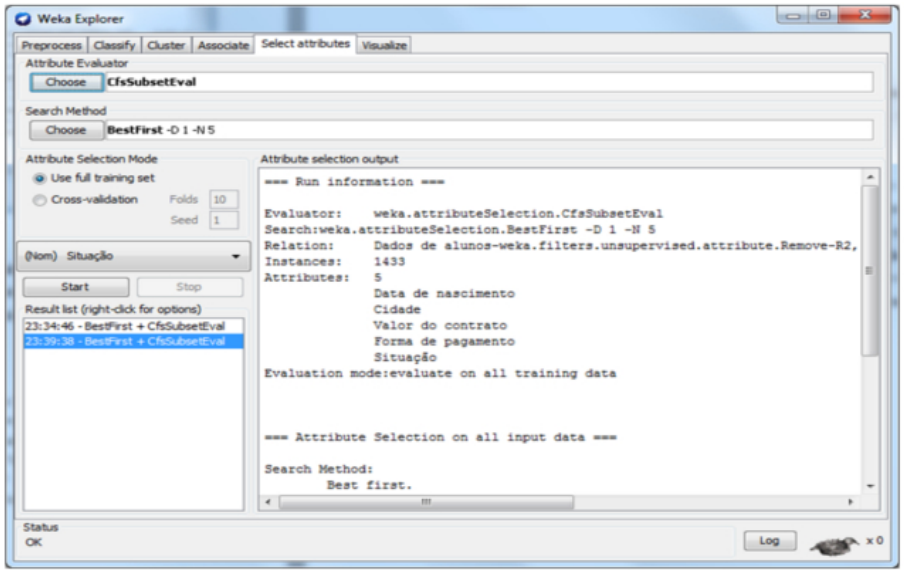

Figura 5.8: *Aba Select atributes do Weka Explorer. Fonte: Screenshot do Software*

- *Visualize*:Aba onde é possível ver, através de gráficos 2D, a dispersão dos dados, normalmente é a aba que demonstra de forma mais clara os padrões/associações encontradas dentro de uma análise de dados.
  Isto é feito através de nós plotados no plano 2D gerado em ladrilho que se for selecionado é exibido em detalhe em uma janela independente.
  As bases de dados podem ser importadas de arquivos em vários outros formatos (arff, csv, c4.5, binário). Além disso, elas podem ser lidas de uma *URL* (*Uniform Resource Locator*) ou site da *internet*. O formato mais comum é o .arff.
  A opção *Simple CLI* permite a realização de experimentos por meio de comandos (*prompt*), a opção Explorer é utilizada para o pré-processamento e aplicação de técnicas de aprendizagem de máquina, a opção *Experimenter* é usada para fins de comparação entre as mais variadas técnicas de inteligência computacional, e finalmente a opção *Knowledge Flow* é uma nova interface gráfica para o *WEKA*, ainda não finalizada, mas que servirá para definir, através de fluxos, o experimento

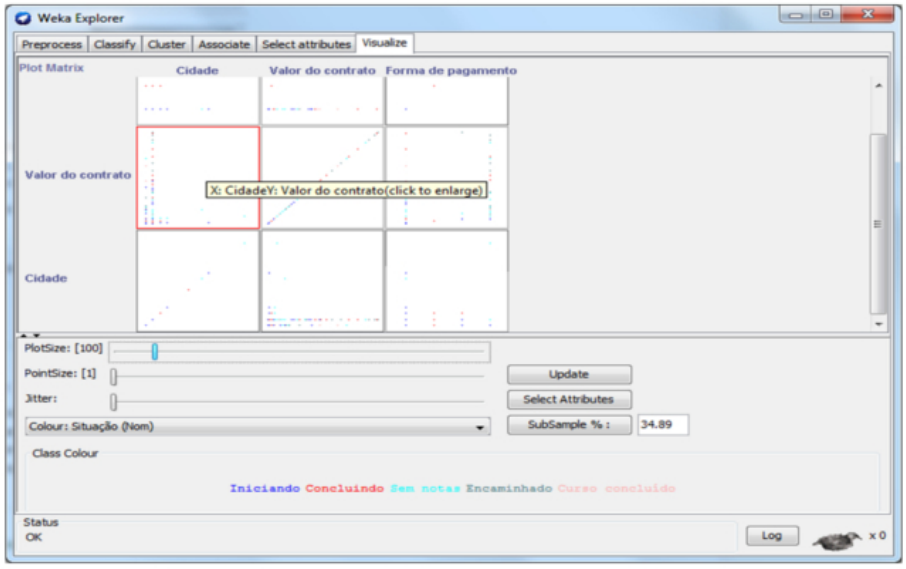

Figura 5.9: *Aba Visualize atributes do Weka Explorer. Fonte: Screenshot do Software*

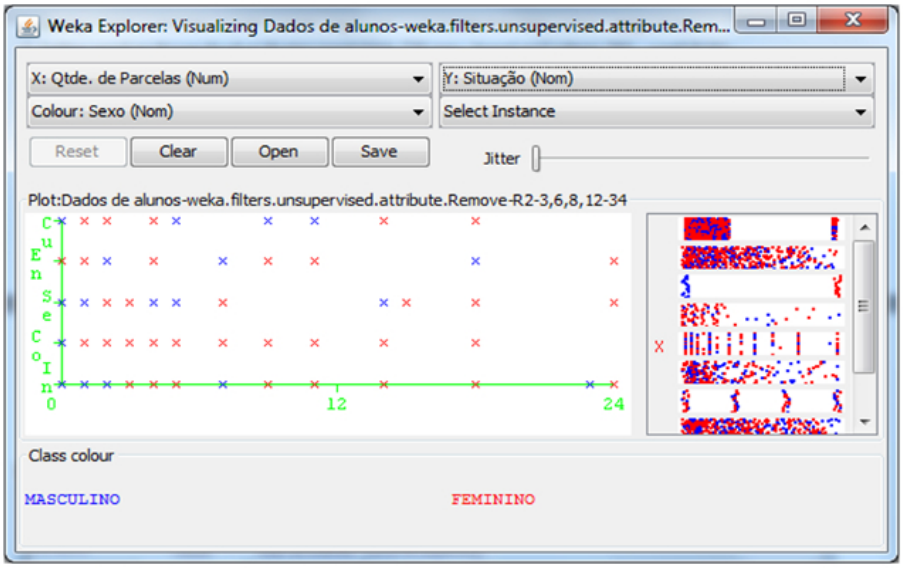

Figura 5.10: *Detalhamento de uma padrão da paleta visualize. Fonte: Screenshot do Software*

a ser executado. Das opções citadas, focarei no *Explorer* por ser a mais usada e a mais importante de ser aprendida inicialmente análise preditiva da DMSS[1]. Segunda maior rede de lojas de departamentos de vestuário.

[1]A DMSS é uma empresa de software que está presente no país desde 1990, oferecendo todas as tecnologias disponíveis para inteligência de negócio e data mining para todo o mercado nacional.

CAPÍTULO 6

Principais algorítimos utilizados na mineração

O centro da mineração de dados envolve o uso de algoritmos avançados, que utilizam de técnicas especiais e são as ferramentas de trabalho que se aplica aos dados já preparados para mineração.

> O algoritmo de mineração de dados é o mecanismo que cria um modelo de mineração de dados. Para criar um modelo, um algoritmo primeiro analisa um conjunto de dados e procura padrões e tendências específicos. O algoritmo usa os resultados dessa análise para definir os parâmetros do modelo de mineração. Esses parâmetros são aplicados pelo conjunto de dados inteiro para extrair padrões acionáveis e estatísticas detalhadas.
>
> O modelo de mineração que um algoritmo cria pode assumir vários formatos, incluindo:
>
> - Um conjunto de regras que descreve como são agrupados produtos em uma transação.
> - Uma árvore de decisão que prevê se um determinado cliente comprará um produto.
> - Um modelo matemático que prevê as vendas.
> - Um conjunto de *clusters* que descreve como os casos em um conjunto de dados estão relacionados.
>
> [27].

6.1 Aplicando os Algoritmos

No contexto de definição e aplicação de algoritmos percebe-se que:

> A definição e uso de um algoritmo para a mineração de dados não é uma tarefa fácil e exige conhecimento avançado das técnicas neles empregadas e dos resultados obtidos através de cada algoritmo, conforme cita a Microsoft.

> A escolha do melhor algoritmo para uma tarefa empresarial específica pode ser um desafio. Embora você possa usar algoritmos diferentes para executar a mesma tarefa empresarial, cada algoritmo produz um resultado diferente e alguns podem produzir mais de um tipo de resultado. Por exemplo, você pode usar o algoritmo Árvores de Decisão da Microsoft não apenas para previsão, mas também como uma maneira de reduzir o número de colunas em um conjunto dados uma vez que a árvore de decisão pode identificar colunas que não afetam o modelo de mineração final.
>
> Você também não tem que usar algoritmos independentemente. Em uma solução de mineração de dados, é possível usar alguns algoritmos para explorar dados e, em seguida, usar outros algoritmos para prever um resultado específico com base nesses dados. Por exemplo, você pode usar um algoritmo de *cluster*, que reconhece padrões, para dividir dados em grupos que são mais ou menos homogêneos e, em seguida, usar os resultados para criar um modelo de árvore de decisão melhor. Você pode usar vários algoritmos em uma solução para executar tarefas separadas, por exemplo, usando um algoritmo de árvore de regressão para obter informações de previsão financeira e um algoritmo com base em regras para executar uma melhor análise de cesta básica. [27].

Sendo assim evidencia-se que o uso de algoritmos de mineração não é um processo fixo ou que possa ser aplicado sem critério conforme afirma Bogorny (2003) A escolha de um método depende do contexto e do domínio da aplicação, bem como, do tipo de conhecimento que se deseja encontrar.

Os modelos de mineração podem prever valores, produzir resumos de dados e localizar correlações ocultas. Para uma melhor definição a tabela 6.1 fornece sugestões para quais algoritmos devem ser usados para tarefas específicas.

6.2 Algoritmo de Árvore de decisão

O algoritmo de árvore de decisão é um algoritmo de classificação e regressão para uso em modelagens de previsão de atributos discretos e contínuos. No caso dos atributos discretos[1], o algoritmo faz previsões fundadas nas relações entre colunas de entrada em um conjunto de dados.

Ele usa os valores, conhecidos como estados, dessas colunas para prever os estados de uma coluna que você define como previsível. Especificamente, o algoritmo

[1] Um atributo discreto significa que a coluna contém um número finito de valores sem continuidade entre eles, códigos de área de telefone são um bom exemplo de dados numéricos discretos [27].

Tabela 6.1: *Sugestões de algoritmos para mineração por tarefa. Fonte: [27]*

Tarefa	**Indicação de algoritmos a serem usados**
Prevendo um atributo discreto: Por exemplo, prever se o destinatário da campanha de mala-direta comprará um produto.	Algoritmo Árvores de Decisão Algoritmo de *Clustering* Algoritmo de Rede Neural
Prevendo um atributo contínuo: Por exemplo, prever as vendas do próximo ano.	Algoritmo Árvores de Decisão
Localizando grupos de itens comuns em transações: Por exemplo, usar a análise de cesta básica para sugerir produtos adicionais a um cliente para compra.	Algoritmo de Associação Algoritmo Árvores de Decisão
Localizando grupos de itens semelhantes. Por exemplo, segmentar dados demográficos em grupos para entender melhor as relações entre atributos.	Algoritmo de *Clustering*

identifica as colunas de entrada que são correlacionadas com a coluna previsível. Por exemplo, em um cenário em que se deseja prever a tendência dos clientes em adquirir uma bicicleta, se 9 de 10 clientes jovens comprarem uma bicicleta, mas apenas 2 de 10 clientes mais velhos fizerem o mesmo, o algoritmo infere que idade é um bom indicador para a compra de bicicletas. A árvore de decisão faz previsões com base nesta tendência para obter um resultado específico.

No caso de atributos contínuos, o algoritmo usa a regressão linear para determinar onde uma árvore de decisão se divide. Se mais de uma coluna for definida como previsível, ou se os dados de entrada tiverem uma tabela aninhada configurada como previsível, o algoritmo criará uma árvore de decisão separada para cada coluna previsível.

6.2.1 Como o algoritmo funciona

> O algoritmo de árvore de Decisão gera um modelo de mineração de dados criando uma série de divisões na árvore. Essas divisões são representadas como nós. O algoritmo adiciona um nó ao modelo toda vez que uma coluna de entrada é considerada significativamente correlacionada a uma coluna previsível. [27].

O algoritmo de árvores de Decisão usa a seleção de recurso para guiar a seleção dos atributos mais úteis. A seleção de recurso é importante para impedir que atributos sem importância usem tempo do processador.

Um problema muito comum nos modelos de mineração de dados é que eles se tornam muito sensíveis a diferenças pequenas nos dados de treinamento. Nesse caso, refere-se a eles como sobrecarregados ou muito treinados.

6.2.2 Prevendo colunas discretas

A forma como o algoritmo de Árvore de Decisão cria uma árvore para uma coluna previsível discreta que pode ser mostrada usando uma imagem. A imagem a seguir mostra um histograma que esboça uma coluna previsível.

Compradores de bicicleta, em comparação com uma coluna de entrada, Idade.

A Figura 6.1 mostra que a idade de uma pessoa ajuda a distinguir se ela comprará uma bicicleta.

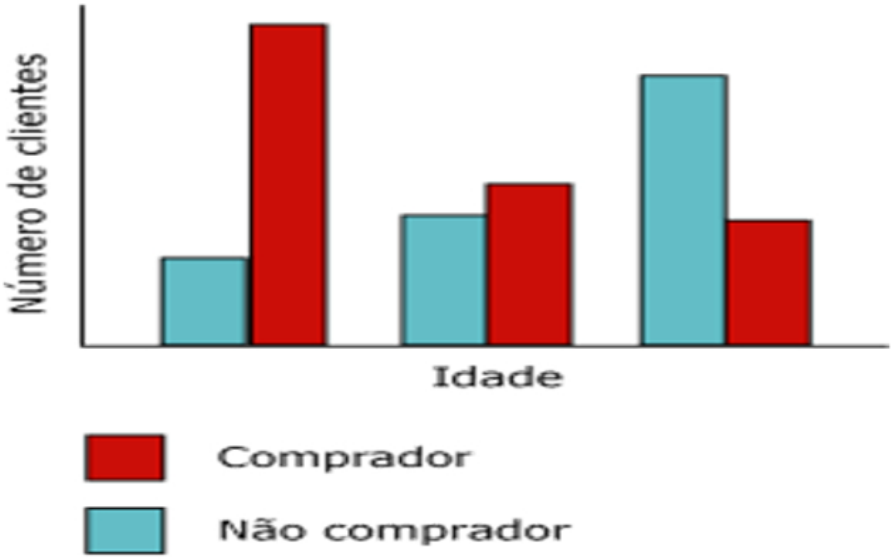

Figura 6.1: *Demonstração de tendência de compra por idade. Fonte: [27]*

A correlação que é mostrada no diagrama faz com que o algoritmo Árvore de Decisão crie um novo nó no modelo, conforme demonstrado na Figura 6.2.

À medida que o algoritmo acrescenta novos nós em um modelo, uma estrutura de árvore é formada, com o nó superior da árvore indicando a divisão da coluna previsível para a média da população de clientes. Como o modelo continua crescendo, o algoritmo considera todas as colunas.

6.2.3 Prevendo colunas contínuas

Quando o algoritmo de Árvore de Decisão cria uma árvore com base em uma coluna previsível contínua, cada nó contém uma fórmula de regressão e uma divisão ocorre em um ponto de não-linearidade na fórmula de regressão.

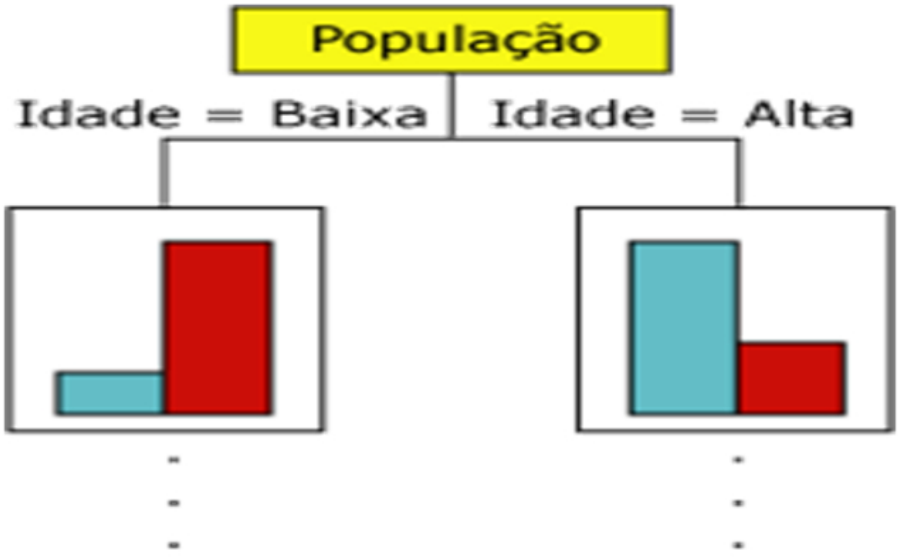

Figura 6.2: *Nó de decisão baseado na tendência por idade. Fonte: [27]*

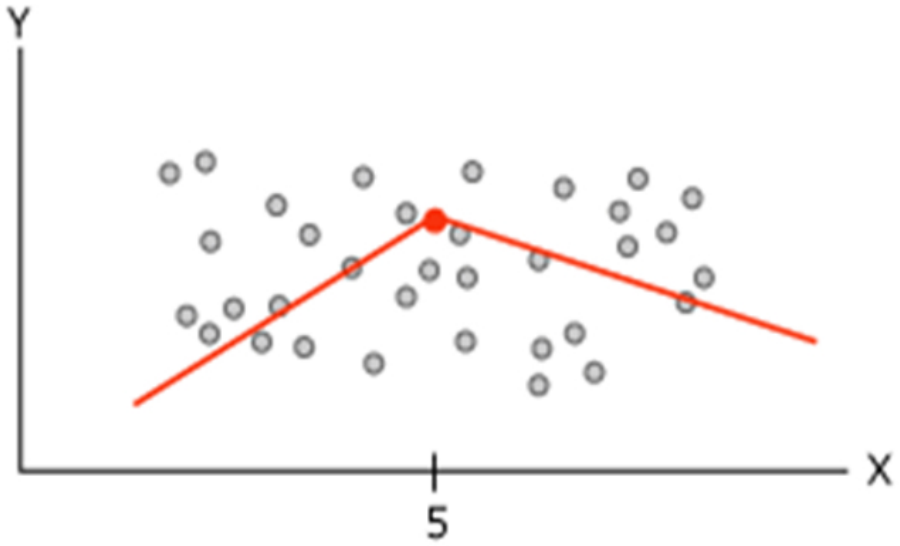

Figura 6.3: *Criação de nós a partir de uma fórmula de regressão. Fonte: [27]*

O diagrama contém dados que podem ser modelados usando uma única linha ou usando duas linhas conectadas. Porém, uma única linha não representaria os dados de forma satisfatória. Mas, se você usar duas linhas, o modelo terá um desempenho muito melhor ao aproximar dados. O ponto onde duas linhas se encontram é o ponto de não-linearidade e é onde o nó de um modelo de árvore de decisão se dividiria. Por exemplo, o nó que corresponde ao ponto de não-linearidade no gráfico anterior poderia ser representado pelo diagrama a seguir. As duas equações representam as equações de regressão para as duas linhas.

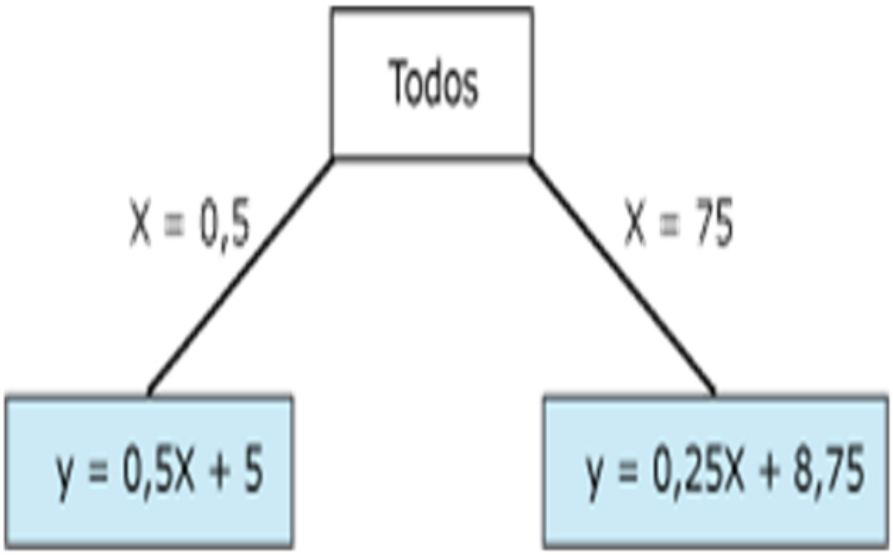

Figura 6.4: *Fórmula de regressão da árvore de decisão. Fonte: [27]*

6.2.4 Dados necessários para modelos de árvore de decisão

Ao preparar dados para usar em um modelo de árvore de decisão, deve-se saber os requisitos do algoritmo específico, incluindo a quantidade de dados necessária e como eles são usados. Os requisitos para um modelo de árvore de decisão são os seguintes:

- Uma única *key* coluna: Cada modelo deve conter uma coluna numérica ou de texto que identifique unicamente cada registro. Não são permitidas chaves compostas;
- Uma coluna previsível:Requer, pelo menos, uma coluna previsível. Você pode incluir vários atributos previsíveis em um modelo, e o atributo previsível pode ser de diferentes tipos, tanto numérico como discreto. Porém, o aumento no número de atributos previsíveis pode aumentar o tempo de processamento;
- Colunas de entrada: Requer colunas de entrada que podem ser discretas ou contínuas. O aumento no número de atributos de entrada afeta o tempo de processamento.

6.3 Algoritmo de *Clustering*

Este algoritmo usa técnicas iterativas para agrupar casos em um conjunto de dados em *clusters* que contenham características semelhantes. Esses agrupamentos são úteis para explorar dados, identificando anomalias nos dados e criar previsões.

Modelos de *clustering* identificam as relações em um conjunto de dados que não podem ser derivados de forma lógica através de observação casual. Por exemplo, você pode discernir logicamente que pessoas que se vão para o trabalho de bicicleta

normalmente não moram longe do local onde trabalham. Porém, o algoritmo pode encontrar outras características dos usuários de bicicleta que não são tão óbvias. Na figura a seguir, o *cluster* "A" representa dados sobre pessoas que pretendem ir de carro para o trabalho, enquanto o *cluster* "B" representa dados sobre pessoas que pretendem ir de bicicleta para o trabalho.

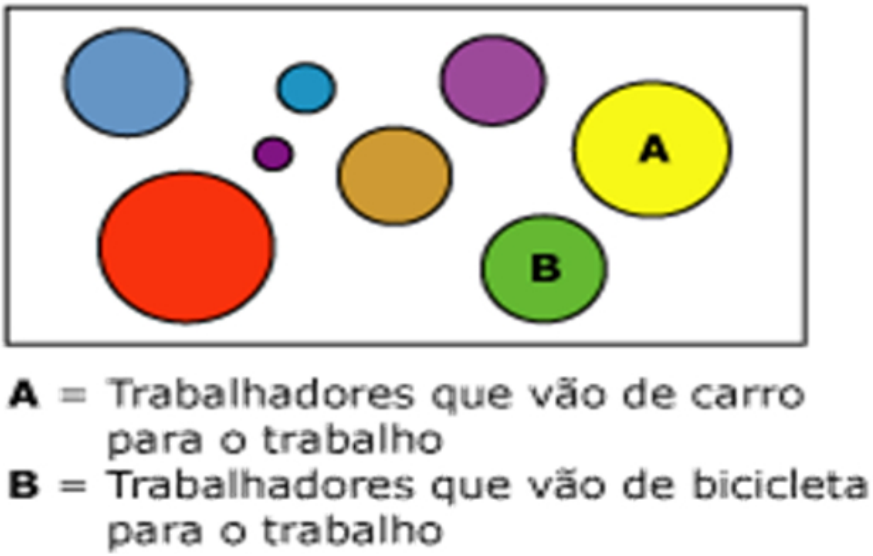

Figura 6.5: *Demonstração de agrupamento por similaridades. Fonte: [27]*

O algoritmo de *clustering* difere dos demais algoritmos de mineração de dados, como o algoritmo de Árvore de Decisão, porque você não precisa designar uma coluna previsível para poder criar um modelo de *clustering*. O algoritmo de *clustering* treina o modelo estritamente a partir das relações existentes nos dados e a partir dos clusters que o algoritmo identifica.

6.3.1 Como o algoritmo funciona

O algoritmo *Clustering* primeiro identifica as relações em um conjunto de dados e gera uma série de *clusters* com base nelas. Uma dispersão é uma maneira útil para representar visualmente como o algoritmo agrupa os dados, conforme mostrado no diagrama a seguir. A dispersão representa todos os caso no conjunto de dados, e cada caso é um ponto no gráfico. Os *clusters* agrupam pontos no gráfico e ilustram as relações que o algoritmo identifica.[27]

Nos algoritmos de *clustering*, que utilizam uma abordagem *bottom-up*, cada elemento do conjunto é, inicialmente, associado a um *cluster* distinto, e novos *clusters* vão sendo formados pela união dos *clusters* existentes. Esta

união ocorre de acordo com alguma medida que forneça a informação sobre quais deles estão mais próximos uns dos outros.

Nos algoritmos de divisão, com uma abordagem *top-down*, inicialmente tem-se um único *cluster* contendo todos os elementos do conjunto e, a cada passo, são efetuadas divisões, formando novos *clusters* de tamanhos menores, conforme critérios pré-estabelecidos.[22]

O seja o mesmo algoritmo que é utilizado para fazer o agrupamento pode ser utilizando para fazer a divisão dos elementos conforme pode ser constatado na imagem a seguir.

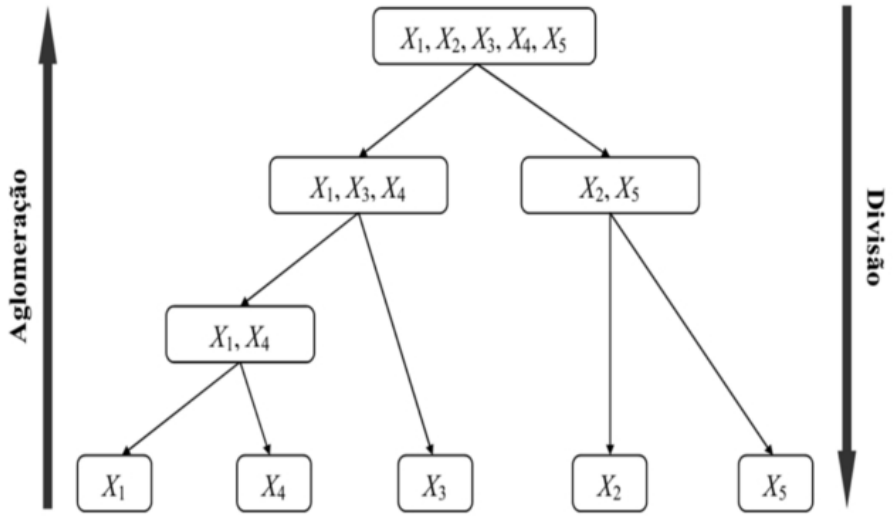

Figura 6.6: *Sentidos do algoritmo para o agrupamento ou divisão dos elementos. Fonte: [22]*

Após definir primeiro os *clusters*, o algoritmo calcula como os mesmos representam satisfatoriamente agrupamentos dos pontos, e em seguida, tenta redefinir os agrupamentos para criar os *clusters* que melhor representem os dados. O algoritmo itera através desse processo até não poder mais melhorar os resultados pela redefinição dos *clusters*.

É possível personalizar o modo como o algoritmo funciona ao selecionar uma técnica de *clustering* específica, limitando o número máximo de *clusters* ou alterando o valor de suporte necessário para criar um *cluster*. Para obter mais informações, consulte Referência técnica do algoritmo *Clustering*.

6.3.2 Dados necessários para modelos de *clustering*

Segundo [27], Ao preparar dados para usar no treinamento de um modelo de *clustering*, é preciso entender os requisitos de um determinado algoritmo, incluindo a quantidade de dados necessária e como eles são usados.

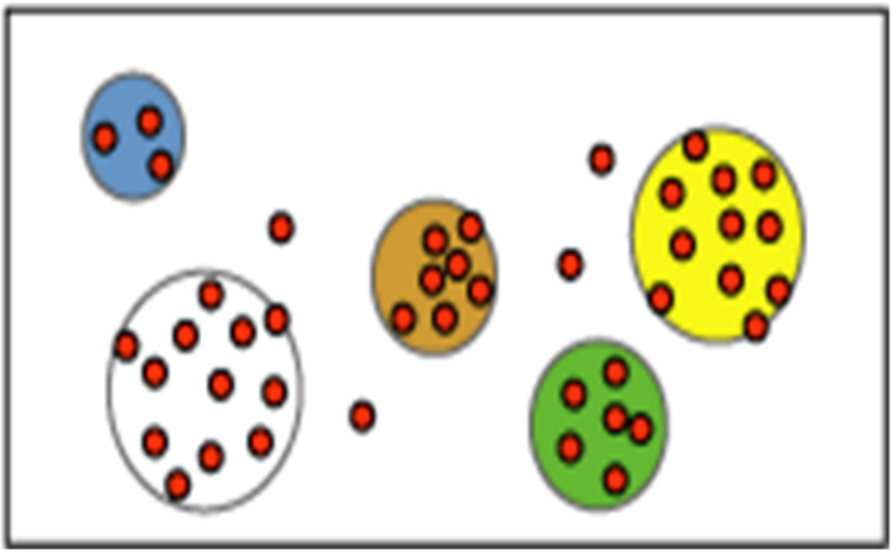

Figura 6.7: *Dispersão de casos em um conjunto de dados. Fonte: [27]*

Os requisitos de um modelo de *clustering* são os seguintes:

- Uma única *key* coluna: Cada modelo deve conter um numérico ou coluna de texto que exclusivamente identifica cada registro, chaves compostas não são permitidas;
- Colunas de entrada: Cada modelo deve conter pelo menos uma coluna de entrada contendo os valores que serão usados para criar os *clusters*. Você pode ter quantas colunas de entrada desejar, mas, dependendo do número de valores em cada coluna, a inclusão de colunas extras pode aumentar o tempo que leva para treinar o modelo;
- Coluna previsível opcional: O algoritmo não precisa de uma coluna previsível para criar o modelo, mas você pode adicionar uma coluna previsível de praticamente qualquer tipo de dados, os valores da coluna previsível podem ser tratados como entrada para o modelo de *clustering* ou você pode especificar que ela deve ser usada somente para previsão. Por exemplo, para prever a receita de clientes ao agrupá-los por dados demográficos, como região ou idade, você poderia especificar a receita como coluna previsível e adicionar todas as outras colunas, como região e idade, como entradas.

6.4 Algoritmo de Rede Neural

Redes neurais é um conceito da computação que visa trabalhar no processamento de dados de maneira semelhante ao cérebro humano. O cérebro é tido como um processador altamente complexo e que realiza processamentos de maneira paralela para isso, ele organiza sua estrutura, ou seja, os neurônios, de forma que eles realizem o processamento necessário. Isso é feito

> numa velocidade extremamente alta e não existe qualquer computador no mundo capaz de realizar o que o cérebro humano faz.
>
> Nas redes neurais artificiais, a ideia é realizar o processamento de informações tendo como princípio a organização de neurônios do cérebro. Como o cérebro humano é capaz de aprender e tomar decisões baseadas na aprendizagem, as redes neurais artificiais devem fazer o mesmo. Assim, uma rede neural pode ser interpretada como um esquema de processamento capaz de armazenar conhecimento baseado em aprendizagem (experiência) e disponibilizar este conhecimento para a aplicação em questão. Um modelo de mineração desenvolvido com o algoritmo de Rede Neural pode conter várias redes, dependendo do número de colunas usadas para a previsão de entrada ou usadas apenas para previsão. O número de redes que um modelo de mineração simples contém depende do número de estados que estão contidos nas colunas de entrada e as colunas previsíveis que o modelo de mineração usa.
>
> O algoritmo de Rede Neural é útil para analisar dados de entrada complexos, tais como de um processo de fabricação, de comercialização ou, problemas comerciais para os quais uma quantidade significativa de dados de treinamento está disponível mas, para os quais regras não podem ser facilmente derivadas usando outros algoritmos. [27].

Segundo a [27], a aplicação deste tipo de algoritmo pode ser feita nas seguintes áreas:

- Análise de promoção e marketing, tais como, medição do sucesso de uma promoção de mala direta ou de uma campanha publicitária radiofônica;
- Prevendo movimento de ações, flutuação de moeda ou demais informações financeiras altamente fluidas a partir de dados históricos;
- Analisando processos industriais e de fabricação;
- Mineração de texto;
- Qualquer modelo de previsão que analisa relações complexas entre muitas entradas e, relativamente, menos saídas.

6.4.1 O aprendizado

O processo de aprendizagem das redes neurais é realizado quando ocorrem várias modificações significantes nas sinapses dos neurônios. Essas mudanças ocorrem de acordo com a ativação dos neurônios. Se determinadas conexões são mais usadas, estas são reforçadas enquanto que as demais são enfraquecidas. É por isso que quando uma rede neural artificial é implantada para uma determinada aplicação, é necessário um tempo para que esta seja treinada.

Existem, basicamente, três tipos de aprendizado nas redes neurais artificiais:

- Supervisionado: neste tipo, a rede neural recebe um conjunto de entradas padronizadas e seus correspondentes padrões de saída, onde ocorrem ajustes nos pesos sinápticos até que o erro entre os padrões de saída gerados pela rede tenha um valor desejado;
- Não-supervisionado: neste tipo, a rede neural trabalha os dados de forma a determinar algumas propriedades dos conjunto de dados. A partir destas propriedades é que o aprendizado é constituído;
- Híbrido: neste tipo ocorre uma "mistura"dos tipos supervisionado e não-supervisionado. Assim, uma camada pode trabalhar com um tipo enquanto outra camada trabalha com o outro tipo.

6.4.2 Como o algoritmo funciona

As redes neurais artificiais são criadas a partir de algoritmos projetados para uma determinada finalidade. É impossível criar um algoritmo desse sem ter conhecimento de modelos matemáticos que simulem o processo de aprendizado do cérebro humano. Por este ser um artigo de introdução a este assunto, abordaremos uma explicação conceitual eliminando ao máximo os princípios matemáticos naturalmente relacionados.

Basicamente, uma rede neural se assemelha ao cérebro em dois pontos: o conhecimento é obtido através de etapas de aprendizagem e pesos sinápticos são usados para armazenar o conhecimento. Uma sinapse é o nome dado à conexão existente entre neurônios. Nas conexões são atribuídos valores, que são chamados de pesos sinápticos. Isso deixa claro que as redes neurais artificiais têm em sua constituição uma série de neurônios artificiais (ou virtuais) que serão conectados entre si, formando uma rede de elementos de processamento.

Tendo uma rede neural montada, uma série de valores podem ser aplicados sobre um neurônio, sendo que este está conectado a outros pela rede. Estes valores (ou entradas) são multiplicados no neurônio pelo valor do peso de sua sinapse. Então, esses valores são somados. Se esta soma ultrapassar um valor limite estabelecido, um sinal é propagado pela saída (axônio) deste neurônio. Em seguida, essa mesma etapa se realiza com os demais neurônios da rede. Isso quer dizer que os neurônios vão enfrentar algum tipo de ativação, dependendo das entradas e dos pesos sinápticos. O algoritmo de Rede Neural cria uma rede que é composta por até três camadas de neurônios. Essas camadas são uma camada de entrada, uma camada opcional oculta e uma camada de saída.

Tabela 6.2: *Camadas de uma rede neural. Fonte: [27]*

Camada	Padrão de funcionamento
Entrada	Os neurônios de entrada definem todos os valores de atributo de entrada do modelo de mineração de dados e suas probabilidades.
Oculta	Os neurônios ocultos recebem entradas de neurônios de entrada e fornecem resultados para os neurônios de saída. A camada oculta é onde as várias probabilidades de entradas são ponderadas. Uma ponderação descreve a relevância ou importância de uma entrada específica para o neurônio oculto. Quanto maior a ponderação atribuída a uma entrada, mais importante será o valor daquela entrada. As ponderações podem ser negativas, o que significa que a entrada pode inibir, em vez de favorecer, um resultado específico.
Saída	Os neurônios de saída representam valores de atributos previsíveis para o modelo de mineração de dados.

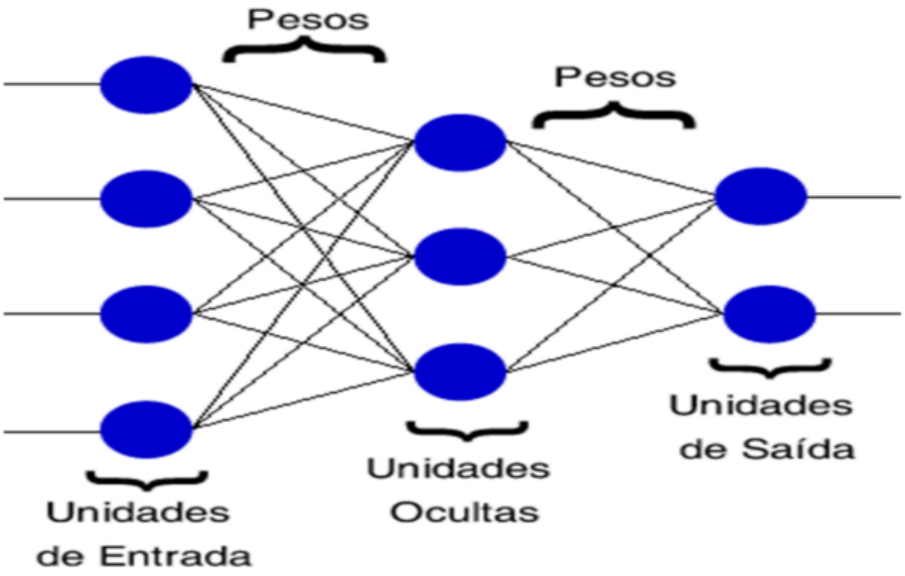

Figura 6.8: *Representação de uma rede neural.Fonte: Laboratório Nacional de Computação Científica (Disponível em: http://www.lncc.br/ labinfo/tutorialRN/)*

6.4.3 Dados necessários para modelos de rede neural

Um modelo de rede neural deve conter uma coluna-chave, uma ou mais colunas de entrada e uma ou mais colunas previsíveis.

Os modelos de mineração de dados que usam o algoritmo Rede Neural da Microsoft são amplamente influenciados pelos valores especificados para os parâmetros disponíveis para o algoritmo. Esses parâmetros definem como os dados são amostrados, são distribuídos ou estimados para serem distribuídos em cada coluna e quando a seleção de recurso é chamada para limitar os valores usados no modelo final.

6.5 Algoritmo de Regressão Linear

O algoritmo de Regressão Linear é uma variação do algoritmo de Árvores de Decisão que o ajuda a calcular uma relação linear entre uma variável dependente e uma independente e, depois, a usar aquela relação para previsão. A relação assume a forma de uma equação para uma linha que melhor represente uma série de dados. Por exemplo, a linha na figura a seguir é a melhor representação linear possível dos dados. [27].

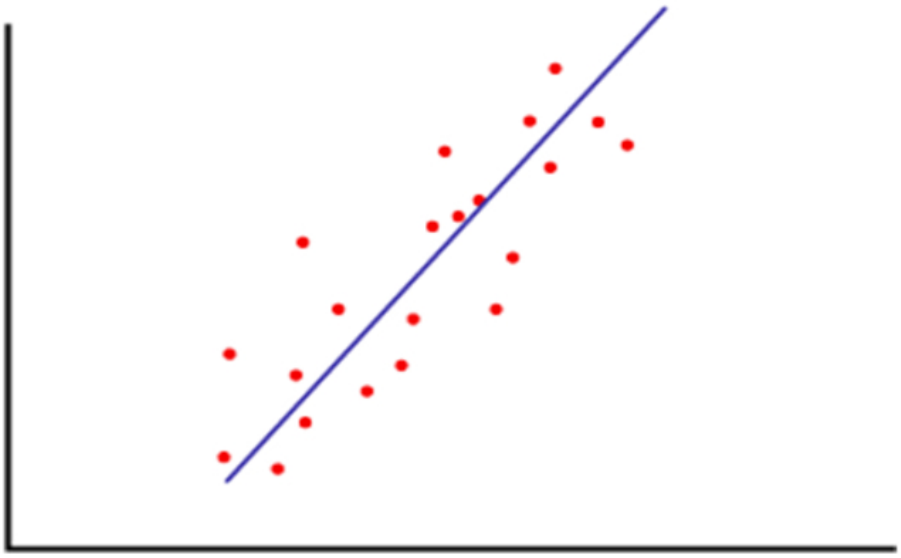

Figura 6.9: *Representação de uma série de dados linear. Fonte: [27]*

Cada ponto de dados no diagrama tem um erro associado à sua distância da linha de regressão. Os coeficientes a e b na equação de regressão ajustam o ângulo e o local da linha de regressão. É possível obter a equação de regressão ajustando a e b até que a soma dos erros associados a todos os pontos atinja o menor número.

Há outros tipos de regressão que usam diversas variáveis e também métodos não lineares de regressão. Porém, uma regressão linear é um método útil e conhecido para modelar uma resposta a uma alteração em alguns fatores subjacentes.

Exemplo:

É possível usar a regressão linear para determinar uma relação entre duas colunas contínuas. Por exemplo, você pode usar a regressão linear para computar uma linha de tendência de dados de fabricação ou de vendas. É possível também usar a regressão linear como um precursor para o desenvolvimento de modelos de mineração de dados mais complexos para avaliar relações entre colunas de dados.

Apesar de haver muitas maneiras para se computar regressão linear sem a necessidade de ferramentas de mineração de dados, a vantagem de usar o algoritmo de Regressão Linear para esta tarefa é que todas as relações possíveis entre as variáveis são automaticamente computadas e testadas, não sendo necessário selecionar um método de computação, como resolver para mínimos quadrados. Porém, a regressão linear pode simplificar muito as relações em cenários onde diversos fatores afetam o resultado.

6.5.1 Como o algoritmo funciona

O algoritmo de Regressão Linear é uma variação do algoritmo de Árvores de Decisão. Além disso, em um modelo de regressão linear, todo o conjunto de dados é usado para computar relações na passagem inicial, enquanto que um modelo de árvores de decisão divide os dados repetidamente em subconjuntos ou árvores menores.

6.5.2 Dados requeridos para modelos de regressão linear

Ao preparar dados para usar em um modelo de regressão linear, você deve entender os requisitos para o algoritmo específico. Isso inclui a quantidade de dados necessária e a forma como os dados são usados. Os requisitos para este tipo de modelos são os seguintes:

- Uma única *key*: coluna Cada modelo deve conter uma coluna numérica ou de texto que identifique unicamente cada registro. Não são permitidas chaves compostas;
- Uma coluna previsível: Requer, pelo menos, uma coluna previsível. Você pode incluir diversos atributos previsíveis em um modelo, mas eles devem ser tipos de dados numéricos contínuos. Não é possível usar um tipo de dados *datetime* como um atributo previsível, mesmo que o armazenamento nativo dos dados seja numérico;
- Colunas de entrada: Colunas de entrada devem conter dados numéricos contínuos e devem ser atribuídas ao tipo de dados apropriado.

CAPÍTULO 7

Conclusão

A Mineração de dados é uma das áreas do conhecimento mais deslumbrante no contexto da Tecnologia da Informação e Comunicação. A principal contribuição deste estudo foi a identificação de ferramentas que permitam a descoberta de conhecimento e a rápida análise em um grande volume de dados, que cresce a cada dia nas organizações ou até mesmo na Internet.

Utilizando ferramentas desenvolvidas com esta finalidade e complexidade específica, junto com conhecimento do funcionamento dos principais algoritmos de Mineração de dados, acredita-se que é possível analisar e gerar conhecimento que até então estavam incobertos, junto à milhares ou milhões de registros em bancos de dados corporativos, mostrando assim novas formas de atuação a partir da descoberta feita.

Com a crescente massa de dados, torna-se evidente o uso de tecnologias mais avançadas para o fornecimento de informações para apoiarem no processo de tomada de decisão. Neste contexto, o uso de técnicas e algoritmos relacionados a Mineração de dados se fazem necessários para viabilizar a geração de conhecimentos e, dessa forma, fornecer subsídios para os gestores maximizarem seus investimentos e alocarem da melhor forma seus recursos observando o custo-benefício. Para isto, percebe-se que existem um conjunto de técnicas, ferramentas e práticas utilizadas em Inteligência de negócio e Mineração de dados, cada qual com seus conceitos, boas práticas, sistemas de computador e os algoritmos mais utilizados para estas finalidades. Assim, destaca-se a importância das organizações tomarem conhecimento da sua realidade para que tenham sucesso e alcance os resultados almejados.

Referências Bibliográficas

[1] ABERNETHY, M. **Mineração de dados com weka**. IBM. Acesso em: 15 de dezembro de 2013.

[2] ALECRIM, E. **Redes neurais artificiais**. Infowester. Acesso em: 21 de dezembro de 2013.

[3] ALMEIDA, E. C. D. **Estudo de viabilidade de uma plataforma de baixo custo para data warehouse**. arxiv.org. Acesso em: 14 de dezembro de 2013.

[4] BERNERS-LEE, R.; SWICK, T. **Semantic web development**. www.w3.org. Acesso em: Agosto de 2013.

[5] BOGORNY, V. **Algoritmos e Ferramentas para Descoberta de Conhecimento em Bases de Dados Geográficos**. Biblioteca UFRGS, Porto Alegre, 2003.

[6] CONSULTORIA, C. **O que é data mart?** Cetax. Acesso em: 19 de dezembro de 2013.

[7] CORPORATION, I. **Ibm spss modeler**. www.pse.pt. Acesso em: 14 de dezembro de 2013.

[8] DA COSTA, W. **A empresa digital**. Wladi Fatec. Acesso em: 08 de dezembro de 2013.

[9] DA COSTA, W. **Lojas renner melhoram resultados de suas campanhas de marketing direto com análise preditiva da spss**. Dmss. Acesso em: 10 de dezembro de 2013.

[10] DA SILVA, F. V. **Decisões com B.I. (Business Intelligence)**. Ciência Moderna, Rio de Janeiro, 2008.

[11] DARIO, B. R. **Data mart**. Dataprix. Acesso em: 20 de dezembro de 2013.

[12] DE ALMEIDA; ANTÔNIO, A. L. **Sistemas de informação geográfica dicionário ilustrado editora hucitec**. www.uefs.br. Acesso em: 08 de Dezembro de 2013.

[13] DE CARVALHO, L. A. V. **DATAMINING: A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração.** Ciência Moderna, Rio de Janeiro, 2005.

[14] DE COME, G. **Os metadados no ambiente de data warehouse**. *IV SEMEAD*, 1999.

[15] DE COMPUTAÇÃO CIENTÍFICA, L. N. **Tutorial de redes neurais aplicação em bioinformática**. www.lncc.br. Acesso em: 21 de dezembro de 2013.

[16] DE JESUS, F. **Pensamento e Planejamento Estratégico: uma abordagem competitiva**. Edição do Autor, Goiânia, 2011.

[17] DE PESQUISA EM ENGENHARIA DE ALGORITMO GPEA, G. **Pré-processamento em data mining**. www.din.uem.br. Acesso em: 05 de dezembro de 2013.

[18] DO NASCIMENTO, L. A. **Data warehouse arquitetura e implantação**. Nascimento. Acesso em: 10 de novembro de 2013.

[19] E EDUARDO O C, O. M. **Sistemas de informação e sistemas de apoio á decisão**. Ebah. Acesso em: 10 de dezembro de 2013.

[20] E FREITAS (H.), P. M. **Características desejáveis de um eis enterprise information system rumo á proatividade**. www.lume.ufrgs.br. A cesso em: 10 de dezembro de 2013.

[21] F, C. **Classificação de sistemas de informação**. Tecspace. Acesso em: 06 de Janeiro de 2014.

[22] FURTADO, L. S. C. R. S. S. **Clusterização em mineração de dados**. www.ic.uff.br. Acesso em: 20 de novembro de 2013.

[23] IBL. **Extração, transformação e carga**. www.infobras.com.br. Acesso em: 15 de dezembro de 2013.

[24] INMON, W. H. **Building the Data Warehouse Third Edition**. John Wiley and Sons Inc, New York, 2002.

[25] JUNIOR, M. A. **Oltp x olap**. MarquinhosNet. Acesso em: 10 de dezembro de 2013.

[26] LORENZI, C. A. **Arquiteturas de data warehousing parte 2**. Blog do Lito. Acesso em: 30 de dezembro de 2013.

[27] MICROSOFT. **Algoritmos de mineração de dados (analysis services mineração de dados)**. MSDN Microsoft. Acesso em: 15 de dezembro de 2013.

[28] MICROSOFT. **Dados hierárquicos (sql server)**. Technet Microsoft. Acesso em: 08 de dezembro de 2013.

[29] MORELLATO, L. **Mineração de dados e web semântica**. iMasters. Acesso em: 13 de dezembro de 2013.

[30] PANAZZO E MARIA LUISA VAZ, S. **Navegando pela historia**. Aprendendo a gostar de historia. Acesso em: 16 de dezembro de 2013.

[31] PERZEPIORSKI, E. **Análise de crédito bancário com o uso de data mining: Redes neurais e árvores de decisa o**. www.ppgmne.ufpr.br. Acesso em: 20 de dezembro de 2013.

[32] RENATO. **Data warehouses: Fundamentos, ferramentas e tendâncias atuais**. www.inf.ufsc.br. Acesso em: 13 de dezembro de 2013.

[33] ROCHA, F. L. **Bussines inteligence: arquiteturas e tecnologias**. Felipelirarocha. Acesso em: 14 de dezembro de 2013.

[34] SHARDA, E. T. J. E. D. K. R. **Business Intelligence: Um enfoque gerencial para a inteligência do negócio**. Bookman, Porto Alegre, 2009.

[35] SIDEMAR, J. O. **Porque business intelligence?** iMasters. Acesso em: 08 de Janeiro de 2014.

[36] SOUZA, M. **Ferramentas ol ap**. iMasters. Acesso em: 13 de Dezembro de 2013.

[37] SYBASE. **Sybase etl**. Sybase. Acesso em: 19 de dezembro de 2013.

[38] SYBASE, S. **Sap sybase iq**. Sybase. Acesso em: 14 de dezembro de 2013.

[39] UNIPRESS. **Bam business activity monitoring: inteligência de negócios em tempo real**. Unipress. Acesso em: 10 de Janeiro de 2014.

[40] VIPIN, P.-N. M. **Introdução ao Data Mining Mineração de dados**. Ciência Moderna, Rio de Janeiro, 2009.

[41] W, R. M. G. **Princípios de Sistemas de Informação: Uma abordagem gerencial**. Pioneira Thomson Learning, São Paulo, 2006.