CENTRO UNIVERSITÁRIO UNIVATES

CURSO DE SISTEMAS DE INFORMAÇÃO

**ANÁLISE DA UTILIZAÇÃO DE PRODUTOS EM UMA COOPERATIVA DE CRÉDITO UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS.**

**“ANÁLISE DE AFINIDADES COM ALGORITMOS DE APRIORI”**

**(a segunda parte, entre aspas podes tirar)**

Marcos Oscar Wassem

Lajeado, setembro de 2014.

Marcos Oscar Wassem

**SISTEMA DE ANÁLISE DE DADOS**

Trabalho de Conclusão de Curso I apresentado ao curso de Sistemas de Informação do Centro Universitário Univates, para obtenção parcial do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Evandro Franzen

Lajeado, novembro de 2014.

**RESUMO**

Atualmente existe uma grande dificuldade das instituições financeiras interpretarem a grande quantidade de dados e informações que são geradas diariamente, movimentações de crédito, prestação de serviços, gestão de pessoas. Muitas destas informações possivelmente úteis estão sendo desperdiçadas, ficando em oculto em relatórios convencionais. Com tais necessidades, o uso de ferramentas poderosas de mineração de dados são cada vez mais comuns, auxiliando na seleção de regras associativas relevantes que auxiliam em tomadas de decisões muito mais rápidas e com muito mais precisão. (este trabalho aborda/apresenta...tens que complementar, dizer um pouco do que será falado no TCC 1)

**ABSTRACT**

Currently there is a major difficulty in financial institutions interpret the vast amount of data and information which are generated diariamen thee, credit transactions, services, people management. Many des-tas possibly useful information are being wasted, getting hidden in conventional reports. With such requirements, the use of tools Mighty-sas data mining are increasingly common, assisting in the selection of relevant association rules that assist in decision-making much faster and much more accurately.

LISTA DE FIGURAS

[Figura 1 - Etapas do Processo de KDD 10](#Figura!0|sequence)

[Figura 2 - Ciclo de gerenciamento de dados em um sistema de Big Data 15](#Figura!1|sequence)

[Figura 3 - Algoritmo básico de um coletor de dados da web 22](#Figura!2|sequence)

[Figura 4 - Estrutura de um documento JSON 31](#Figura!3|sequence)

[Figura 5 - Componentes do MongoDB 32](#Figura!4|sequence)

[Figura 6 - Diagrama de casos de uso 36](#Figura!5|sequence)

[Figura 7 - Diagrama de classes 37](#Figura!6|sequence)

[Figura 8 - Diagrama de atividade 41](#Figura!7|sequence)

[Figura 9 - Modelo do banco de dados da ferramenta proposta 42](#Figura!8|sequence)

LISTA DE TABELAS

Tabela 1 - Cronograma de desenvolvimento da proposta 12

Tabela 2 - Lista de requisitos funcionais 34

Tabela 3 - Lista de requisitos não funcionais 35

Tabela 4 - Classe TipoColetor 37

Tabela 5 - Classe ConfiguracaoTipoColetor 38

Tabela 6 - Classe Coletor 38

Tabela 7 - Classe AtributosColetor 38

Tabela 8 - Classe DriverNoSQL 39

Tabela 9 - Classe TipoBanco 39

Tabela 10 - Classe ConfiguracaoTipoBanco 40

Tabela 11 - Classe Banco 40

Tabela 12 - Classe AtributosBanco 40

**LISTA DE ABREVIATURAS**

DM: Data Mining

JDBC Java Database Connectivity

API Aplication Programing Interface

BD: Banco de dados

SGDB: Sistema Gerenciador de Banco de Dados

CRM Gestão de relacionamento com o cliente

ARFF, Attibute Relation File Formar

JDBC Java Database Connectivity

API Aplication Programing Interface

SUMÁRIO

1 INTRODUÇÃO 9

1.1 Objetivos 10

1.2 Justificativa e relevância 11

1.3 Delimitação 11

1.4 Estrutura do documento 12

1.5 Cronograma 12

2 BIG DATA 13

2.1 Tipos de dados do Big Data 14

2.1.1 Dados estruturados 14

2.1.2 Dados não-estruturados 15

2.2 Requisitos de um sistema de Big Data 15

2.3 Exemplos de aplicação de um sistema de Big Data 16

3 MINERAÇÃO DE DADOS (DATA MINING) 17

3.1 Técnicas de mineração de dados 17

3.1.1 Classificação 18

3.1.2 Estimativa 18

3.1.3 Previsão 18

3.1.4 Análise de afinidades 18

3.1.5 Análise de agrupamentos 19

3.2 Mineração da internet (Web Minning) 19

3.2.1 Coletores de dados da web 20

3.2.1.1 Algoritmo básico de um coletor 21

4 BANCO DE DADOS 23

4.1 Sistemas de banco de dados relacionais 24

4.1.1 PostgreSQL 25

4.2 Sistemas de banco de dados NoSQL 25

4.2.1 Características dos sistemas NoSQL 26

4.2.1.1 Escalabilidade horizontal 26

4.2.1.2 Alta disponibilidade 27

4.2.1.3 Ausência de esquema ou esquema flexível 27

4.2.2 Modelos de dados NoSQL 27

4.2.2.1 Modelo chave-valor 28

4.2.2.2 Modelo orientado a colunas 28

4.2.2.3 Modelo orientado a documentos 28

4.2.2.4 Modelo orientado a grafos 29

4.2.3 Implementações de sistemas NoSQL 29

4.2.3.1 Cassandra 29

4.2.3.2 MongoDB 30

4.2.3.3 Apache CouchDB 32

5 PROJETO DA SOLUÇÃO PROPOSTA 33

5.1 Tecnologias utilizadas 33

5.2 Descrição do sistema 33

5.3 Levantamento de requisitos 34

5.3.1 Requisitos funcionais 34

5.3.2 Requisitos não funcionais 35

5.4 Casos de uso 35

5.5 Diagrama de classes 36

5.6 Diagrama de atividade 40

5.7 Modelagem do banco de dados 41

5.8 Captura de dados 42

5.8.1 Coletor de páginas da internet 42

5.8.2 Coletor de mensagens do Twitter 43

5.9 Interfaces do sistema 43

6 CONSIDERAÇÕES FINAIS 44

1 INTRODUÇÃO

O mundo vem passando por profundas modificações no que diz respeito aos aspectos sociais e econômicos. Neste cenários de competições globais as instituições financeiras tem buscado na tecnologia recursos que agreguem valor aos seus negócios, seja agilizando operações ou viabilizando inovações. A grandes quantidade de informações geradas diariamente tem dificultado muito a análise destes dados, o que leva estas instituições a buscar ferramentas que possam auxiliar na análise dos mesmos, com eficiência no resultado, e agilidade no processamento do mesmo.

Em um mercado extremamente competitivo, faz-se necessário melhorar o processo de aquisição de novos clientes. Tendo em vista que a baixa produtividade das campanhas publicitárias das instituições financeiras traz como principal motivo o de tentar atingir um grande número de pessoas com promoções muito genéricas, que muitas vezes não agrada um determinado grupo de clientes. Em todas as intuições financeiras, o fator é agravado pelo fator de as áreas de negócios ser dividida em produtos, o que dificulta a visão em um todo sobre todas as informações dos seus clientes.

Considerando esta realidade e a necessidade de analisar dados para elaboração de campanhas de produtos, este trabalho aborda um tema que ganha muita importância a cada ano que passa, mineração de dados (Data Minning), que é um conjunto de ferramentas e técnicas que busca por padrões, associações nos diversos dados analisados. Estes resultados que podem ser apresentados em diversas formas, com muita precisão e relevância, formando grupos com potencial para determinados produtos, evitando desperdício de dinheiro e tempo em propagandas desnecessárias para pessoas que não vão utilizar determinado serviço ou produto.

Este estudo visa apresentar os conceitos e técnicas de Mineração de Dados, e aplicar os conhecimentos adquiridos para achar um novo grupo de clientes, utilizando Inteligência Artificial para achar semelhanças ou padrões. As informações sobre os clientes serão coletadas de uma base de dados previamente cadastrada, para depois serem utilizadas na definição de perfil do cliente ideal para cada grupo de produtos oferecidas pela intuição, assim reduzindo tempo, oferecendo produtos que cabe a cada grupo.

Tal estudo foi realizado em uma cooperativa de crédito localizada na cidade de Teutônia.

**1.1 Objetivos**

**(objetivo tem que ser mais sucinto e tem que ter alguns objetivos específicos, tem texto nesta parte que vai na introdução ou na justificativa. Tens que dizer qual ojetivo geral que é analisar dados e demonstrar resultados de forma que estes sejam interpretados por gestores e outros objetivos gerais como contribuir para melhorias, compreender aspectos de mineração de dados, entre outros)**

Atualmente existe uma grande dificuldade das Instituições financeiras em conseguir analisar sua grande quantidade de informações de uma maneira total, tendo em vista que a base de dados é muito grande, e é praticamente impossível conseguir achar alguma regra de associação sem a ferramenta correta.

O presente trabalho tem como objetivo principal fazer a análise de uma base dados através de rotinas, compreendendo as tecnologias utilizadas para a mineração dos dados, e assim buscando uma regra de associação entre diferentes clientes, trazendo assim um novo grupo de clientes, com mesmo perfil, assim contribuindo para o desenvolvimento de aplicações de mineração de dados.

Será feito uma análise em uma base de um arquivo CRM que contém dados de mais de 40 mil clientes, divididos em 31 produtos. Nesta base de dados é discriminado quais produtos que cada cliente possui, assim podendo fazer uma análise para busca de grupos de clientes com determinado perfil. (no objetivo não vai o que será feito, pode ser na introdução ou justificativa)

As instituições financeiras buscam oferecer vários tipos de produtos, neles inclusos débitos em conta, que tem como principais objetivos a comodidade para seu cliente, e o mais importante, desafogar a grande quantidade de pessoas que buscam essas instituições para pagar boletos de conta de água, luz, telefone, etc. Com a análise pretende-se achar uma relação nos clientes que possuem estes serviços, e poder visualizar qual os perfis dos mesmos. A análise dos clientes que possuem seguro de vida, podendo achar uma relação nos clientes que possuem tal serviço, e oferecendo o mesmo ao mesmo grupo de clientes que não possua o mesmo. (aqui também tem coisa fora dos objetivos, coloca algo na justificativa)

Com este estudo pretende-se tonar os clientes mais rentáveis, entender as características de cada cliente, reter clientes, melhor a satisfação e aumentar o número de serviços que cada cliente possui.

**1.2 Justificativa e relevância**

Diante da grande quantidade de informações que as instituições financeiras possuem, faz se necessário um estudo para de que maneira estas informações podem ser úteis para oferecer novos produtos a clientes já existentes.

Para a empresa, o estudo será de suma importância, pois através do mesmo poderemos conhecer melhor nossos clientes, e planejar novas ações específica em cada novo grupo de clientes formado, economizando tempo, dinheiro e atuando especificamente para cada grupo de clientes que possuem determinado perfil.

Este trabalho faz parte de uma iniciativa que busca ajudar os gestores a ver a grande quantidade de dados de uma maneira mais simples, e não ficando somente a cargo de uma pequena quantidade de informações as decisões tomadas, ações de marketing que podem ser mudadas, assim economizando tempo e dinheiro.

A mineração de dados é importante para que possamos manipular esta grande quantidade de informação, assim ganhando tempo e gerando novos conhecimentos para seus modelos de negócio.

Hoje a instituição possui uma grande quantidade de serviços e produtos, onde muitas vezes os atendentes não sabem qual produto oferecer para os clientes que procuram nossa instituição para qualquer assunto. Um exemplo para o mesmo: Suponhamos que um cliente venha até a nossa instituição para fazer uma solicitação de empréstimo. Seria muito mais prático o atendente saber a qual grupo de clientes ele se encaixa e quais produtos estes grupo de cliente normalmente possui, assim além do empréstimo, oferecer mais algum produto que possa ser de interesse do mesmo.

**1.3 Delimitação**

Observando-se as diversas técnicas utilizadas para a análise da grande quantidade de dados, com o objetivo de achar uma relação entre clientes, nota-se que para a maioria destes processos não são utilizados softwares específicos para esta análise, ou técnicas corretas para Mineração de dados, assim dificultado que os analistas forneçam resultados precisos, de forma clara e gerencial para a tomada de decisão.

Com este princípio o presente trabalho busca a descoberta de novos conhecimentos sobre dados que encontram-se em diversas planilhas. Será utilizado vários algoritmos de Mineração de Dados a fim de trazer de forma organizada e eficiente os resultados procurados.

O presente trabalho, não terá como objetivo e resultado o desenvolvimento de uma ferramenta de mineração de dados, e sim um estudo de ferramentas, técnicas e softwares de mineração de dados, além de critérios de avaliação de uma solução, é adequado que a pessoa que fizer o uso das ferramentas alguém já envolvido no processo e que possa aliar a sua experiência com os dados obtidos, para ter soluções finais de melhor qualidade.

(aqui tem que ter alguns parágrafos indicando que no capítulo 2 serão descritos fundamentos, etc, no capítulo 3....e assim por diante)

(nova página)

**2 REFERENCIAL TEÓRICO**

O presente referencial apresenta o estudo sobre as técnicas de mineração de dados e a descoberta de conhecimento em base de dados, utilizando exemplos de algoritmos, destacando a particularidade para cada área da Mineração de Dados.

2.1 Mineração de dados

Atualmente a grande quantidade de informações que são geradas diariamente nas mais diversas formas de texto, sons, imagens etc, ultrapassou a capacidade humana de analise, interpretar e utilizar as mais diversas informações contidas nelas, assim tornando-se necessário a utilização de ferramentas que são capazes de manipular esta grande quantidade de informações Fayyad (1996).

Para análise, os dados precisam primeiramente serem capturados, e após esta fase, os dados podem ser analisados, processados e transformados com base no problema a ser resolvido. Através dos resultados destas análises, produzimos o conhecimento que é capaz de nos auxiliar nas tomadas de decisões, assim podendo decidir a melhor ação a ser tomada para determinada situação.

De acordo com Carvalho (2005) e Fayyad (1996), a Descoberta de Conhecimento em Base de Dados (Knowledge Discovery in Databases - KDD) é um processo que envolve a automação da identificação e do reconhecimento de padrões em um banco de dados. Sua principal característica é a extração não-trivial de informações a partir de uma base dedados de grande porte. Essas informações são necessariamente implícitas, previamente desconhecidas e potencialmente úteis.

Segundo Carvalho(2005), mesmo que as técnicas de mineração de dados sejam antigas, apenas nos últimos anos passaram a ser utilizadas como exploração de dados pelos seguintes motivo:

- O volume de dados disponível atualmente é enorme:

- Os dados estão sendo organizados:

- A competição empresarial exige técnicas mais modernas de decisão:

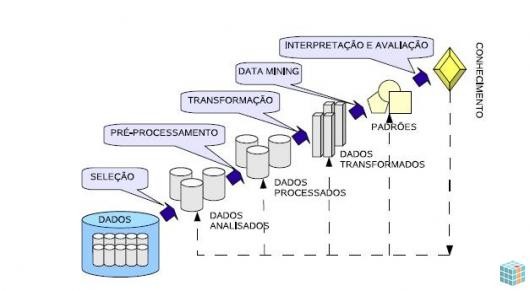
Um dos principais fatores para aceitação de softwares de Data Mining é a facilidade de uso das mesmas, funções sofisticadas que se encontram no próprio software, desta forma o usuário não precisa ter um grande conhecimento das técnicas para ter seus resultados em telas, ou por meio de relatórios impressos.

**2.2 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS KDD**

Segundo FAYYAD (1996), os modelos tradicionais de transformação dos dados em descoberta de informação, consiste em um processo manual onde todas as informações são analisadas por especialistas que produzem relatórios que poderão ser analisados. Na grande maioria das vezes, devido à grande quantidade de dados, esses processos manuais tornam-se impraticáveis. Ainda de acordo com FAYYAD (1996), o KDD ou descoberta de conhecimento nas Bases de Dados é uma tentativa de solucionar o problema de sobrecarga de dados.

Para FAYYAD (1996), existe uma grande diferença entre Data Mining e KDD, muitas vezes considerados mesmos. KDD consiste ao processo de conhecimento, e Mineração de dados, uma das atividades do processo.

Na figura 01 podemos ver representado o processo do KDD

Figura 1: Etapas do Processo de KDD ,

Fonte: STEINER et al., 2006

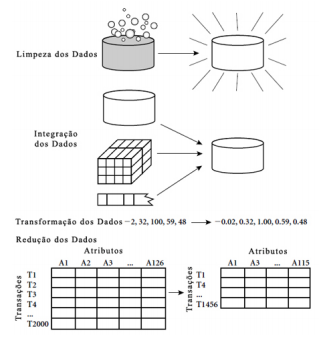
De acordo com Fayyad o KDD é definido como “um processo não trivial de identificação de novos padrões válidos, úteis e compreensíveis”.

**2.1.1 Os dados**

Para escolher o(s) melhor(es) método(s) de busca e processamento dos dados é fundamental conhecer os tipos de dados a serem trabalhados. Podendo classificar os dados em dois tipos: qualitativos e quantitativos. Onde os dados qualitativos contêm valores nominais, e dados quantitativos são representados por valores numéricos. Portanto para poder aplicar algum algoritmo de mineração de dados precisamos explorar, conhecer e trabalhar os dados.

Segundo Han (2006), preparação dos dados para a mineração, também pode ser chamado de pré-processamento, que consiste em:

Figura 02: Atividades do pré-processamento



Fonte: Han et al.

Os quatro processos de transformação dos dados serão descritos no capítulo a seguir:

**2.1.1.1 Limpeza de dados**

Muitas vezes os dados são encontrados com diversos erros de registros, incompletos, dados inconsistentes ou incorretos. Esta etapa visa eliminar estes registros incorretos, impedindo que eles influam no resultado das pesquisas. Existem várias técnicas para a limpeza destes registros incorretos, desde a remoção do registro, quanto a aplicação de técnicas de agrupamento que possam auxiliar na descoberta de novos valores. De acordo com Han (2006), devido ao grande esforço para nesta etapa, devemos utilizar um processo específico para a limpeza dos dados.

**2.1.1.2 Integração dos dados**

Em alguns casos, os dados a serem minerados vem de diferentes fontes, como planilhas, arquivos de texto, data Warehouse, imagens, entre outros. Para tal integração faz se necessário uma profunda análise dos mesmos, observando redundância, valores conflitantes, e dependência entre as variáveis e regras diferentes para os mesmos dados.

**2.1.1.3 Transformação dos dados**

Esta pode ser considerada uma das mais importantes etapas da pré-processamento, pois alguns algoritmos somente trabalham com valores numéricos e outros somente com valores categóricos. Neste caso precisamos transformar dados numéricos em categóricos ou dados categóricos em numéricos. Existem várias técnicas empregadas nesta etapa: Suavização (Remove dados incorretos), generalização (Transforma valores muito específicos em valores genéricos), normalização (coloca os valores nas mesmas escalas), criação de novos atributos (gera dados a partir de outros dados já existentes) e agrupamento (agrupa os dados nas mesmas faixas).

**2.1.1.4 Redução dos dados**

O volume de dados utilizado para a mineração de dados normalmente é muito grande, em alguns casos tornando-se inaplicável. Nestes casos, a redução de dados na massa de dados original pode ser aplicada, tornando uma massa de dados menor, porém sem perder a representatividade dos dados originais. Assim ganhando em agilidade no processamento dos dados, mantendo a qualidade do resultado.

**2.2 TÉCNICAS (técnicas de que? Só técnicas não é um bom título, acho que já tinha comentado isso)**

**(aqui parece faltar uma introdução a data mining, alguns parágrafos indicando o que é,porque dentro do KDD tem uma etapa de data mining e aí podes falar das técnicas)**

Devido à grande quantidade de dados, e diferentes tipos de informações, não a uma técnica de mineração que possa ser utilizada para analisar todos os tipos de dados. Cada problema possui sua peculiaridade e assim possui sua forma de resolução.

Para Dias (2001), a familiaridade entre as técnicas é necessário para que possamos escolher a mais adequada para solução do problema apresentado.

As técnicas para Mineração de dados consiste basicamente em extensões neurais ou generalização de métodos analíticos, que auxiliam os gestores nas tomadas de decisões, estratégias para os negócios.

Existem várias técnicas de mineração de dados, que consiste em especificação de métodos que nos garantam uma maneira de como descobrir os padrões que nos interessam.

A seguir são descritas algumas técnicas de mineração de dados:

**2.2.1 Classificação**

Para Carvalho 2001, o método mais utilizado para Mineração de Dados é o de classificação, pelo fato de ser uma das tarefas mais utilizadas pela humanidade para compreensão do ambiente em que vivemos. Segundo Fayyad (2001), ela pressupõe características que classifica ou associa um item de acordo com classes pré-definidas.

A tarefa de classificação visa identificar a qual classe o registro pertence. Neste modelo faz se a análise de registros fornecidos, com cada registro qual já tem indicação a classe pertencente, a fim de classificar um novo registro.

As tarefas de classificação podem ser usada para:

- Identificar para qual turma determinado aluno se encaixa.

- Diagnosticar onde uma doença pode estar presente.

Os algoritmos clássicos utilizados para a classificação baseiam-se em árvores de decisão, regras de decisão e análise discriminante.

**2.2.2 Sumarização**

É uma das primeiras funções de Mineração de dados, visa principalmente orientar e motivar análises posteriores, mais complexas. Análises exploratórias de dados, estatística descritiva.

Nos processos de mineração de dados, a sumarização tem como principais funções a visualização para observar alguma característica estrutural em sua base de dados. A visualização é uma das mais importantes e poderosas ferramentas para análise de dados, muitas vezes sendo o suficiente para uma tomada de decisão.

**2.2.3 Associação**

Segundo Agrawal (1994), este modelo procura descrever dependências significativas entre variáveis, podendo ser dividido em dois níveis: quantitativo e estrutural. No modelo quantitativo são utilizados escalas numéricas para achar as dependências entre as variáveis, enquanto no modelo estrutural, as variáveis são especificadas localmente, sendo uma dependente da outra.

Um exemplo para a utilização de uma regra de associação, seria em uma loja, onde a partir de uma base de dados criar a seguinte regra de associação: {cinto, calça} = {sapato}, com isto podemos definir que com um determinado grau de certeza, o cliente que compra cinto e calça provavelmente compre um sapato. Este grau é definido por dois índices, o grau de suporte e o de confiança.

Um outro bom exemplo é o dos grandes varejos, que conseguem através da venda de seus milhares de produtos adquirir conhecimento através da mineração de dados, organizar seus processos de marketing e promover catálogos que possam promover melhor cada grupo de clientes.

**2.2.4 Regressão**

É similar a classificação e estimação, porém ela ainda busca descobrir o valor futuro de um determinado item, como por exemplo, descobrir o valor de uma ação três meses adiante.

Com a popularização dos computadores, tornou-se possível a utilização de técnicas de regressão. Atualmente estas ferramentas são encontradas em diversas plataformas de computação, até mesmo no Microsoft Excel.

**2.2.5 Análise de cluster ou agrupamento**

Segundo GOLDSCHMIDT, 2005 com a Análise dos componentes principais, podemos obter conjuntos e coordenadas para descrever de forma simplificada os dados. Assim reduzindo as dimensões dos dados e sem perder informações importantes.

Com a aplicação de Análise de componentes principais podemos determinar as variáveis que possam explicar o conjunto de dados encontrados, assim reduzindo a dimensão do conjunto encontrado.

Para a análise de cluster, podemos usar o seguinte exemplo: Imagine que você seja dono de uma grande loja, onde você vende vários tipos de produtos diferente, e onde tenha em seu banco de dados cadastrados vários campos que descrevem seu cliente, como por exemplo, nome, endereço, profissão, renda mensal. E a sua empresa está querendo enviar um material promocional de eletrônicos que estão em promoção, para evitar despesas desnecessárias, gostaria de enviar estas promoções somente aos clientes que sejam potenciais compradores. Tomando por base as informações que dispõe no banco de dados, você quer saber quais clientes podem ser potencialmente compradores, isto é, dividir sua grande base de dados em clientes que compram e que não compram produtos eletrônicos, podendo ter como base uma regras aplicadas em idade, profissão e renda mensal, onde por exemplo podemos colocar que se o cliente está entre 20 e 30 anos, e renda mensal acima de R$ 2.000,00 mensais, então ele se torna um potencial cliente para compra de eletrônicos.

A fase de clustering frequentemente está presente nas primeiras fazes de mineração de dados, com o intuito de reunir os registros de grupos semelhantes com o intuído de utilizar nas fazes sequentes.

Seu maior objetivo é classificar, com base em um conjunto de variáveis consideráveis, selecionar indivíduos para formação de um novo grupo relativamente homogêneo.

Segundo Malhotra (2001), a fase de agrupamento também pode ser chamada de análise de classificação, taxonomia numérica ou análise Q.

**2.4 ALGORITMOS PARA MINERAÇÃO DE DADOS**

Os algoritmos de Mineração de dados são heurísticas e cálculos que criam os modelos de Mineração de Dados. Para criar um algoritmo de Mineração de Dados, primeiro precisa-se fornecer os dados, para então ser feita uma análise do padrões e tendências existentes. A partir desta análise dos resultados se verifica os parâmetros ideais para criar o modelo de Data Mining

(colocar uma parágrafo indicando que algoritmos podem seguir diversas abordagens e que a seguir serão descritos alguns destes...)

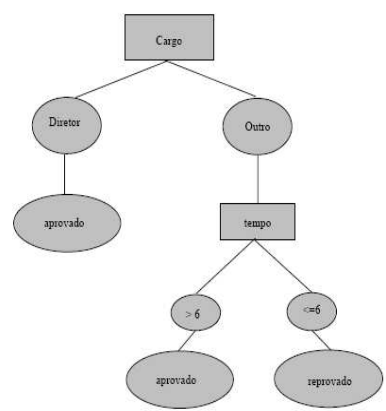
**2.4.1 ÁRVORE DE DECISÃO**

É possível definir como árvore de cisão um modelo representado geograficamente por nós e ramos, como o próprio nome diz, uma estrutura em árvore, mas no sentido invertido Han & Kamber (2006). Onde o nó da raiz é o primeiro nó da árvore, no começo da estrutura. E a parte que contém os nós internos são os nós de decisão.

A arvore de decisão utiliza a técnica de recursividade para montar a base da árvore. Separando classes diferentes a fim de serem alocadas em subconjuntos diferentes, cada qual com a sua regra.

Na figura 03 apresenta um modelo no qual se aprovaria ou reprovaria algum pedido, conforme o cargo e tempo de serviço na empresa.

Figura 3: Estrutura de uma árvore de decisão



Fonte: Cunico (2005 página 29)

Uma das principais vantagens para o uso da árvore de decisão em relação as demais técnicas, é uma melhor apresentação dos resultados.

**(não tem nenhum algoritmo de árvore de decisão, falastes da fundamentação, mas não de algum algoritmo, tens que citar pelo menos um)**

**2.4.2 Descoberta de Regras de Associação**

A descoberta de regras de Associação é muito utilizada na área do comércio, que busca por regras e padrões na compra, buscando orientar os gestores a buscar as melhores ações de venda. A Descoberta de Regras de associação busca a relação em um grande conjunto de dados que ocorrem com muita frequência. Um dos pontos positivos do uso da DRA é a confiança pois ela limita a quantidade de regras a serem extraídas, trazendo a descrição destas regras.

A técnica de DRA é muito utilizada em alguns algoritmos que implementam as regras de associação, como por exemplo o algoritmo de Apriori.

**2.4.3 ALGORITMO DE APRIORI**

**(apriori é um algoritmo para associação, não coloca em um novo subitem, mas dentro de associação)**

No contexto de mineração de dados, um dos algoritmos de busca por associação mais utilizado é o *Apriori*AGRAWALL (1994), que simplesmente faz uma varredura no conjunto de dados, e busca associações que tenham relações e que sejam frequentes.

O algoritmo de Apriori faz a busca por associação em duas etapas (Freita et al.,2001)

- Geração: aonde faz-se uma varredura na base de dados, e gera todos os conjuntos de possíveis combinações de campo aonde as colunas tenham uma frequência maior ou superior a mínima.

- Poda: considera-se somente o conjunto de dados que possua um grau de confiança mínimo, e os demais se descarta.

Um exemplo prático para a regra de associação, seria que em um conjunto de dados, N% de registros que tenham o atributo A e B, constem no C. Assim podemos constatar que a N% representa o fator de confiança da regra, e assim elimina tendências mais fracas e mantém as tendências mais forte.

O algoritmo faz a busca dos itens mais frequentes, que possuam suporte ao valor máximo ou ao mínimo exigido, desta forma, é necessário fornecer suporte ao valor mínimo

e outro referente a confiança mínima além do arquivo de itens de transação.

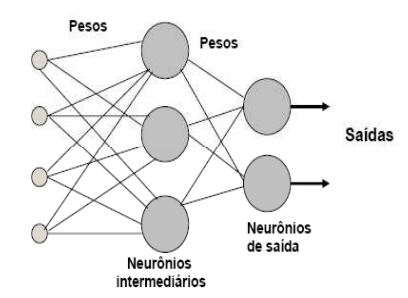
**2.4.4 REDES NEURAIS E ARTIFICIAIS**

Segundo Almeida (2009), as Redes Neurais procurar imitar o cérebro humano para a solução de problemas, assim como o ser humano, aplica conhecimento adquirido em experiências anteriores para resolver os problemas ou situações criados, de igual modo utiliza exemplos previamente resolvidos para resolver os novos problemas.

A técnica de Redes Neurais é indicada para a resolução de problemas que envolvam clusterização, classificação, estimativa. Uma das principais vantagens da utilização das RNA é sua variedade de aplicações, mas um das principais dificuldades é a formação da entrada dos dados, e os modelos produzidos são difíceis de serem entendidos.

Na figura 04 mostra-se um exemplo de uma rede neural, que está composta pelas entradas, compostas pelos dados da base de dados, os pesos que são atribuídos para a avaliação destas entradas, onde são calculados conforme a rede é treinada para melhorar os dados obtidos, os pesos intermediários que são vistos pelo usuário, e a saída que já possuem os pesos relacionados, e que é visto pelo usuário.

Figura 04: Exemplo de uma Rede Neural



Fonte: Lemos (2003 p55)

**2.4.4.1 MODELOS DE REDES NEURAIS**

**(cuidar da formatação, o tamanho de letra deste título está incorreto)**

Quando as redes neurais foram propostas, muitos modelos foram sugeridos para a implantação deste tipo de rede. Cada modelo oferecendo sua própria arquitetura atingindo problemas e aplicações especificações. No capítulo seguinte será descrito o modelo Backpropagation.

**2.4.4.1.1 Modelo Backpropagation**

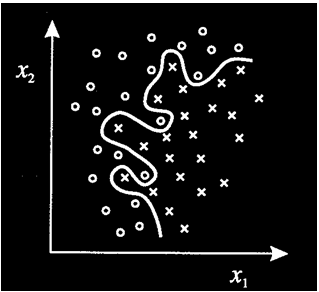
A rede neural Backpropagation utiliza a regra de Delta para o processo de aprendizagem. Uma descrição para o nome Backpropagation é propagação de erro para trás. O algoritmo se baseia na retro propagação dos erros para o ajustes dos pesos nas camadas internas da rede Pádua (2000). Simplificando, toda vez que ela recebe algum padrão, ela tenta adivinhar aleatoriamente o que entrada deve ser

Os padrões podem ser visto como um conjunto de pares ordenados {(x1, y1), (x2, y2), (x3, y3)...}, onde y representa o valor de padrão de saída, associado ao valor de entrada x.

A principal vantagem em se utilizar o algoritmo backpropagation, é que o mesmo trabalha com multicamadas, assim é capaz de resolver problemas “não linearmente separáveis”, problemas que alguns algoritmos não conseguem resolver.

Na figura 05 podemos observar claramente que não é possível separar o padrão “o” do padrão “x” apenas traçando uma reta.

Figura 5: Problema não linearmente separável



Fonte: Desconhecido

(só falastes de um modelo, mas o título do subitem é modelos de redes neurais, terias que falar de mais um, pelo menos, tem vários, até mesmo mapas auto organizáveis, que são para clusterização, mas tem outros...)

**2.4.1 ANÁLISE DE AGRUPAMENTO OU CLUSTER**

Segundo Haldiki (2001), o algoritmo de análise de cluster é um dos mais utilizados para a descoberta de agrupamentos, identificação e padrões para entendimento dos dados.

(não é um algoritmo de análise de cluster e sim, vários, como k-medias, x-medias, entre outros, muda este parágrafo)

Sem fazer o agrupamento das informações de um banco de dados, é praticamente impossível entender todas as informações que nele existe. Este processo busca separar o conjunto de dados em padrões que refletem algum tipo de comportamento, separando as informações do banco de dados de forma que cada grupo tenha alguma similaridade. A partir do momento em que os padrões tenham sido estabelecidos, podemos utilizar estes critérios para desmontar nossa base de dados em diversos grupos, tornando assim as informações mais compreensíveis.

Na clusterização, as classes são resultantes da Mineração de Dados, diferente da classificação, aonde as classes são pré-determinadas pelo usuário. Para melhor compreensão, podemos utilizar o seguinte exemplo: agrupar clientes de uma loja que possuam similaridades em suas compras.

Porém um dos maiores desafios para a utilização da clusterização, é que algumas vezes os resultados não são compreensíveis, do ponto de vista de análise, à medida que a quantidade de informações e de atributos aumentam, aumenta a complexidade da análise.

De acordo com Han (2006), o algoritmo de clusterização pode ser dividido em cinco categorias principais.

- *Métodos de particionamento:* a pesquisa é iniciada com n objetos, o método constrói k partições, onde cada partição representa um grupo, e k ≤ n. Este conjunto de dados satisfaz 2 requisitos. Cada grupo deve ter ao menos um objeto. E cada objeto pertence ao mesmo grupo.

- *Métodos Hierárquicos:* este método decompõe a base de dados por base da hierarquia, formando grupos que podem ser divisiva ou aglomerativa. Na divisiva cada objeto pertence a um grupo, que vai sendo repartido a cada interação. Na aglomerativa, cada objeto inicia como sengo um grupo, e com o avanço da análise ele é ligado a um grupo similar.

- *Métodos baseados na intensidade:* este método se baseia de aumentar um dado grupo, até que o mesmo exceda algum limite ou quantidade estabelecida.

- *Métodos baseados em grade:* os objetos são separados em um número finito de células, formando uma estrutura em grade. As operações de agrupamento são executadas baseados nesta grade. Uma grande vantagem é a velocidade de execução do mesmo.

- *Métodos baseados em modelo:* é criado um modelo hipotético para cada grupo, assim encontrando o melhor ajuste para o objeto ao modelo dado.

**(aqui também falta descrever os algoritmos de agrupamento, como citei acima)**

**2.5 BANCO DE DADOS**

Os primeiros softwares desenvolvidos nos primórdios da computação, tinham em seus sistemas todas as funcionalidades para o funcionamento do mesmo, interação com o usuário e até o acesso e gravação de dados, porém com a evolução dos sistemas percebeu-se a dificuldade de compartilhamento de informações entre aplicações, uma vez que os códigos de gerência ficava dentro das aplicações.

Podemos definir Banco de dados como um conjunto de dados que busca atender um ou mais usuários ao mesmo tempo (HEUSER, 1998.), sendo projetado com os dados para um propósito específico.

O software de gerenciamento de banco de dados, mais conhecido como SGBD (Sistema gerencial de banco de dados), surgiu na década de 70 com o intuído de auxiliar e facilitar o desenvolvimento de sistema que utilizem dados armazenados. Os primeiros softwares de SGBD eram caros e difíceis de utilizar. Mas com o passar dos anos, a popularização da informática, e o barateamento das tecnologias, houve um grande aumento nas pesquisas na área de abanco de dados. Estes investimentos em pesquisa resultou em um tipo de SGBD, que praticamente tornou-se padrão para todas as implantações do mercado de SGBD relacional.

(só isso de banco de dados? Ficou meio perdido, poderias falar um pouco do modelo relacional e deste tipo de armazenamento, justificando que muitos dados encontram-se em outro formato, como planilha e por isso é necessário converter para BD)

**2.6 GESTÃO DE RELACIONAMENTO COM O CLIENTE CRM**

O CRM ou (Customer Relationship Management) é uma tecnologia que direciona o relacionamento com os clientes, trazendo o máximo de informação de cada cliente, permitindo manipulação de dados, e assim buscando o atendimento mais personalizável para cada cliente.

De acordo com SWIFT (2001) a definição de CRM incluiria todas as atividades que transformam clientes eventuais em clientes leais, satisfazendo ou excedendo as exigências deles, de tal forma que eles voltem a comprar.

Para a área de Tecnologia da Informação, é um software que integra vários setores da empresa, possibilitando fazer a análise de todos os dados obtidos, possibilitando traçar estratégias de negócios voltados ao encantamento do cliente.

De acordo com SWIFT (2001) CRM não diz respeito a preços, ao envio de grande quantidade de correspondências ou muitas ligações irritantes para clientes em potencial. Definitivamente, não diz respeito a utilização dos canais para direcionar os clientes para os concorrentes e sim fornecer a capacidade de gerar produtos, serviços, respostas, individualização, personalização em massa e satisfação do cliente.

(também ficou muito sucinto, é ruim ter um assunto abordado de forma tão básica, teria que ter mais detalhes, também acho que anteriormente tínhamos comentado um pouco sobre isso)

**(nova página)**

**3 FERRAMENTAS PARA MINERAÇÃO DE DADOS**

O foco deste trabalho não é o desenvolvimento de uma ferramenta para DM, e sim a análise de softwares de DM e dos vários algoritmos que podem ser utilizados. Neste capítulo, é dada uma visão geral dos softwares capazes de interpretar estes algoritmos.

Existem muitas ferramentas de mineração de dados, mas foram escolhidas duas ferramentas em específico para um estudo mais aprofundado, Weka e o RapidMiner, ambas ferramentas gratuitas desenvolvidas na plataforma Java.

Para a presente pesquisa, será utilizar a ferramenta do WEKA, pelo fato de ser uma ferramenta grátis, desenvolvida em Java e de fácil utilização. Outra motivação para o desenvolvimento do presente trabalhos, é a dificuldade em encontrar artigos com exemplos de utilização do WEKA no data mining, de regras de associação.

**3.1 WEKA**

Para a análise dos dados foi utilizado a ferramenta Weka *(Waikato Environment*

*for Knwoledge Analysis),* desenvolvida pela universidade de Waikato, Nova Zelandia. De acordo com Witten & Frank (2000) o Weka é um pacote implementado em Java, segundo o paradigma de orientação ao objeto, é composto de uma série de algoritmos de aprendizagem para solucionar problemas de mineração de dados. Dentre os vários algoritmos existentes, para este estudo foi utilizado o algoritmo de Apriori, implementado pela ferramenta e que faz o uso de associação.

Na figura 06 temos a tela inicial do Weka. Ela fornece as opções de Explorer, Experimenter, KnowledgeFlow, Simple CLI, todas estas opções descritas são opções que fornecem interação com o usuário, mas a única que não tem representação gráfica é a opção Simple CLI. A opção mais utilizda é a opção Explorer, que pode será estudada nos tópicos a seguir.

Figura06: Tela inicial Software Weka



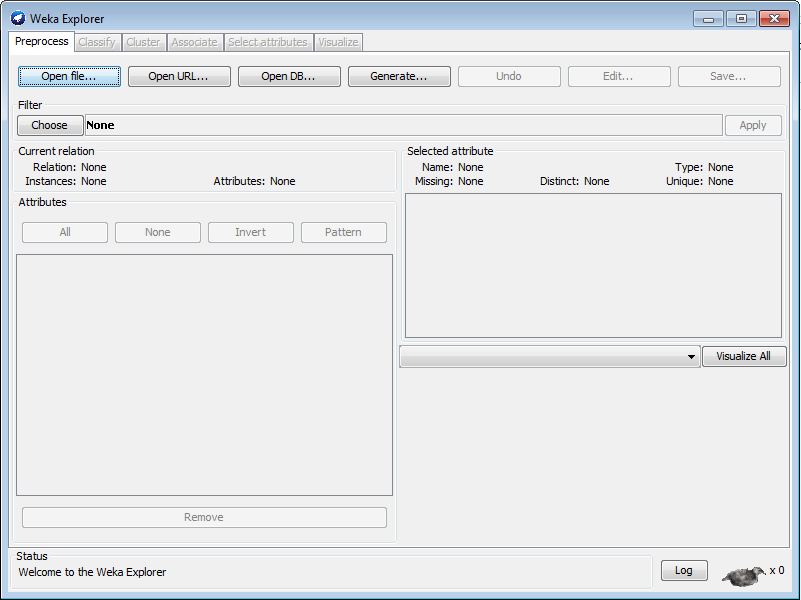
Fonte: Weka (2010).

Nesta aba de pré-processamento (Process), pode ser aberto um arquivo que tem formato ARFF, que é suportado pela ferramenta do WEKA, a extensão ARFF é descrita no tópico 3.1.1, nesta aba, com os dados carregados é possível realizar a tarefa de Mineração de dados.

Na aba de Cluster, utiliza os arquivos para achar semelhança entre os arquivos processados, formando novos grupos com perfil semelhantes.

A aba associação serve para agregar as regras encontradas, a partir da utilização dos algoritmos é feita a associação de atributos. Para selecionar as regras de relevâncias dos dados, utiliza-se a aba Seleção de Atributos. Por fim a visualização do conhecimento é feito na aba Vizualize. A figura 07 apresenta a tela do Explorer com a aba Preprocess selecionada.

Figura 07: Tela Preprocess



Fonte: Weka 2010

O Weka é uma ferramenta de Mineração de dados com uma grande crescente de utilização nos últimos anos, e foi escolhida por vários motivos: facilidade de instalação e implementação, ferramenta desenvolvida em Java, e encontra-se na Internet no site (<http://www.cs.waikato.ac.nz/~ml/weka/>). A equipe lança periodicamente atualizações de correções para o aplicativo, além de haver vários grupos de discuções sobre o mesmo.

O sistema possui uma interface amigável, seus relatórios fornecem dados analíticos e estatísticos em domínio minerado.

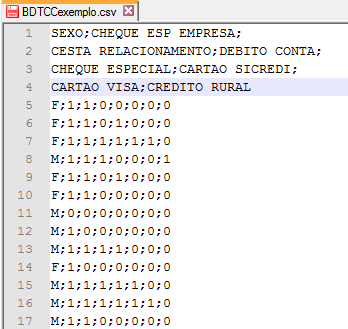
**3.1.1 FORMATO ARFF**

Para a análise pela ferramenta Weka, é necessário transformar os arquivos de um banco de dados em um arquivo com extensão \*.arff, (Attibute Relation File Formar). Esta extensão foi desenvolvida na universidade de Waikato, para ser utilizada pela ferramenta do WEKA. Primeiro é preciso exportar a base de dados para um formato com delimitação, separado por vírgulas, e ser salvo com a extensão .CSV. Após o arquivo ser salvo nesta extensão ele é aberto em um simples editor de texto e salvo na extensão. arff. E para transformar este arquivo em um arquivo que o Weka possa interpretar precisamos colocar no início do arquivo colocar um nome para a base de dados, como por exemplo @DMCRM, nas subsequentes, coloca-se os atributos para o conjunto de dados, como por exemplo numeric, para valores reais ou numéricos.

Após a descrição de todos os atributos, coloca-se @data, para que o algoritmo de mineração de dados possa entender que após vem os dados a ser minerados.

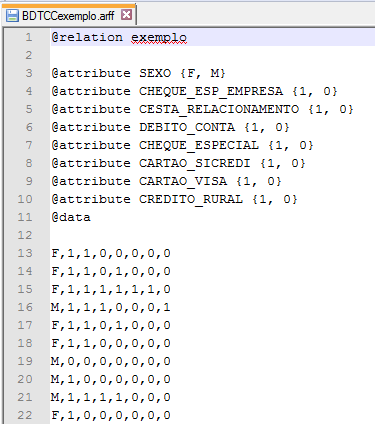
Conforme mostra a figura 07 apresenta um arquivo extraído de uma base de dados, não está na formatação que a ferramenta precisa, neste arquivo contem uma coluna com a delimitação de sexo, M, masculino ou F para feminino, cheque especial empresa, cesta de relacionamento, debito em conta, cheque especial, cartão Sicredi, cartão visa, credito rural, para estas colunas são estabelecidos valores de 1 para quem possui ou 0 para quem não possui o serviço contratado, estando somente aberto como texto simples, já na figura 08 mostra o conjunto formado para que a ferramenta possa executar a Mineração de Dados.

Figura 07: representação dos dados



Fonte: Próprio autor

Figura 08: Dados formatados extensão .arff



Fonte: Próprio autor

A figura XXX representa o arquivo convertido para extenção .arff, onde no inicio é apresentado o nome para o prjeto exemplo, delimitado cada atributo para cada coluna do arquivo orginal, entre colchetes aparece o valor que cada atributo pode conter.

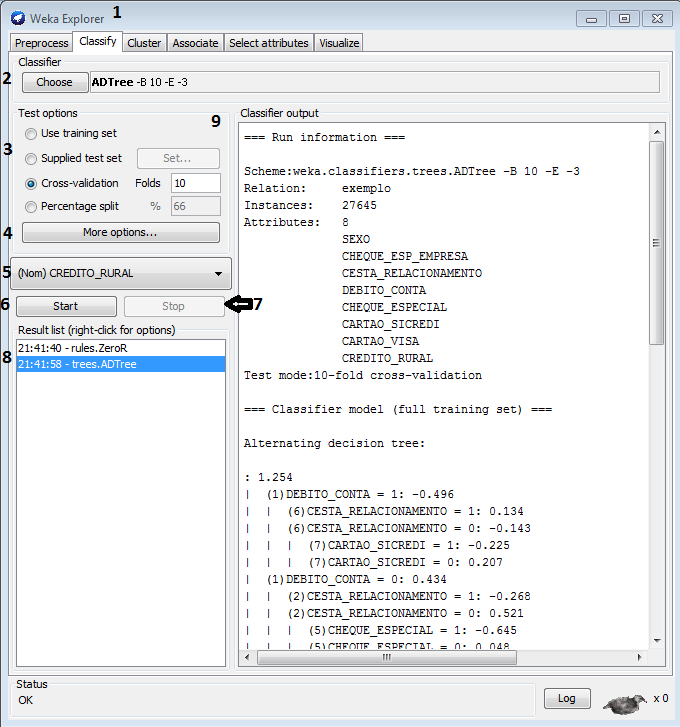
E após delimitado o @data, vem a base a ser minerada, com toadas as informação

**3.1.2 Classificação (Classify)**

Na figura 09 podemos ver a tela da aba Classify, onde permita que o usuário possa classificar os dados processados pelo WEKA, indicado pelo número 2, o usuário pode definir o algoritmo de classificação que será utilizado.

O Test options que é identificado pelo numero (3) na figura XXX define como será feito o teste e tipo de saída após a mineração de dados (4). O número 5 apresenta o atributo de predição que pode ser utilizado. O número 6 indica o botão de início da execução da tarefa, e caso queira parar a execução da tarefa existe o botão de Stop indicado no número 7. O Result List indicado no número 8 mostra os algoritmos utilizados para a mineração de dados, que será visualizado no Classifer Output (9).

Figura 09: Tela Classify



Fonte: Weka 2010

O Classifier, da aba Classify é a parte que dispões de vários algoritmos de classificação, técnicas de aprendizagem de máquina tais como por exemplo árvore de decisão, redes neurais.

Pelo fato do weka ser um framework de código aberto, muitas aplicações utilizam o Weka como base, mesmo ele tendo seu próprio formato de entrada, a extensão

**3.2 CONEXÃO COM UMA BASE DE DADOS**

A ferramenta do WEKA possibilita a conexão com uma base de dados, sem precisar converter arquivos para a extensão .aff. Para que possa ser executado o mesmo, precisamos do driver JDBC, Java Database Connectivity – é uma API – Aplication Programing Interface – que permite que o java conecte com qualquer tipo de Banco de dados.

(não tem uma figura ou exemplo de conexão?)

**3.2 RapidMiner**

A ferramenta de Mineração de RapidMiner é uma ferramenta de código aberta, implementado na plataforma JAVA. Esta ferramenta começou a ser desenvolvida em 2001 na Universidade de Dortmund, na unidade de Inteligência Artificial, sendo sua primeira versão lançada em 2002.

Esta ferramenta funciona em plataformas Windows, Linux e Macintosh, em duas versões:

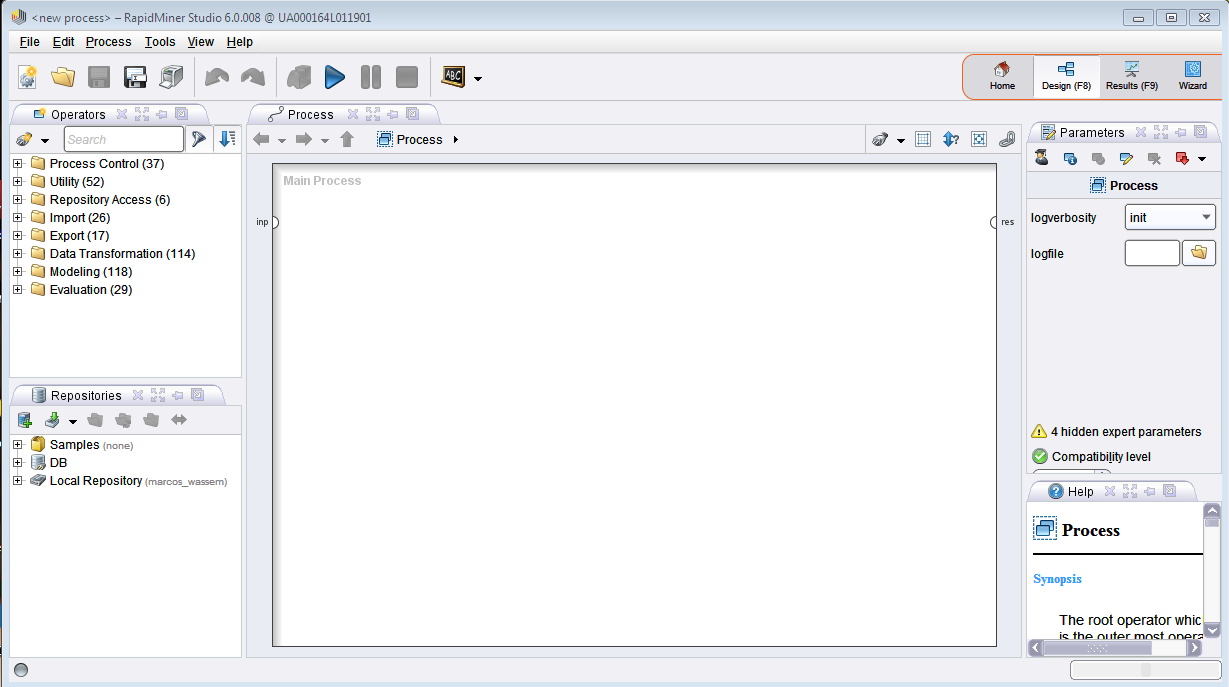
- Comunity Edition: Gratuito, mas com suas funcionalidades reduzidas, mas é muito utilizada em todo o mundo.

- Enterprise Edition: É a versão profissional do software, além de contar todas funcionalidades do Comunity, ainda possui soluções empresariais específicas

**3.2.1 Ambiente Operacional**

Como a ferramenta WEKA, o software RapidMiner oferece vários algoritmos implementados para as fases de pré-processamento. A manipulação de dados se faz de modo transparente, fazendo como tarefa do usuário organizar os mesmos de forma coerente, em forma de uma árvore de processamento, do qual a plataforma executa de forma descendente e sequencial. Na figura 10 pode ser visto o ambiente da ferramenta.

Figura 10 – Ambiente operacional



Fonte: RapidMiner

A ferramenta disponibiliza vários recursos, como por exemplo: análise inteligente dos dados, seleção de atributos, comando de utilização da ferramenta OLAP que é responsável pela geração dos gráficos para melhor entendimento das análises, técnicas e algoritmos para classificação, análise inteligente de dados.

**3.2.2 LEITURA DOS DADOS**

A leitura dos dados é feita através do operador ‘Read CSV’, estrutura de dados que precisa estar separada por caracter separador. O arquivo precisa estar na extenção .CSV e pelo menos com 2 colunas para que seja feito a análise.

**3.2.3 FUNÇÕES DISPONÍVEIS**

As principais funções de grupos operadores encontram-se no lado esquerdo da interface, são elas:

- *Import:* possui os operadores para leitura da base de dados, como por exemplo arquivos .CSV ou conexões direto a base de dados pelo JDBC.

- *Export:* utilizado na escrita dos dados, manipula praticamente os mesmos operadores da leitura.

*- Data Transformation:* em muitas situações é necessário a conversão de dados, filtragem, limpeza antes de ser aplicado algum algoritmo de aprendizagem. Neste grupo se encontram os operadores desta tarefa.

*-* *Modeling:* este grupo possui os algoritmos de aprendizagem da máquina, como geração de regras de produção ou classificação.

*- Text Processing:* neste grupo se encontram operadores especiais para a manipulação de texto, na maioria dos casos são operadores de pré-processamento de texto.

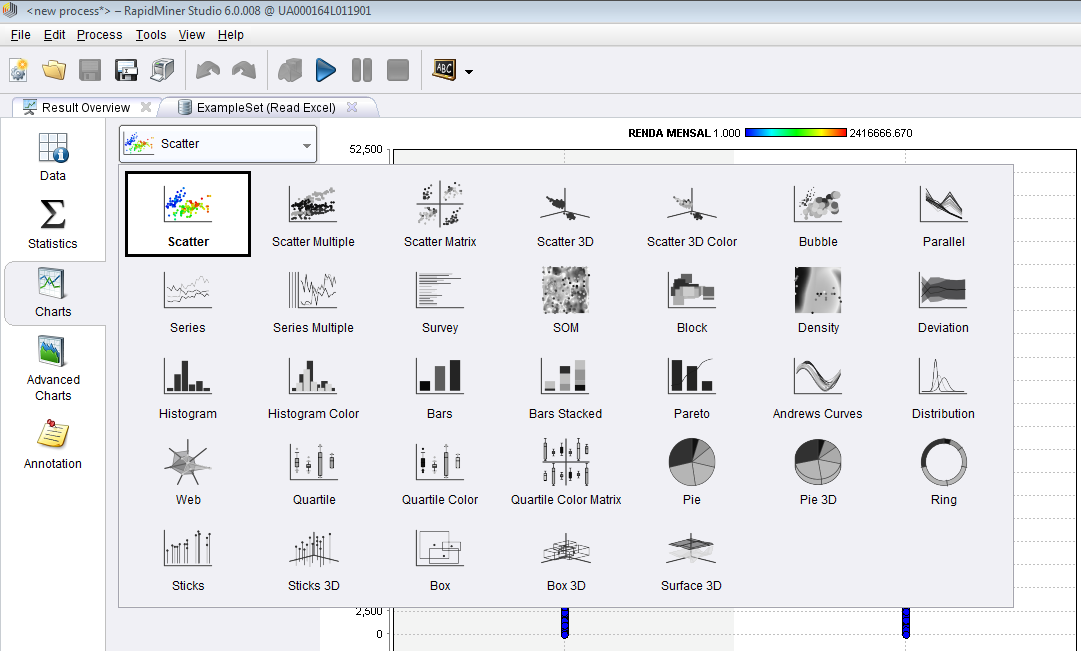
*- Web Minig:* este grupo possui operadores específicos para projetos WebMinig.

Para melhor entendimento da ferramenta, foi analisado uma pequena quantidade de informações na ferramenta do RapidMiner, tendo em vista que para a versão grátis não podemos ter uma base de dados muito grande.

Para o teste, não foi incluso nenhum operador na ferramenta, apenas o importe de um arquivo .xlsx, que é uma planilha Excel.

Após os dados serem importados, o RapidMiner traz uma grande quantidade de gráficos, como demonstrados na figura XXX que podem ser utilizado para a melhor visualização das informações processadas.

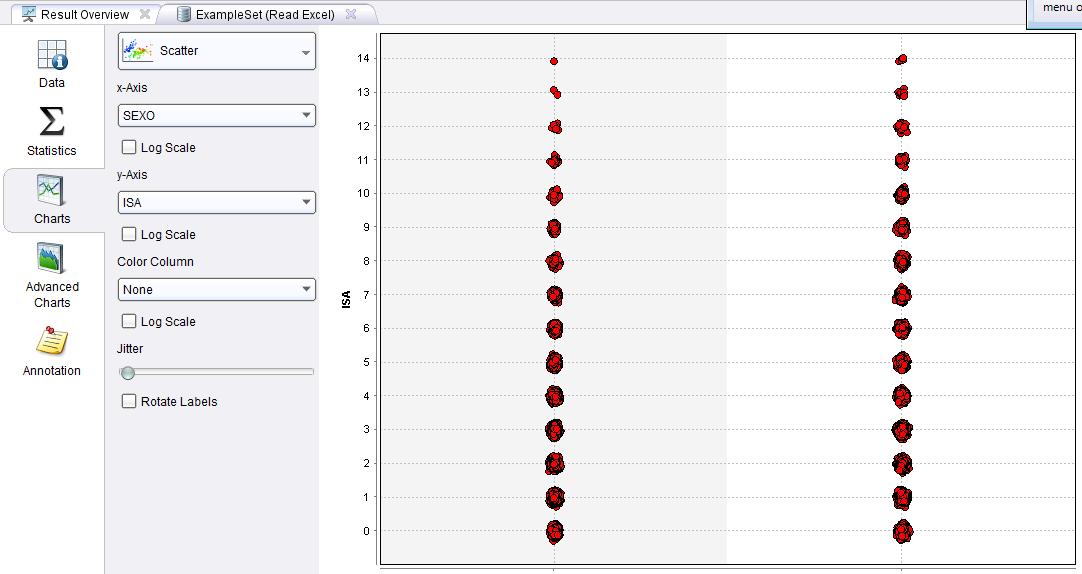
Figura06: Seleção de gráfico



Fonte Rapid Miner

Na análise realizada, foi seleciona o Sexo em relação ao ISA, Quantida de Soluções por Associado, e observa-se que pessoas do sexo masculino tem mais serviços contratados do que as do sexo feminino, como demonstrado na figura XXX

FiguraXXX: Resultado obtido



Fonte: RapidMiner

**3.3 COMPARAÇÃO ENTRE AS FERRAMENTAS**

Foram analisados várias ferramentas, e neste trabalho foram analisadas duas em específico, o RapidMiner e o WEKA. Foram encontrados muitos artigos e trabalhos relacionados a estas duas ferramentas, como o trabalho realizado por LEMOS (2004), que faz a análise de crédito bancário através de redes neurais, e árvore de decisão, fazendo a análise de todas as técnicas de mineração e dados, e para aplicação do algoritmo utilizou a ferramenta do WEKA.

O trabalho realizado por Kampff (2008), Mineração de dados educacionais para a construção de alertas em ambientes virtuais de aprendizagem como apoio à prática docente, faz a análise de alunos que fazem curso a distância, e buscou identificar alunos que tenham perfil para reprovação ou evasão do curso, fez a análise de dados e aplicou as regras utilizando o software RapidMiner.

Devido à grande quantidade de artigos, trabalhos realizados e a facilidade de utilização, foi escolhido a ferramenta do WEKA para busca de conhecimento, utilizando as técnicas e algoritmos de mineração de dados.

**4 PROJETO PARA SOLUÇÃO DA PROPOSTA**

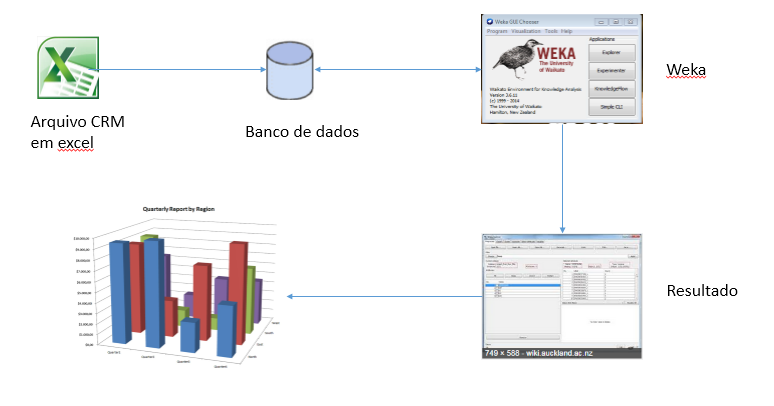
Neste capítulo serão apresentados os artefatos, documentos e técnicas que serão utilizados na implementação da solução proposta, e definições das tecnologias que serão utilizadas para a implementação do trabalho.

(talvez fosse interessante um ou dois parágrafos falando que a metodologia do trabalho é baseada em estudo de caso, análise exploratória de dado, tem isso em livros de metodologia, seria interessante fazer uma introdução)

Na figura 13 é apresentado os passos para o desenvolvimento do trabalho, que consiste na análise de uma base de dados de um arquivo Excel, transformado para extensão .arff para que o WEKA possa fazer a análise do mesmo, com os resultados obtidos, os mesmos são utilizados em uma planilha de Excel a fim de uma melhor visualização dos resultados, em forma de gráficos

(vai ser convertido para arff ou vais acessar direto a base a partir do weka? A segunda alternativa talvez seja melhor para evitar problemas de memória, que são comuns no weka ou podes indicar que serão feitos testes com ambas as abordagens...é que considerando que vais jogar dados para uma base, tem sentido conectar com o weka, senão poderias gerar direto do excel ou algo assim)

Figura 13: Técnicas Mineração de dados



Fonte: Próprio autor

Um fator muito importante para a realização deste preposto, é a identificação de quais informações utilizar, devido à grande quantidade de base de dados, optou-se pela utilização de uma base de entrada via arquivo .CSV. Com isto se tem uma grande independência do banco de dados. E outro fator que pode ser considerado, e pelo fato do arquivo .CSV ser considerado texto limpo, que possui pouco codificação em relação a outras estruturas.

(o csv seria conversão do excel? Seria para importar para o BD?)

**4.1 ANÁLISE DOS DADOS DA EMPRESA.**

Analisando a estrutura de dados utilizado pela cooperativa, constatamos ser somente necessário a utilização de apenas um arquivo do CRM, que contém a quantidade de produtos que cada associado possui, e descriminado quais os produtos que o mesmo utiliza.

Uma das primeira etapas para o desenvolvimento do projeto é o de selecionar os dados que realmente são úteis, na qual se objetiva minerar os dados em busca de um padrão de associação.

Para o correto funcionamento da ferramenta WEKA, é necessário que os arquivos estejam em tabelas, ou qualquer outro arquivo que possibilite ele ser exportado para a extenção .CSV¹

*¹CSV: é um formato de arquivo que putiliza vírgula para separar os valores em uma linha, e quebra de linha para separar as linhas, e aspas duplas para valores que possuem virgula.*

**4.1.1 ESTRUTURA TABELA**

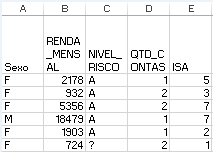
Nesta seção vamos ver a estrutura de dados da tabela que será executado os algoritmos de mineração de dados. Este arquivo obedece a seguinte ordem.

- cada linha refere-se a um cliente.

- cada coluna descreve se o cliente possui ou não determinado produto.

Na tabela 12 apresenta-se parte da estrutura da tabela que será examinada.

Figura 12: Representação tabular dos dados

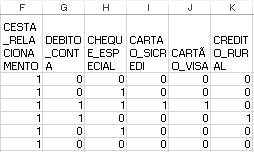


Fonte: Próprio autor

No exemplo proposto na tabela 12 podemos ver a estrutura de dados, representados pelas suas colunas e respectivos valores.

Coluna A, somente pode ter dois valores, M ou F, referindo-se ao sexo masculino ou feminino, a coluna B refere-se ao valor de renda mensal de cada cliente, a coluna C, consiste no nível de risco para empréstimos, sendo letras de A até G, onde a letra A representa cliente com risco baixo de não pagar alguma dívida, e a letra G representa clientes com risco alto para empréstimo, a coluna D representa a quantidade de contas que cada cliente possui, normalmente representados pelos números 1 ou 2, a coluna E é o ISA, índice de solução por associa, onde os valores descritos é a quantidade de produtos ou serviços que o associado tem contratado.

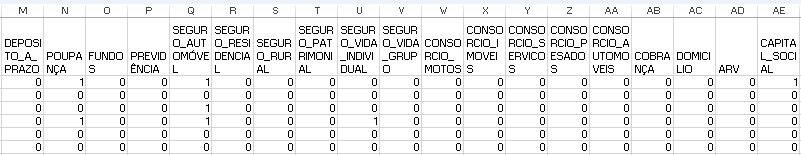
Figura 13: Representação tabular dos dados



Fonte próprio autor

Coluna F representa se possui ou não cesta de relacionamento, que consiste no valor pago pelo cliente por mês para manutenção da sua conta, a colunas G representa debito\_conta, valores que podem variar, dependendo da quantidade de débitos em conta que cada cliente possui, nesta base de dados o maior valor é treze, a coluna H representa cheque\_especial, a coluna I representa cartão\_sicredi, a coluna J representa cartão\_visa, a coluna k representa credito\_rural, sendo que para estas colunas o número 1 representa que possui o serviço contratado e 0 não possui.

Figura 14: Representação tabular dos dados



A coluna L representa credito\_geral, a coluna M representa deposito\_a\_prazo, a coluna N representa poupança, a coluna O representa fundos, a coluna P representa previdência, a coluna Q representa seguro\_automóvel, a coluna R representa seguro\_residencial, a coluna S representa seguro\_rural,a coluna T representa seguro\_patrimonial, a coluna U representa seguro\_vida\_individual, a coluna V representa seguro\_vida\_grupo, a coluna W representa consorcio\_motos, a coluna X representa consorcio\_imóveis, a coluna Y representa consórcio\_servicos, a coluna Z representa consorcio\_pesados, a coluna AA representa consórcio\_automóveis, a coluna AB representa cobrança, a coluna AC representa domicílio, a coluna AD representa arv, a coluna AE representa capital\_social, sendo que para estas colunas o número 1 representa que possui o serviço contratado e 0 não possui.

Nas células que aparece o ?, representa um erro de cadastro, estando esta célula em branco.

**4.2 PRÉ PROCESSAMENTO DOS DADOS**

Esta etapa, como mencionado na seção 2.1.1.1, é utilizada para efetuar a limpeza dos dados antes do processamento que utiliza a técnica para a descoberta de conhecimento em uma base de dados. Neste sentido está técnica envolve maior análise para a verificação de registros incorretos.

Neste preposto foi avaliado uma base de dados com 31 colunas e mais de 40 mil linhas. Fazendo a verificação destes dados, foi verificado algumas inconformidades, como por exemplo dados que não estão preenchidos, ou cadastrados de forma incorreta. Para evitar análises incorretas, nos valores que estavam em branco, foi colocado o símbolo ?, para que o WEKA não tomasse como base este resultado.

**4.3 FERRAMENTA EXPLORATÓRIA**

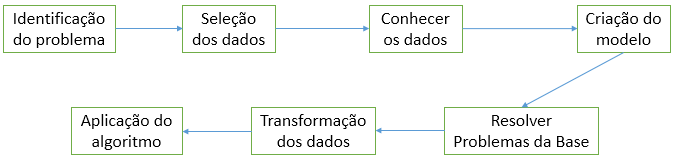
**(título estranho. Esta parte tem muitas referências a autores, normalmente a fundamentação vai no capítulo 2 e aqui descreves o que tu vais fazer, não os fundamentos do autor... esta parte do método pode ir antes, como comentei...tenta separar o que é fundamentação para colocar no capítulo dois, pode ser com aquela figura original do Berry e aqui colocas a tua figura adaptada com textos elaborados por ti.)**

Neste capítulo será abordado o método utilizado na realização deste ensaio, segundo Lakatos e Marconi(2001), métodos, é o conjunto das atividades sistemáticas racionais que, com maior segurança, permite alcançar objetivos propostos inicialmente, traçando o caminho a ser seguido, detectando erros e auxiliando as decisões cientistas.

Para este estudo será executado uma análise exploratória dos dados, utilizando métodos estáticos para análise dos mesmos, a fim de identificar aplicações e técnicas capazes de auxiliar na gestão de conhecimento.

As etapas processos e metodologia da mineração de dados são descritos em seis itens, ilustrados na figura 11, que serão descritos nos capítulos seguintes.

Figura 11: Etapas do processamento de dados



Fonte: Próprio autor

**4.3.1 IDENTIFICAÇÃO DO PROBLEMA DO NEGÓCIO**

Uma das primeiras fazes para transformar um problema comum em um problema para Mineração de Dados, é saber qual a melhor técnica deve ser aplicada para a análise (classificação, estimativa, previsão, análise de afinidade, análise de agrupamento ou sumarização).

De acordo com Berry(2004), no momento da decisão de qual técnica utilizar, é ter presente para que os resultados serão utilizados.

**4.3.2 SELEÇÃO DOS DADOS APROPRIADOS**

Para Berry (2004), muitas vezes os resultados podem ser afetados por estar utilizando base de dados desatualizada, ou possuir mais de uma base de dados, não tendo informações centralizadas.

Para a seleção de dados, deve-se levar em conta que novos dados não podem substituir dados antigos, e sim somente serem inseridos na base de dados, assim obtendo resultados coerentes cada vez que a técnica é aplicada.

**4.3.3 Conhecer os dados**

Para Berry(2004), a pessoa que for responsável pela mineração de dados, deve conhecer a base analisada, tanto em nível estrutural quanto em nível lógico, assim evitando que haja resultados não satisfatórios.

**4.3.4 CRIAÇÃO DO MODELO**

De acordo com Berry(2004), este modelo é utilizado em todas as etapas da mineração de dados, pois se cria a estrutura de dados que será utilizado num todo. Neste modelo estuda-se a possibilidade de criar alguns atributos específicos, que possa resultar em combinações na base de dados estudada.

Por exemplo, para buscar a idade média dos seus clientes, para isto é preciso ter uma entrada no sistema a data de nascimento, para que possa ser mensurado a idade média dos seus clientes.

**4.3.5 RESOLVER PROBLEMAS DOS DADOS**

Para Berry(2004), todos os dados possuem problemas, podem ser coletados em bases diferentes e terem valores diferentes. Por esta razão deve-se reparar todos estes problemas antes de executar a Mineração de Dados, a fim de evitar qualquer resultado que não seja real.

**4.3.6 TRANSFORMAÇÃO DOS DADOS PARA CONTRUÇÃO DE INFORMAÇÕES PARA MINERAÇÃO DE DADOS**

Segundo Berry (2004), neste processo é feito a análise dos atributos a serem analisados, se os atributos estiverem corretos não é necessário fazer nenhuma alteração quantos ao mesmos. Como atributos passiveis para alteração, podemos citar como exemplo a idade que os clientes possuem, onde na base de dados temos a data de nascimento e o dia atual, então precisamos criar mais um campo com a subtração da data de nascimento com a data atual. Assim pode-se criar um campo com idade total antes de executar a mineração de dados.

**4.3.7 CONSTRUÇÃO DO MODELO**

Para Berry(2004), nesta etapa iremos utilizar as técnicas descritas no capítulo 2.2, a fim de obter resultados satisfatórios após a execução da Mineração de Dados

**5 CRONOGRAMA**

O trabalho será desenvolvido no segundo semestre de 2014. A seguir é apresentada a tabela demonstrando o cronograma de cada etapa do desenvolvimento.

Tabela 1 - Cronograma de desenvolvimento da proposta

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tarefa** | **Jul/14** | **Ago/14** | **Set/14** | **Out/14** | **Nov/14** | **Dez/14** |
| Ajustes na revisão bibliográfica | X |  |  |  |  |  |
| Ajustes no projeto da ferramenta | X |  |  |  |  |  |
| Estudo dos dados a ser analisado |  | X |  |  |  |  |
| Técnicas de Mineração de dados |  | X | X |  |  |  |
| Análise do dados |  |  | X | X |  |  |
| Documentar resultados |  | X | X | X |  |  |
| Analisar e documentar os resultados |  |  |  |  | X |  |
| Finalizar e ajustar a monografia |  |  |  |  |  | X |

Fonte: Elaborado pelo autor.

**6 CONSIDERAÇÕES FINAIS**

O estudo realizado com base em diversas fontes bibliográficas evidenciou a oportunidade da instituição financeira em adquirir novos conhecimentos sobre seus clientes, a fim de melhorar suas análises negociais, através da grande quantidade de informações existentes em vários arquivos diferentes.

No referencial teórico foram apresentados vários casos onde a aplicação desta base de conhecimentos é capaz de trazer bons resultados, estes capazes de auxiliar os gestores nas tomadas de decisões.

Ao propor a análise dos vários algoritmos e técnicas de Mineração de Dados, e aquisição de conhecimento, o autor busca apresentar as melhores soluções para a busca de determinadas soluções de Mineração de Dados.

(melhorar consideravelmente as considerações finais, tem pouca coisas...)

**REFERÊNCIAS**

CARVALHO, L. A. V. **Data Mining**: A Mineração de dados no Marketing. Rio de Janeiro: Editora Ciência Moderna, 2005.

FAYYAD, USAMA; PIATETSKY-SHAPIRO, GREGORY; SMYTH, PADHRAIC. **From Data Mining to Knowledge Discovery: An Overview.** American Association for Artificial Intelligence, 1996.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Mining:** um guia prático.Rio de Janeiro/RJ: Campus, 2005.

HEUSER, Carlos A. **Projeto de Banco de Dados**. 6. ed. Porto Alegre: Artmed, 1998.

SWIFT, Ronald. **CRM Customer Relationship Management**. 13. ed. Rio de Janeiro/RJ: Elsevier, 2001.

HAN, J; KAMBER, M. **Data Mining: Concepts and Techniques.** Elsevier, 2006.

WITTEN, I. H.; FRANK, E.; 2000. Data Mining: Practical Machine Learning Tools and

Techniques with Java Implementations. USA: Morgan Kaufmann Publishers

AGRAWAL, R.; SRIKANT, R.; 1994. Fast Algorithms for Mining Association Rules. In

PROCEEDINGS OF THE 20TH INTERNATIONAL CONFERENCE ON VERY LARGE

DATABASES (1994: Santiago, Chile).

FREITAS, O. G.; et al.; 2001. Sistema de Apoio à Decisão usando a Tecnologia Data Mining com

Estudo de Caso da Universidade Estadual de Maringá. In I CONGRESSO BRASILEIRO DE

COMPUTAÇÃO – CBComp 2001. Anais

MALHORTA, Naresh K, Pesquisa de Marketing: Uma orientação Aplicada.3. ed. Porto Alegre, Bookman 2001

DIAS, Maria M. Um Modelo de Formalização do Processo de Desenvolvimento de

Sistemas de Descoberta de Conhecimento em Banco de Dados. 2001. 212 f. Tese

(Doutorado em Engenharia da Produção) – Universidade Federal de Santa Catarina – UFSC,

Florianópolis, 2001.

GOLDSCHMIDT, R.R.; PASSOS, E. Data Mining: Um guia prático, conceitos, técnicas,

ferramentas, orientações e aplicações. São Paulo: Elselvier 2005.

SCOSS, Anne M.. A Clusterização e Classificação no Processo De Data Mining para

Análise do Desempenho Docente no Ensino de Graduação. 2006. 86 f. Trabalho de

Conclusão de Curso (Especialização) - Universidade do Extremo Sul Catarinense - UNESC, Criciúma, 2006.

CUNICO, Luiz H.B.. Técnicas em Data Mining Aplicadas na Predição de Satisfação de Funcionários de uma Rede de Lojas do Comércio Varejista. Curitiba, 2005. Tese de Mestrado, Setores de Tecnologia e de Ciências Exatas, Universidade Federal do Paraná

LEMOS, Eliane P.. Análise de Crédito Bancário com uso de Data Mining: Redes Neurais e Arvore de Decisão. Curitiba 2003. Tese de Mestrado, Setores da Tecnologia, Departamento da Construção Civil e de Ciências Exatas, Universidade Feral do Paraná.

PADUA, A. et al. Redes Neurais Artificiais-Teoria e Aplicações. Rio de Janeiro, LTC,

2000.

HALKIDI, Maria; BATISTAKIS, Yannis; VAZIRGIANNIS, Michalis. On clustering

validation techniques. Journal of Intelligent Information Systems, v. 17, n. 2-3, p. 107-145,

Dec. 2001.

Mineração de dados educacionais para a construção de alertas em ambientes virtuais de aprendizagem como apoio à prática docente Adriana Justin Cerveira Kampff - akampff@gmail.com 1,2 Eliseo Berni Reategui - eliseoreategui@gmail.com 1 José Valdeni de Lima - valdeni@inf.ufrgs.br 1 1 PGIE – UFRGS 2 ULBRA – Tecnologia e Computação / EAD dezembro de 2008