

Roberto de Souza Santos

**APLICAÇÃO DE UM MODELO PREDITIVO DE
MINERAÇÃO DE DADOS PARA APOIO À
DECISÃO DE CRÉDITO**

Belo Horizonte
Universidade Federal de Minas Gerais
2006

Roberto de Souza Santos

**APLICAÇÃO DE UM MODELO PREDITIVO DE
MINERAÇÃO DE DADOS PARA APOIO À
DECISÃO DE CRÉDITO**

Dissertação apresentada ao Programa de
Pós-Graduação em Ciência da Informação
da Universidade Federal de Minas Gerais,
como requisito parcial à obtenção do título
de Mestre em Ciência da Informação.

Orientador: Prof. Dr. Marcello Peixoto
Bax

Belo Horizonte
Universidade Federal de Minas Gerais
2006

Dedico este trabalho ao meu querido avô Henrique.

AGRADECIMENTOS

Agradeço aqui a alguns, entre tantos outros que de diversas formas me apoiaram durante a realização deste trabalho.

Ao Prof. Dr. Marcello Peixoto Bax, por acreditar que poderíamos fazer um bom trabalho.

A todos os colegas e amigos do Mercantil do Brasil, em especial à Marcela Tavares Fidelis pela sua ajuda.

Aos vários professores e funcionários da Escola de Ciência da Informação da UFMG.

A tantos colegas do curso, como a Sandra, Enilda, Maria Lúcia, Frederico, Paulo, Delfim, entre tantos outros que terão sempre a minha amizade e consideração.

Aos meus pais Evando e Cristina, pela formação e pelas portas que me abriram.

À Viviane e ao Marco Aurélio pelo incentivo, pelas revisões e diversas formas de ajuda.

À Tereza, Antonio e à Rúbia pelo constante apoio.

À minha amada esposa Caroline pelo amor, pela compreensão, pela dedicação, pelas noites e finais de semana ocupados.

RESUMO

Este estudo busca compreender a contribuição da descoberta de conhecimento em bases de dados ao suporte à decisão de concessão de crédito. Para tanto, tem como objetivo principal a criação de um modelo preditivo capaz de identificar informações utilizando mineração de dados, bem como explorar as possibilidades de uso de tal modelo. São seus objetivos específicos explorar como se planeja e se realiza cada uma das etapas e atividades necessárias para a criação de um modelo preditivo e implementá-lo através de uma metodologia para mineração de dados. Buscou-se também compreender como o uso do modelo se encaixa no âmbito de processos mais amplos de descoberta de conhecimento sobre o risco de crédito em instituições financeiras. Identificou-se um processo de conhecimento em que os dados históricos representam a memória organizacional, a mineração de dados representa o processo de aprendizagem e o conhecimento extraído é um novo conhecimento de domínio. Conclui-se que o uso de mineração de dados, que se concentra nos processos de extração e aplicação da descoberta de conhecimento em bases de dados, apoiada nos conceitos e definições para escoragem de crédito pode auxiliar no processo de decisão de crédito.

Palavras-chave: informação; conhecimento; descoberta de conhecimento em bases de dados; mineração de dados; escoragem de crédito.

ABSTRACT

The present study searches the understanding of the contribution of the knowledge database discovery on supporting the credit decision. For this purpose, its general objective is to create a predictive model capable of identifying information through data mining, as well as to explore how the model can be used. The study's specific objectives are to explore how to plan and execute each one of the tasks required to create the predictive model and to implement it using a data mining methodology. This study also seeks to understand the role of this model on a broader knowledge discovery process over the credit risk at financial institutions. A knowledge process was recognized in which the historical data represents organizational memory, the data mining represents the learning process and the knowledge extracted is the new domain knowledge. The use of data mining, which focuses on the extraction and application processes of knowledge database discovery, can support the credit decision when guided by the credit scoring definitions.

Keywords: information; knowledge, knowledge discovery in databases, credit score

LISTA DE FIGURAS

Figura 1 – Fases do Modelo CRISP-DM	20
Figura 2 – A espiral do conhecimento.....	24
Figura 3 – Arquitetura de um modelo para descoberta de conhecimento.	28
Figura 4 – Etapas da DCBD	34
Figura 5 – Etapas do processo de MD.....	35
Figura 6 – Atividades e tarefas de MD.....	37
Figura 7 – Exemplo de visualização de regras de associação	38
Figura 8 – Exemplo de um agrupamento	39
Figura 9 – Exemplo de árvore de decisão.....	41
Figura 10 – Representação de uma análise discriminante.....	55
Figura 11 – Curvas características da distribuição normal (a) e logística (b)	56
Figura 12 – Comportamento médio das operações nas faixas de inadimplência	69
Figura 13 – Distribuição da taxa juros/ativo por faixas de risco	70
Figura 14 – Distribuição diária da taxa juros/ativo	71
Figura 15 – Carteira de operações por tipo de cliente	73
Figura 16 – Distribuição de clientes por estado civil	74
Figura 17 – Cubo OLAP criado e exemplo de consulta disponível	76
Figura 18 – Interface para criação da árvore de decisão	77
Figura 19 – Comandos utilizados para na criação e no treinamento do modelo	78
Figura 20 – Visualização da árvore de decisão	78
Figura 21 – Ferramenta para criação de pacotes DTS.....	80
Figura 22 – Pacote DTS criado para a predição	81
Figura 23 – Consulta à tabela com o atributo previsto	82
Figura 24 – Fluxo operacional com a adição do conhecimento descoberto.....	83

LISTA DE QUADROS

Quadro 1 – Exemplo de escore para definição do ponto de corte	59
Quadro 2 – Relação entre as metodologias para DCBD, MD e EC	62
Quadro 3 – Classes de risco para provisionamento segundo a CMN 2682.....	68
Quadro 4 – Distribuição da carteira em relação às faixas de risco.....	69
Quadro 5 – Critério para a classificação dos clientes.....	72
Quadro 6 – Distribuição dos empréstimos por tipo de cliente	72
Quadro 7 – Distribuição dos clientes por sexo	73
Quadro 8 – Resumo das definições para a implementação da MD.....	84

LISTA DE SIGLAS E ABREVIATURAS

BIC - Bureau de Informação de Crédito

CI - Ciência da Informação

CRISP-DM - Cross-Industry Standard Process for Data Mining

DCBD - Descoberta de Conhecimento em Bases de Dados

DCC - Departamento de Ciência da Computação

DTS - Data Transformation Services

EC - Escoragem de Crédito

ECI - Escola de Ciência da Informação

KDD - Knowledge Discovery in Databases

MD - Mineração de Dados

OLAP - Online Analytical Processing

SQL - Structure Query Language

UFMG - Universidade Federal de Minas Gerais

SUMÁRIO

1.	Introdução.....	11
1.1	Problema e justificativa	14
1.2	Objetivos.....	16
1.3	Classificação da Pesquisa	18
1.4	Metodologia.....	18
1.5	Organização do trabalho	21
2.	Referencial teórico.....	22
2.1.	Definições e conceitos	22
2.1.1.	Dado, informação e conhecimento	22
2.1.2.	Regras organizacionais para tomada de decisão.....	25
2.2.	Descoberta de conhecimento em bases de dados	26
2.2.1.	A natureza interdisciplinar da DCBD.....	29
2.3.	Mineração de dados	32
2.3.1.	Tarefas de mineração de dados.....	35
2.3.1.1.	Regras de Associação	37
2.3.1.2.	Agrupamento	39
2.3.1.3.	Classificação	39
2.3.1.4.	Árvore de Decisão	40
2.4.	Concessão de crédito e escoragem de crédito	41
2.4.1.	Métodos subjetivos para concessão de crédito	42
2.4.2.	O uso de modelos de escoragem de crédito.....	43
2.4.3.	Desenvolvimento de um sistema de escoragem de crédito	46
2.4.3.1.	Planejamento e definições	47
2.4.3.2.	Identificação das variáveis potenciais	49
2.4.3.3.	Planejamento amostral e coleta de dados	51
2.4.3.4.	Determinação da fórmula de escoragem	53
2.4.3.5.	Determinação do ponto de corte.....	57
2.4.3.6.	Determinação das regras de decisão.....	60
2.4.3.7.	Validação do sistema de escoragem de crédito	61
2.5.	Síntese do referencial teórico	61
3.	Criação do modelo de mineração de dados	65
3.1.	Caracterização do ambiente pesquisado	65
3.2.	Desenvolvimento do modelo	66
3.2.1.	Construção da árvore de decisão	75
3.2.2.	Predição da inadimplência dos proponentes.....	79
4.	Conclusões.....	85
5.	Referências Bibliográficas.....	89

1. Introdução

O mercado de crédito brasileiro sofreu muitas mudanças na última década após o plano real. Conforme apontam PINHEIRO & MOURA (2001), parte da atual ineficiência dos mercados de crédito se deve aos poucos incentivos para o investimento em instrumentos de análise de crédito para concessões de médio e longo prazo para empresas e indivíduos até o início do plano real em 1994. Até esta data, cerca de 50% dos lucros dos bancos era proveniente de receitas com operações de *floating*¹ e o crédito era tradicionalmente concedido por bancos estatais. Isto levava os bancos privados a concederem pouquíssimos empréstimos, concentrando tais atividades em operações de curto prazo como cheques especiais e financiamento com capital de giro, garantidos pelos fluxos de caixa dos tomadores que eram gerenciados pelas mesmas instituições. A decisão de crédito era julgamental, baseada em fatores subjetivos e na experiência do analista de crédito.

Com a redução nas taxas de inflação após o Plano Real, reduziu-se a receita com *floating*, a economia se reaqueceu e deu-se um aumento da demanda por crédito. Entretanto, a redução das receitas com *floating* e a falta de mecanismos adequados para mensurar o risco de crédito trouxeram sérios problemas de sobrevivência para os credores, levando à extinção ou absorção de vários deles.

A explosão da oferta de crédito ocorreu sem que tivesse havido qualquer mudança essencial nas práticas de concessão de crédito, que continuavam a utilizar métodos antiquados para a seleção dos tomadores. [...] Assim, não foi surpresa quando os índices de inadimplência aumentaram dramaticamente nesse segmento do mercado, levando à falência credores que apenas há alguns meses exibiam recordes de vendas e de receitas de juros. [...] Os bancos também não ficaram imunes: tanto os públicos quanto os privados – nacionais e estrangeiros – testemunharam um aumento das taxas de inadimplência nos seus empréstimos para pessoas físicas. (PINHEIRO E MOURA, 2001, p.68).

¹ *Floating* - aplicação feita pelos bancos no período de hiperinflação no Brasil. Dentro das 72 horas em que os depósitos levavam para entrar na conta dos clientes, os bancos se utilizavam desses valores para aplicações no mercado especulativo, ganhando dinheiro e devolvendo ao cliente seu depósito original acrescido de uma parcela de lucros menor do que conseguiam no mercado.

O desenvolvimento dos mecanismos de apoio à decisão sobre o crédito é importante para permitir a expansão das decisões de gasto privado e para o investimento produtivo, mas segundo relatório do Ministério da Fazenda no Brasil os indicadores de volume de crédito, custo e os instrumentos de financiamento existentes ainda estão bem aquém do que seria adequado para o nível de desenvolvimento da nossa economia. (DUARTE et al, 2004).

Lourenço (2005) afirma que o segmento varejista de crédito ao consumidor tem como características “o grande volume de transações, o baixo valor unitário por transação, *spread*² elevado e a necessidade de velocidade na decisão”.

Segundo GALINDO e MILLER (2001), em muitos países o mercado financeiro passou por um período de fusões e consolidações, onde ocorre a perda do conhecimento sobre os clientes na absorção das instituições menores pelas maiores, que freqüentemente centralizam o processo de crédito. Estes fatores, entre outros, aumentam a importância dos modelos de extração de informações e geração de conhecimento a partir dos registros de dados sobre as operações de crédito.

Borko (1968) aponta na Ciência da Informação (CI) não apenas componentes de ciência pura, que estuda um assunto sem considerar sua aplicação, mas também componentes de uma ciência aplicada, que desenvolve serviços e produtos. Questões como a geração, interpretação e uso da informação e do conhecimento em sistemas naturais e artificiais devem ser estudados pela CI.

Esta dissertação propõe a pesquisa sobre a construção e a interpretação de informações a partir da mineração de dados (MD), que consiste no conjunto de técnicas e tecnologias que possibilitam a descoberta de novas informações, procurando, em certos casos, “aprender” a

² Diferença entre o preço de compra e de venda de um título ou moeda.

partir de relacionamentos escondidos, padrões³, correlações e interdependências existentes em grandes bases de dados.

³ Padrões – abstrações de um subconjunto dos dados em alguma linguagem descritiva de conceitos.

1.1 Problema e justificativa

Uma das principais dificuldades para a concessão de crédito é a identificação do risco de inadimplência ou não cumprimento dos contratos. Isto decorre da incapacidade em estimar-se adequadamente a probabilidade dos agentes envolvidos em cumprir os compromissos assumidos.

Por vezes, as instituições de crédito não sabem avaliar risco de crédito e se baseiam apenas em listas de crédito negativo mantidas pelos BIC⁴.

Os ambientes de análise de crédito são caracterizados pela dinâmica na tomada de decisões e pela grande variedade de informações vindas das mais diversas fontes. Essas informações muitas vezes podem ser incompletas, ambíguas, parcialmente incorretas ou de relevância dúbia. A forma subjetiva como se dá a análise dessas informações faz com que não se consiga explicar o processo de tomada de decisões embora seja sabido que existem fatores que influenciam essas decisões. (SENGER e CALDAS JUNIOR, 2001, p.19).

Senger e Caldas Junior (2001) afirmam que as decisões de crédito devem ser criteriosas, pois podem provocar prejuízos às instituições financeiras, além de prejuízos morais aos clientes. Quanto mais amplas e precisas as informações sobre a probabilidade de pagamento, melhores serão as condições para a correta avaliação do risco de cada operação. Com isso, as instituições financeiras terão maior propensão a emprestar, os custos dos empréstimos serão menores e todos os agentes econômicos obterão condições de realizar melhores resultados.

As empresas reconhecem a necessidade de extrair informação e conhecimento presentes em suas bases de dados. Métodos manuais são dispendiosos e inviáveis quando aplicados a

⁴ BIC – Bureau de Informação de Crédito. Existem vários no Brasil há várias décadas. Tradicionalmente mantêm essencialmente informações negativas, obtidas dos cartórios judiciais e de títulos, das associações de diretores lojistas e do registro do Banco Central sobre cheques devolvidos. O cadastro apenas de informações negativas desqualifica esta fonte, tornando suas informações pouco úteis no contexto deste trabalho. Entre os maiores BIC do Brasil estão o SERASA e o SPC.

grandes bases de dados. A mineração dos dados registrados sobre os clientes e o histórico das operações realizadas nas empresas é uma alternativa para suprir esta necessidade de informação.

Existem atualmente poucos estudos aprofundados na literatura brasileira sobre as etapas necessárias à criação de um modelo preditivo de MD para apoio à decisão de crédito sob a ótica dos conceitos da Ciência da Informação. Sendo a CI, de acordo com Saracevic (1996), um campo dedicado às pesquisas científicas dos problemas da comunicação do conhecimento e seus registros, encontra-se, no seu contexto, terreno fértil para o estudo dos processos de descoberta de conhecimento em bases de dados (DCBD).

A Ciência da Informação é um campo dedicado às questões científicas e à prática profissional voltada para os problemas da efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, institucional ou individual do uso e das necessidades de informação. No tratamento destas questões são consideradas de particular interesse as vantagens das modernas tecnologias informacionais. (SARACEVIC, 1996, p.7).

Wersig (1993) aponta que a CI necessita de uma compreensão básica dos “atores” e dos processos de criação e transformação do conhecimento. Eles podem ser indivíduos, organizações, culturas e até configurações tecnológicas. Além disso, os modelos e conceitos devem ser confrontados com os processos da realidade para a avaliação de seus potenciais propósitos na CI.

Chen e Liu (2004) afirmam que a MD abriu novos caminhos para a CI, possibilitando a pesquisa por informações valiosas em grandes volumes de dados, assim apoiando as organizações na tomada de decisões mais rápidas e com maior grau de confiança. Segundo os autores, as tarefas de MD têm sido implementadas em muitos domínios de aplicação, podendo apoiar a CI em ambientes personalizados, comércio eletrônico e mecanismos de busca. Portanto, a pesquisa dos métodos de MD revela-se pertinente no contexto dos

processos de decisão de crédito, que, conforme visto anteriormente, têm influência econômica e social sobre o cotidiano.

Saracevic (1996) diz que os estudos sobre sistemas inteligentes, bases de conhecimento, hipertextos e sistemas relacionados, interfaces inteligentes e interação homem-computador e mesmo a reutilização de software têm significativo componente informacional, associado com a representação da informação, sua organização intelectual e encadeamentos; busca e recuperação de informação; qualidade, valor e uso da informação - todos tradicionalmente tratados pela Ciência da Informação.

1.2 Objetivos

O objetivo deste trabalho é criar um modelo preditivo capaz de identificar padrões úteis, i.e., informações, para suporte à decisão utilizando MD, bem como explorar as possibilidades de uso de tal modelo. Pretendeu-se pesquisar as etapas necessárias para a preparação e para a transformação dos dados em informação e geração de conhecimento.

Os seguintes objetivos específicos foram destacados:

- Explorar como se planeja e se realiza cada uma das etapas e atividades necessárias para a criação de um modelo preditivo para a decisão de crédito a partir de dados históricos de operações de crédito realizadas por uma instituição financeira;
- Implementar o modelo preditivo através de uma metodologia para MD.

Na investigação do uso e validação do modelo pretendeu-se verificar, principalmente, se:

- O modelo criado promove ganhos ao processo de decisão, com boa aplicabilidade e bom coeficiente de aderência para este tipo de aplicação, ou seja, se o uso de MD pode apoiar a decisão sobre o risco de crédito.
- Uma vez que se verifique positivamente a utilização de MD, buscar-se-á compreender como o uso do modelo se encaixa no âmbito de processos mais amplos de descoberta de conhecimento sobre o risco de crédito em instituições financeiras.

Em resumo, pretendeu-se investigar a seguinte questão: é possível construir, a partir da MD, um modelo inspirado na escoragem de crédito⁵ que apóie efetivamente a decisão sobre o risco de crédito? Em seguida, busco-se verificar como o modelo de apoio à decisão se relaciona como com os processos de descoberta de conhecimento.

⁵ Escoragem de crédito (EC) – atribuição de pontos aos candidatos ao crédito para classificá-los em grupos de risco. Tradução do termo em inglês *credit scoring*.

1.3 Classificação da Pesquisa

Trata-se de uma pesquisa aplicada que visa identificar elementos para a construção de um modelo de DCBD, utilizando como caso o histórico dos empréstimos pessoais de uma determinada instituição financeira brasileira. O seu nome foi omitido por questões de política de segurança. O universo da pesquisa consiste dos dados registrados no histórico desta instituição e dos usuários e analistas que detêm o conhecimento do domínio necessário à criação do modelo.

1.4 Metodologia

Para atender aos objetivos propostos, as seguintes etapas básicas foram desenvolvidas:

Etapas 1 – Revisão bibliográfica sobre MD e sua contribuição para o ciclo de vida da informação.

Estudou-se o ciclo de vida da informação sob o ponto de vista da Gestão da Informação passando de dado à informação e ao conhecimento e as contribuições das ferramentas de MD apoiando a DCBD e o ciclo de vida da informação. Estudou-se a metodologia para MD CRISP-DM de Chapman et al (1999) e a metodologia proposta por Fayyad et al (1996).

Etapas 2 – Estudo do papel da informação no apoio à decisão da concessão de crédito pessoal e ao entendimento do negócio.

Caracterizou-se o registro e o uso das informações sobre os clientes no contexto da instituição financeira pesquisada e identificaram-se as necessidades de informação para melhor apoiar as

decisões e proporcionar o entendimento do cenário de informação e dos processos de negócio envolvidos. Foram estudados os conceitos e uma metodologia para a criação de modelos de EC para relacionar seus fundamentos e práticas aos objetivos desta pesquisa.

Nesta etapa foram realizadas análises de documentos e informações, além de entrevistas com especialistas do domínio na instituição pesquisada.

Etapas 3 – Construção do modelo de MD

Criou-se um modelo de MD inspirado na EC utilizando ferramentas de *software* para explicitar as informações e o conhecimento contidos nos registros do histórico de transações da instituição financeira. Os trabalhos de Sicsú (1998) e Herszkowicz (1998) foram as principais referências para os conceitos sobre EC.

Foram seguidas as metodologias CRISP-DM, apresentada na figura 1, e a proposta por Fayyad et al (1996) que definem que o processo envolve a realização de várias atividades como:

- Identificação do domínio;
- Seleção e limpeza dos dados;
- Redução do volume de dados;
- Escolha da função de MD;
- Escolha do algoritmo de MD;
- Busca por padrões;
- Avaliação do conhecimento;
- Refinamento do conhecimento

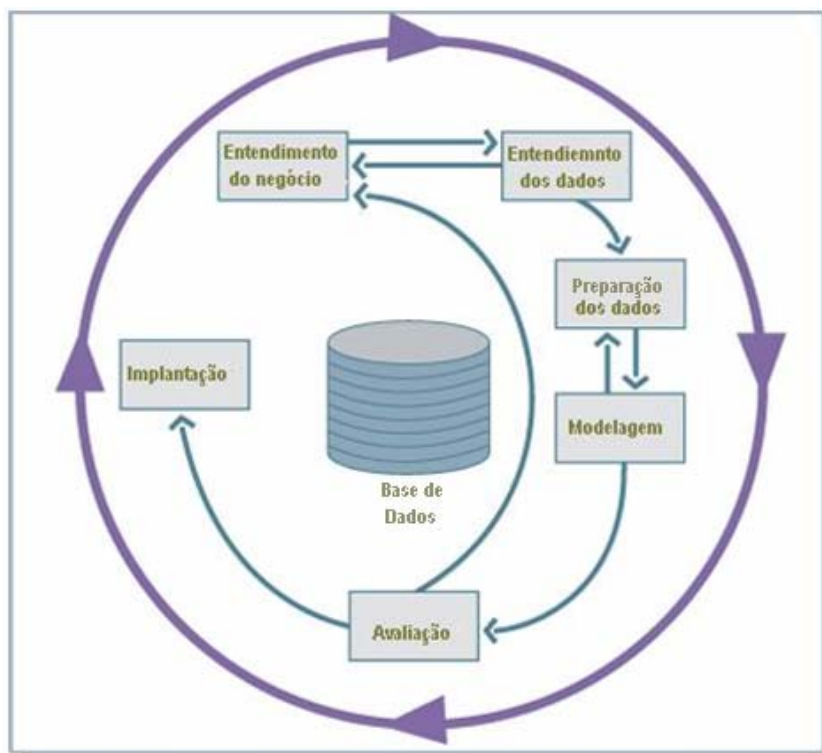


Figura 1 – Fases do Modelo CRISP-DM

Fonte: CHAPMAN, P. et al (1999)

Etapa 4 – Validação do modelo

Foi avaliado o desempenho do modelo quando aplicado a uma base de dados de testes onde o comportamento dos tomadores de empréstimo é conhecido.

Nesta etapa foram feitos testes do modelo de MD e avaliação da precisão das informações geradas. Foi avaliado estatisticamente o grau de acerto do modelo criado comparando o risco previsto com o resultado de operações realizadas.

Etapa 5 – Discussão sobre o modelo.

Discussão e conclusão sobre sua eficiência, sua aplicabilidade e sua viabilidade no suporte à geração de informação.

1.5 Organização do trabalho

Esta dissertação é estruturada em quatro capítulos. A introdução, aqui apresentada no primeiro capítulo, é seguida por um referencial teórico sobre descoberta de conhecimento em bases de dados, mineração de dados, tomada de decisão sobre concessão de crédito e escoragem de crédito, no segundo capítulo. A construção do modelo de mineração de dados é discutida no terceiro capítulo. As conclusões obtidas são expostas no quarto capítulo, que é seguido pelas referências bibliográficas, no quinto capítulo.

2. Referencial teórico

2.1. Definições e conceitos

2.1.1. Dado, informação e conhecimento

Os conceitos relativos aos termos “dado”, “informação” e “conhecimento” serão inicialmente estabelecidos, sob a ótica deste trabalho, posto que são várias as suas definições na literatura da Ciência da Informação.

Miranda (1999) define dado como “um conjunto de registros qualitativos ou quantitativos conhecido que, organizado, agrupado, categorizado e padronizado adequadamente, transforma-se em informação”. Davenport (1998) afirma que um dado corresponde a "observações sobre o estado do mundo".

Segundo Barreto (1996), informações “são estruturas significantes com a competência de gerar conhecimento no indivíduo, em seu grupo, ou na sociedade”, e a qualifica como “um instrumento modificador da consciência do homem e de seu grupo social”. Miranda (1999) define informação como dados organizados de modo significativo, sendo subsídio útil à tomada de decisão.

Para Barreto (1996) o conhecimento é “a alteração provocada no estado cognitivo do indivíduo, isto é, no seu estoque mental de saber acumulado, proveniente de uma interação positiva com uma estrutura de informação”.

Para Davenport e Prusak (1998), o conhecimento é a mistura fluida de experiência condensada, valores, informação contextual e *insight* experimentado, a qual proporciona uma estrutura para a avaliação e incorporação de novas experiências e informações.

Sirihal e Lourenço (2002) resumem, em sua pesquisa sobre vários autores da Ciência da Informação, que “informação é conhecimento fragmentado para sua melhor assimilação e disseminação, principalmente em meio eletrônico; e conhecimento é todo o saber existente nos seres e na natureza que é explicitado através de sua fragmentação em informação”. Os autores afirmam também que “é impossível falar-se em informação sem contemplar o termo conhecimento. Para grande parte dos autores da Ciência da Informação, a informação se inter-relaciona com termos tais como comunicação, mensagem, conhecimento”.

Nonaka e Takeuchi (1997), afirmam que “o conhecimento tácito é o conhecimento pessoal, que é difícil formalizar ou comunicar a outros. É constituído do *know-how* subjetivo, dos *insights* e intuições que uma pessoa tem depois de estar imersa numa atividade por um longo período de tempo. Conhecimento explícito é o conhecimento formal, que é fácil transmitir entre indivíduos e grupos. É frequentemente codificado em fórmulas matemáticas, regras, especificações, e assim por diante. Os dois conhecimento são complementares”.

Conhecimento tácito é o acúmulo de saber prático sobre um determinado assunto, que agrega convicções, crenças, sentimentos, emoções e outros fatores ligados à experiência e à personalidade de quem o detém. O Conhecimento explícito é o conjunto de informações já em algum suporte (livros, documentos, etc.) e que caracteriza o saber disponível sobre tema específico.

Nonaka e Takeuchi (1997) estruturaram um modelo para as conversões possíveis do conhecimento entre tácito e explícito através dos processos de exteriorização, combinação, internalização e socialização. Estas conversões nas dimensões ontológica, epistemológica e temporal dão origem à Espiral do Conhecimento, representada na figura 2.

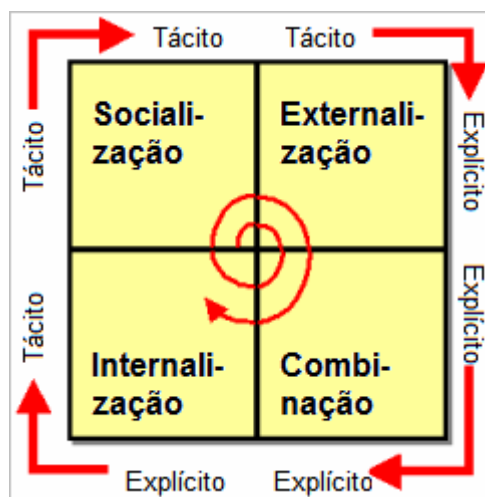


Figura 2 – A espiral do conhecimento.

Fonte: Nonaka e Takeuchi (1997).

As pessoas das diferentes unidades de trabalho que compõem uma organização têm necessidade de dados, informação e conhecimento para desenvolverem suas tarefas cotidianas, bem como para traçarem estratégias de atuação. Portanto, dados, informação e conhecimento são insumos básicos para que essas atividades obtenham resultados satisfatórios ou excelentes. Para gerenciar esses fluxos informacionais, quer formais ou informais, é necessário realizar algumas ações integradas objetivando prospectar, selecionar, filtrar, tratar e disseminar todo o ativo informacional e intelectual da organização, incluindo desde documentos, bancos e bases de dados etc., produzidos interna e externamente à organização até o conhecimento individual dos diferentes atores existentes na organização. (VALENTIM, 2002, p.4).

Davenport e Prusak (1998) dizem que o conhecimento pode se desenvolver através da experiência, que se refere ao que acontece e ao que aconteceu no passado. O conhecimento nascido da experiência reconhece padrões familiares e pode inter-relacionar os acontecimentos atuais com os passados.

Sistemas informatizados prestam-se muito mais para gerenciar o conhecimento explícito do que o tácito. Apesar deste fato deve-se considerar que a criação do conhecimento tácito é facilitada com o uso de sistemas de informação, pois à medida que o conhecimento explícito é mais bem gerenciado, sua disponibilização

é melhorada e há maior interação entre os dois tipos de conhecimento, o que é essencial para o processo de criação de conhecimento. (FERREIRA, 2003, p.44).

Segundo, Ferreira (2003), a tecnologia da informação lida principalmente com o conhecimento explícito, mas seria incorreto afirmar que a gestão do conhecimento tácito não é apoiada. O autor sugere também que ocorre “desnecessária dicotomia entre gestão e tecnologia”, afirmando que a tecnologia “alavanca e potencializa os instrumentos e práticas de gestão”.

2.1.2. Regras organizacionais para tomada de decisão

Slovic et al (1977) dizem que “devido à limitada capacidade de processamento de informação e por ignorância de regras para processamento de informações na tomada de decisão, o julgamento é sujeito a desvios sistemáticos”.

Choo (2003) afirma que como há limitações de tempo, de recursos e, acima de tudo, de energia intelectual para identificar as alternativas, prever as conseqüências e esclarecer as preferências, a atenção torna-se um recurso escasso, que influencia a decisão, assim como a quantidade e a qualidade da informação na qual ela se baseia. O mesmo autor diz que a realidade é comparada com padrões baseados em tendências passadas, projeções, situações comparadas, expectativas de outros e modelos teóricos.

Mintzberg et al (1996) afirmam que “os tomadores de decisão buscam reduzir a decisão em subdecisões nas quais eles aplicam conjuntos de procedimentos ou rotinas gerais”.

Simon (1976) afirma que os homens são racionalmente limitados: quando tentam agir de maneira racional, a racionalidade de seu comportamento é limitada por suas capacidades cognitivas e por restrições organizacionais. Algumas características da racionalidade humana relacionadas ao processo decisório se organizam por programas ou estratégias, devendo os tomadores de decisão construir modelos mentais.

Segundo Kim (1993), a análise da memória ativa da organização define em que se deve prestar atenção, como decidir e qual parte da sua experiência selecionar e lembrar. Assim se constituem os modelos mentais compartilhados e individuais. A explicitação e o compartilhamento destes modelos facilitam a criação do conhecimento através da experiência e apóiam o aprendizado organizacional. O aprendizado ocorre quando as organizações codificam inferências da história em rotinas que orientam o comportamento.

2.2. Descoberta de conhecimento em bases de dados

A automatização dos processos de negócio das empresas criou um fluxo enorme de dados que se multiplica em poucos meses. A análise destes dados é impossível a olho nu, sendo necessária a criação de sistemas automatizados para a descoberta de informação e conhecimento nestas bases de dados.

Diversas empresas e instituições utilizam e demandam melhores métodos e ferramentas para analisar suas bases de dados à procura de padrões úteis e interessantes.

A DCBD, ou o termo em inglês KDD⁶, é a extração da informação implícita, não trivial, previamente desconhecida de uma base de dados. Dado um conjunto de dados F , uma linguagem L , e uma medida de certeza C , define-se um padrão como uma declaração S em L que descreve relacionamentos entre um subconjunto F_S de F com uma certeza c , na qual S é mais simples (de alguma forma) do que a enumeração de todos os dados em F_S . Um padrão que é interessante (de acordo com uma medida de interesse do usuário) e certo o suficiente (novamente de acordo com critérios do usuário) é chamado conhecimento. A saída de um sistema que monitora um conjunto de dados em uma base de dados e produz padrões de acordo com o significado definido acima é chamada conhecimento descoberto. A arquitetura de um sistema DCBD está ilustrada na figura 3.

O campo de estudos do KDD é, em um nível abstrato, voltado para o desenvolvimento de métodos e técnicas objetivando dar sentido aos dados. O problema básico trabalhado pelo processo do KDD é o mapeamento de dados brutos (que tipicamente são muito volumosos para se entender e assimilar) em formas que devem ser mais compactas (como um relatório), mais abstratas (como um modelo de aproximação descritivo do processo que gerou os dados), ou mais útil (como um modelo preditivo de estimação de valores para casos futuros). (FAYYAD et al, 1996, p.37).

A DCBD exibe quatro características principais:

- Deve ser representada em linguagem de alto nível, compreensível para os usuários;
- Exatidão: as descobertas devem representar o conteúdo da base com exatidão. A extensão desta representatividade é expressa pela certeza;
- Resultados interessantes: o conhecimento descoberto será interessante de acordo com critérios definidos pelo usuário. Normalmente o conhecimento será interessante se for novo, potencialmente útil, e se o procedimento de descoberta for não-trivial;

⁶ KDD – *Knowledge Discovery in Databases*, termo inglês muito utilizado na literatura, traduzido como descoberta de conhecimento em bases de dados.

- Eficiência: os tempos de execução para grandes bases de dados devem ser previsíveis e aceitáveis.

A figura 3 ilustra a arquitetura de um sistema para DCBD.

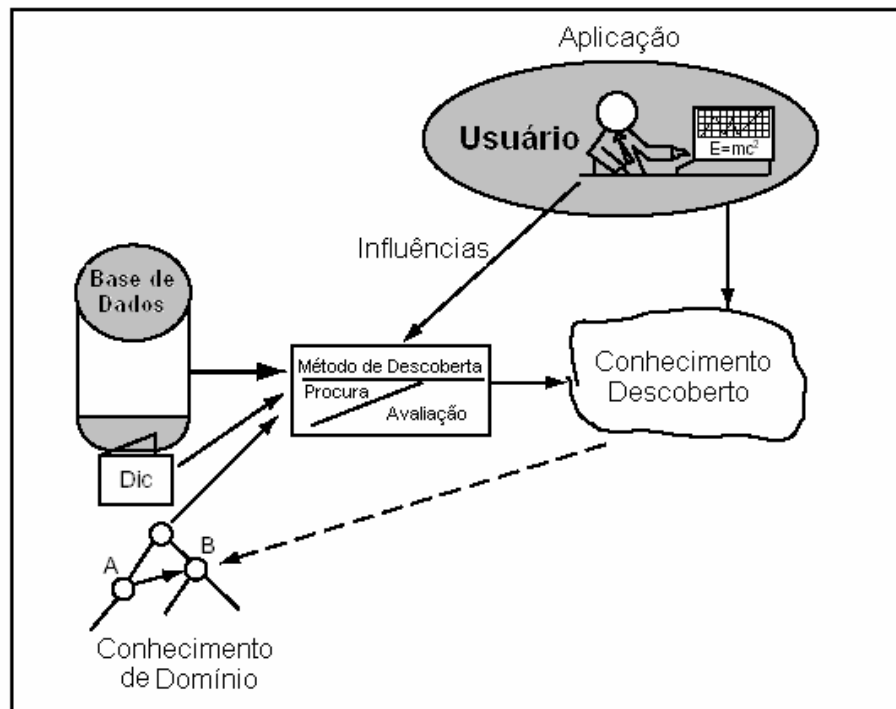


Figura 3 – Arquitetura de um modelo para descoberta de conhecimento.

Fonte: FRAWLEY et al (1992).

No núcleo do sistema está o método de descoberta, que processa e avalia os padrões à medida que estes são reconhecidos. A entrada para o método de descoberta inclui registros de dados da base de dados, informação do dicionário de dados, conhecimento de domínio⁷ adicional, e influências definidas pelo usuário que definem o foco de análise. A saída do processo é o conhecimento descoberto que pode ser devolvido para o usuário ou, em alguns casos, pode ser devolvido para o sistema como um novo conhecimento de domínio.

⁷ Conhecimento de domínio - é aquele que auxilia o processo de descoberta restringindo e direcionando a pesquisa.

Inicialmente uma base de conhecimento pode ser construída com poucas regras, mas, dependendo da complexidade do ambiente e das necessidades de informações variadas, esta base poderá eventualmente crescer para milhares de regras e fatos. Assim, é preciso que se tenha o cuidado de implementar instrumentos internos de refinamento que possibilitem podas na árvore de decisão e cortes na base de conhecimento, para que o processo de busca localize segmentos cujas regras e fatos contemplem os instrumentos necessários que conduzam à solução dos problemas em questão. (MENDES, 1997, p.41).

O sistema pretende apoiar a tomada de decisões pelo analista, ou seja, pelo usuário. A resposta se relaciona com o conhecimento de domínio específico de uma certa comunidade profissional. Desta forma, o receptor do sistema idealmente deve compartilhar uma base de conhecimento de domínio para que possa interpretar e usar a resposta do sistema construtivamente, ou até rejeitá-la, se esta lhe parecer fora do contexto. Ou seja, o usuário deve ter a capacidade de tomar decisões informadas sobre a relevância do conhecimento descoberto.

2.2.1. A natureza interdisciplinar da DCBD

Segundo González de Gómez (2001), o encontro interdisciplinar é o momento da comunicação e do intercâmbio entre áreas do conhecimento diferenciadas, mas com um interesse compartilhado que impulsiona a convergência em direção à reflexão sobre um problema comum.

Segundo Fayyad et al (1996), a DCBD evolui da interseção de campos de pesquisa sobre estatística, aprendizado de máquina⁸, reconhecimento de padrões, bases de dados, inteligência

⁸ Do inglês *machine learning*: é o estudo de algoritmos cujo desempenho em determinada tarefa melhora com a experiência.

artificial, geração de conhecimento para Sistemas Especialistas, visualização de dados e performance computacional, entre outros. O objetivo comum é extrair conhecimento de alto nível partindo de grandes volumes de dados. Enquanto estes campos de pesquisa fornecem métodos aplicados no processo de MD, a DCBD visualiza o processo em um nível mais amplo, incluindo:

- Como os dados são registrados e acessados;
- Como os algoritmos podem ser eficientemente usados em grandes bases de dados;
- Como os resultados podem ser interpretados e visualizados;
- Como a interação homem-máquina pode ser modelada e apoiada.

As interpretações dos padrões e das regras descobertas têm um interesse especial no processo de DCBD. Algumas técnicas como as redes neurais produzem modelos mais difíceis de serem apresentados aos usuários, mas são competentes na redução de ruído nos dados analisados. Por outro lado, técnicas como as árvores de decisão, que será estudada nesta pesquisa, apresentam resultados de fácil compreensão.

As técnicas estatísticas fornecem linguagem e padrões utilizados na quantificação da incerteza inferida sobre as amostras de dados analisadas. Apoiada sobre técnicas estatísticas para a seleção de hipóteses, a DCBD deve seguir métodos estatisticamente corretos e deve ser aplicada sobre amostras estatisticamente válidas ou o resultado será comprometido.

O desenvolvimento de algoritmos para manipulação eficiente, acesso, agrupamento e otimização de consultas a bases de dados apóia o processo de manipulação de grandes bases de dados necessários para a DCBD.

Inmon (1992) define o conceito de armazém de dados como uma coleção de dados orientada ao assunto, integrada, variante no tempo e não volátil para suporte a decisões. Iniciativas de armazém de dados contemplam a análise das áreas de negócio da instituição para a definição dos seus indicadores de performance e de suas dimensões⁹. Os dados presentes em diversos sistemas operacionais da instituição são integrados em um modelo de dados único alinhado à visão tática e estratégica das atividades da instituição. A disponibilidade deste ambiente é favorável para a DCBD, pois esta base de dados é rica quanto aos conhecimentos que dali podem ser extraídos.

Segundo Mendes (1997), os Sistemas Especialistas são sistemas baseados em conhecimento, construídos, principalmente, com regras que reproduzem o conhecimento do perito e são utilizados para solucionar determinados problemas em domínios específicos, possibilitando aumento na produtividade de um especialista na execução de tarefas especializadas, quando assistido por estes sistemas. Os Sistemas Especialistas são constituídos pelos mesmo três elementos fundamentais do DCBD: base de conhecimento, motor de inferência e interface com o usuário.

Sob nosso ponto de vista, a Ciência da Informação e muitas outras áreas podem encontrar nos Sistemas Especialistas, eficientes ferramentas para o gerenciamento da informação. Disponibilizar ferramentas para suporte à tomada de decisão, neste caso, vai mais além do que fornecer gráficos e tabelas ao usuário significa prestar-lhes orientação, na identificação de suas necessidades, simulando cenários e possibilitando maior exatidão e confiabilidade nos seus resultados. (MENDES, 1997, p.40).

Furnival (1995) diz que os melhores Sistemas Especialistas foram construídos com o usuário em mente, codificam conhecimento de um domínio especializado bem delimitado e foram

⁹ As dimensões são as entidades que caracterizam as operações de negócio (fatos) da instituição no ambiente de análise. Como exemplo em uma companhia aérea um bilhete (fato) é emitido para: um trecho, um cliente, em uma data (dimensões).

construídos para serem usados por especialistas destes domínios, que possuiriam a base de conhecimento tácito necessário para interpretar as respostas do sistema.

2.3. Mineração de dados

A MD é um processo de suporte à decisão no qual se procura por padrões escondidos em grandes volumes de dados. Seus objetivos principais são a previsão e a descrição. A MD preditiva faz uso de variáveis existentes na base de dados para prever valores desconhecidos ou futuros. A MD descritiva é voltada para a busca e a apresentação de padrões que descrevem os dados.

O conhecimento extraído deve ser compreensível, potencialmente útil e válido para um novo conjunto de dados com um determinado grau de incerteza¹⁰, podendo ser utilizado no apoio às decisões.

A MD é uma fase da DCBD. O processo tem natureza iterativa (para melhor ajuste dos parâmetros a cada iteração) e interativa (baseado no conhecimento dos especialistas e dos usuários).

¹⁰ Deve-se distinguir o “risco” da “incerteza”: o primeiro existe quando a tomada de decisões é baseada em probabilidades objetivas para a estimação de diferentes resultados. Desta forma a expectativa se fundamenta em dados históricos, permitindo que as decisões sejam tomadas a partir de estimativas consideradas aceitáveis. A incerteza é observada quando não se tem à disposição os dados históricos acima mencionados, levando a decisões baseadas em observações subjetivas.

Segundo Fayyad et al (1996), a descoberta de conhecimento de bases de dados envolve a realização possivelmente iterativa de várias atividades como:

- Identificação do domínio: define-se o escopo, objetivos, metas e o conjunto de dados;
- Seleção e limpeza dos dados: obtenção, unificação e transformação para eliminar ruídos e garantir a qualidade dos dados;
- Redução do volume de dados: seleção de exemplos (através de métodos como a amostragem aleatória), redução de atributos e de seus valores (através de discretização de seus valores, por exemplo);
- Escolha da função de MD: função preditiva ou descritiva;
- Escolha do algoritmo de MD: ocorre de acordo com a função e com a linguagem de representação dos padrões encontrados;
- Busca por padrões: consiste na aplicação dos algoritmos selecionados para a busca dos padrões contidos nos dados;
- Avaliação do conhecimento: aplicação de medidas para critérios como precisão e compreensibilidade sobre o conhecimento gerado;
- Refinamento do conhecimento: combinação de conhecimentos obtidos através de vários modelos e a classificação destes visando obter melhores resultados a cada iteração.

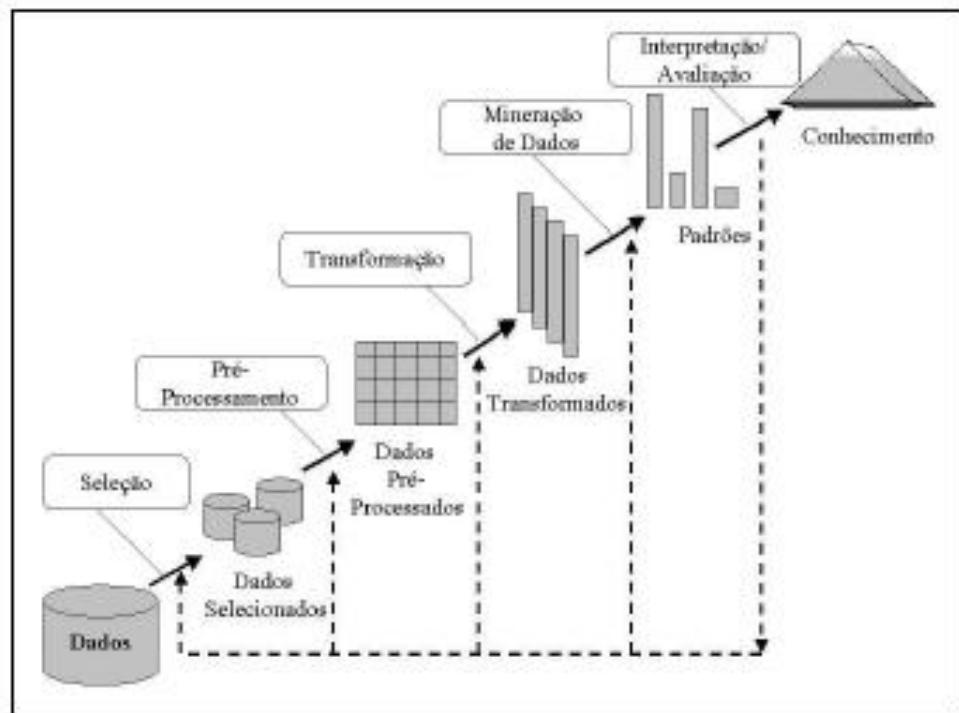


Figura 4 – Etapas da DCBD

Fonte: Fayyad et al (1996).

Inicialmente uma base de conhecimento pode ser construída com poucas regras, mas, dependendo da complexidade do ambiente e das necessidades de informações variadas, esta base poderá eventualmente crescer para milhares de regras e fatos. Assim, é preciso que se tenha o cuidado de implementar instrumentos internos de refinamento que possibilitem podas na árvore de decisão e cortes na base de conhecimento, para que o processo de busca localize segmentos cujas regras e fatos contemplem os instrumentos necessários que conduzam à solução dos problemas em questão. (MENDES, 1997, p.41).

Segundo Rezende et al (2003), o processo de MD é um ciclo dividido em três grandes etapas: Pré-processamento, Extração de Padrões e Pós-Processamento, que são precedidas pela fase de Identificação do Problema, que são sucedidas pela fase de Utilização do Conhecimento. A figura 5 ilustra o ciclo das etapas do processo de MD.



Figura 5 – Etapas do processo de MD.

Fonte: Rezende et al (2003).

2.3.1. Tarefas de mineração de dados

A MD pode ser usada para vários tipos de tarefas. Baseado no tipo de conhecimento a ser descoberto, o processo pode ser supervisionado ou não supervisionado. O primeiro tipo requer que os dados tenham atributos que indiquem a qual classe os registros na base de dados pertencem. O algoritmo de extração faz uma identificação das características comuns entre os membros de uma classe e fornece uma descrição para cada classe. O modelo é construído através dos exemplos, da definição de classes e do fornecimento de exemplos para cada uma. O processo deve achar uma descrição de cada classe, tal como as propriedades comuns dos exemplos. Uma vez que a descrição tenha sido formulada, a descrição e a classe formam uma regra de classificação que pode então ser utilizada para a previsão de classes de objetos ainda não vistos. Esta técnica é similar à análise discriminante em estatística. Um tipo de descrição

utilizada, que usa uma representação a partir de regras é a descrição característica, como se sugere no exemplo:

$$\acute{E_animal}(x), tem_asas(x), pode_voar(x) \rightarrow \acute{e_pássaro}$$

O segundo tipo não requer pré-classificação. Serão formados grupos com características comuns. O modelo é construído através de observação e descoberta. O processo é suprido com os dados, mas nenhuma classe é definida, de forma que o modelo deve observar os exemplos e reconhecer padrões (descrição das classes) por si mesmo. O processo resulta em um conjunto de descrições de classes, uma para cada classe descoberta. Novamente, isto é similar à análise de grupos em estatística.

A figura 6 ilustra algumas das tarefas de MD, classificando-as em atividades preditivas e descritivas. É usual a utilização de várias delas em uma mesma base de dados visando à combinação do conhecimento extraído por cada tipo de tarefa.

A presente pesquisa se baseia no uso de árvores de decisão através de regressão e classificação para a criação do modelo de decisão de crédito. Além das técnicas discutidas nas próximas seções, existem outras tarefas de MD fora do escopo desta pesquisa, como redes neurais, por exemplo.

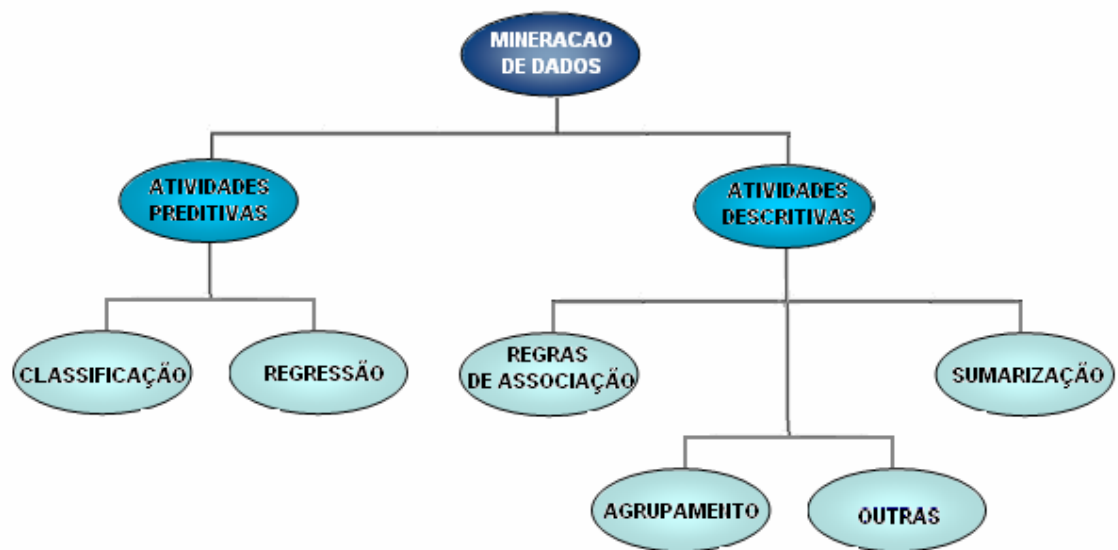


Figura 6 – Atividades e tarefas de MD.

Fonte: Rezende et al (2003).

2.3.1.1. Regras de Associação

Regras de Associação são usadas para encontrar relacionamentos significativos ou características que ocorrem juntas em uma grande base de dados. O objetivo é encontrar e descrever as correlações escondidas nos dados registrados. Segundo Agrawal & Srikant (1994), uma regra de associação caracteriza o quanto um conjunto de itens nos registros de uma base de dados é associado, com certo grau de segurança, à presença de algum outro conjunto distinto de itens nos mesmos registros.

O formato de uma Regra de Associação pode ser representado como uma expressão da forma: $LHS \rightarrow RHS$, onde LHS e RHS são, respectivamente, o lado esquerdo e o lado direito da regra, definidos por conjuntos distintos de itens.

Para ilustração, em um problema hipotético em que se investigam consumidores de pipoca e guaraná, a regra de associação poderia ser:

- 75% dos consumidores que compram pipoca também compram Guaraná.

O índice 75% refere-se à confiança, uma medida do poder preditivo da regra. O item do LHS é pipoca, enquanto que Guaraná está no RHS. O algoritmo produz uma grande quantidade destas regras e deve-se observar a percentagem dos registros que observam esta regra (suporte) e seleccionar os subconjuntos de regras que têm maiores graus de confiança.

No exemplo da figura 7, as regras são exibidas em coordenadas, onde o inicial é o eixo horizontal que representa a confiança da regra e o eixo vertical representa o seu suporte.

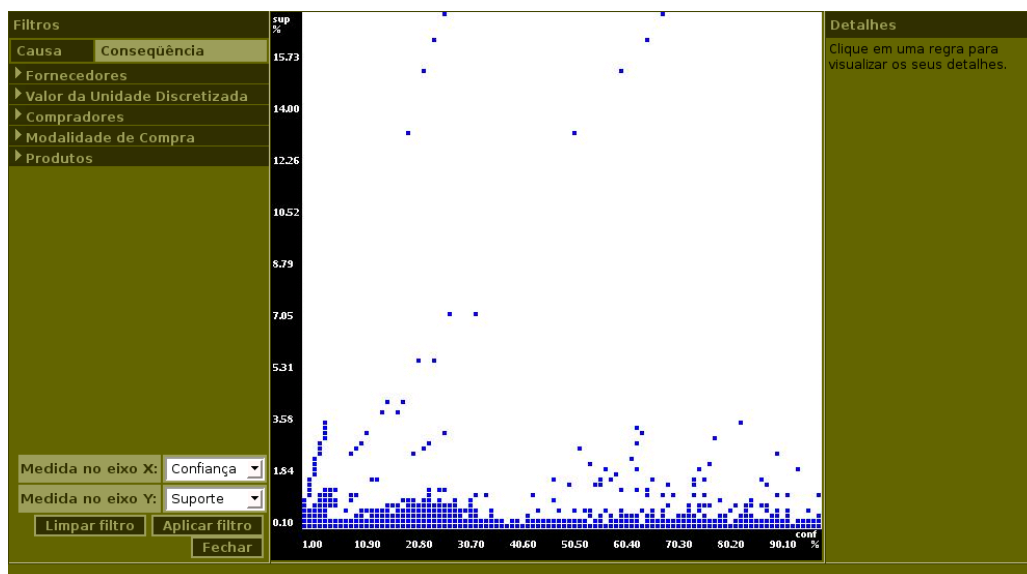


Figura 7 – Exemplo de visualização de regras de associação

Fonte: DCC/UFMG - Projeto Tamanduá - Manual de Ajuda (2006).

2.3.1.2. Agrupamento

Agrupamento consiste na divisão dos dados em grupos de objetos similares. Cada grupo consiste de objetos similares entre si e diferentes dos objetos em outros grupos. São identificadas classes de itens em uma base de dados de acordo com alguma medida de similaridade. Agrupamento é então uma tarefa não supervisionada para a descoberta de padrões nos dados, e pode ser usado em situações onde os registros não estejam pré-classificados. O objetivo é descrever os dados disponíveis. Esta técnica tem a vantagem de revelar tendências desconhecidas, correlação, padrões e não são necessárias suposições sobre a estrutura dos dados. A técnica de agrupamento costuma ser utilizada em Gerenciamento de Relacionamento com Clientes como uma etapa para a segmentação de clientes, e está ilustrada na figura 8.

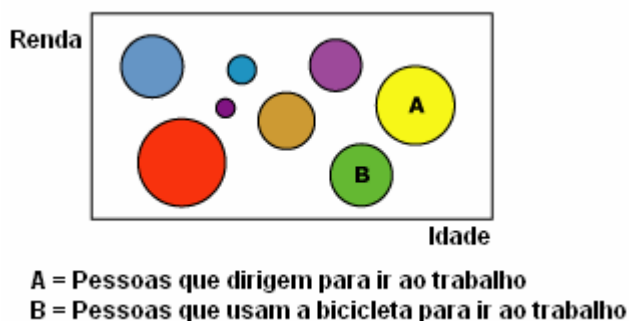


Figura 8 – Exemplo de um agrupamento

Fonte: Microsoft SQL Server 2005 Books Online (msdn.microsoft.com)

2.3.1.3. Classificação

A Classificação, um processo supervisionado, tenta descobrir padrões preditivos baseados nos atributos dos registros. Deve ser identificado o conjunto mínimo das características conhecidas de um determinado conjunto de dados que seja suficiente para prever uma

característica desconhecida. O objetivo é prever as características dos dados futuros, com base nos dados disponíveis.

2.3.1.4. Árvore de Decisão

A Árvore de Decisão possui uma estrutura ramificada onde as folhas representam classificações e os ramos representam as conjunções das características que levam a estas classificações, baseadas nos atributos considerados (variáveis de entrada). Dentre as várias vantagens no seu uso, destacam-se:

- Em geral, não é necessária uma rigorosa pré-seleção das variáveis de entrada para o modelo, pois os processos de construção de árvores de decisão possuem mecanismos robustos de seleção de atributos;
- O modelo resultante é de simples compreensão, interpretação e explicação, facilitando o compartilhamento do conhecimento descoberto;
- Após a criação do modelo, observa-se uma boa performance na predição. Isto permite a tomada de decisão baseada no conhecimento produzido.

A figura 9 representa uma árvore de decisão gerada pela análise de registros relativos à decisão sobre a realização de uma partida de *golf*.

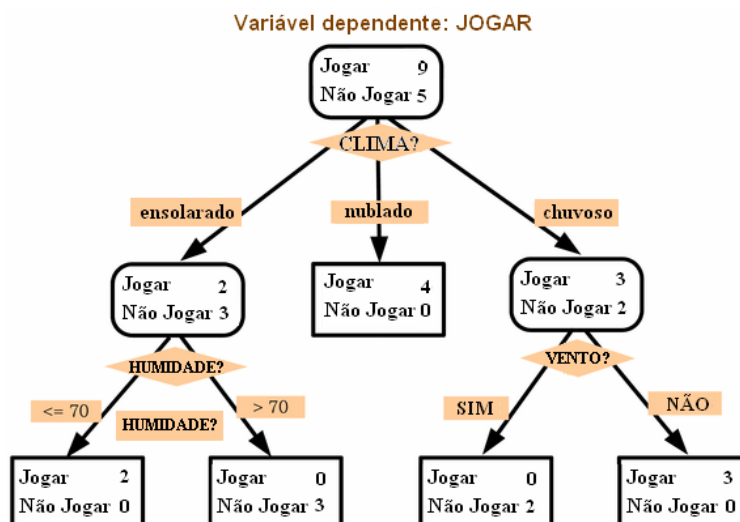


Figura 9 – Exemplo de árvore de decisão

Fonte: Wikipedia contributors (2006a)

A realização da partida (variável dependente) se dá em função das condições climáticas (variáveis de entrada). É possível realizar previsões simplesmente verificando em qual posição da árvore um cenário se encaixa. Em modelos maiores e mais complexos, o sistema automaticamente posiciona cada evento questionado na posição correta e informa a previsão ao usuário.

2.4. Concessão de crédito e escoragem de crédito

Segundo Guimarães (2000), a concessão de crédito significa a transferência da posse de um bem ou de uma quantia em dinheiro, mediante a promessa de pagamento futuro. O crédito à disposição de uma pessoa, física ou jurídica, é a capacidade da mesma em obter dinheiro, produtos ou serviços mediante compromisso de pagamento em determinado período. A

decisão de colocar certo valor à disposição de um cliente, com a promessa de seu recebimento em data futura, levanta a questão da capacidade de pagamento do cliente. Este é um aspecto que deve ser avaliado numa análise de concessão de crédito, sendo necessária uma avaliação do potencial atual e futuro de recursos do tomador em relação ao montante de recursos solicitado e de seus encargos.

Mario (2000) diz que cada oportunidade de retorno avaliada em operações de crédito está condicionada a um nível de risco, sendo que, para a obtenção de um retorno maior, tem-se que assumir um risco maior. Entretanto, muitas vezes o risco que se tem que assumir não é compensado pelo retorno.

2.4.1. Métodos subjetivos para concessão de crédito

A avaliação do risco de crédito dos proponentes pode acontecer por processo de avaliação subjetiva ou por métodos quantitativos, como a escoragem de crédito (EC). Segundo Schrinkel (1997), os modelos subjetivos são baseados no julgamento apoiado pela experiência acumulada dos especialistas na avaliação de crédito. O método pode apresentar vários problemas, como se mostrar preconceituoso: um analista de crédito pode ter restrições, por motivos pessoais, a uma certa categoria profissional. Isto pode levá-lo a recusar uma proposta que seria aceita por outro analista. Não existem medidas objetivas de risco.

Saunders (1999) aponta problemas na comparabilidade das avaliações de crédito e na ponderação subjetiva dos fatores de risco. Quanto ao problema da comparabilidade, ele decorre do fato de diversos analistas lançarem mão de diferentes critérios para avaliar o risco de crédito em uma mesma instituição. Quanto ao problema da ponderação, raramente são

estabelecidos pesos similares a serem aplicados de modo coerente em toda a organização para os fatores incluídos dentre os critérios de seleção de créditos.

2.4.2. O uso de modelos de escoragem de crédito

Thomas (2002) relata que nos Estados Unidos, até a segunda guerra mundial, as decisões quanto a conceder ou não empréstimos eram feitas pelos analistas de crédito. A convocação destes para o serviço militar provocou uma escassez de pessoas com a experiência necessária. Assim, as empresas fizeram com que os analistas escrevessem as regras que usavam para decidir a quem conceder empréstimos. Essas regras foram, então, usadas por pessoas que não dominavam o assunto para ajudar na tomada de decisões de crédito — um dos primeiros exemplos de Sistemas Especialistas. Pouco depois da guerra, isto foi relacionado com as pesquisas feitas na biologia para diferenciação de grupos em populações através de estatística. Na década de 50 surgiu a primeira empresa de consultoria nesta área atendendo a financeiras, varejistas e empresas de compras por mala direta. No final da década de 60 a explosão da demanda pela emissão de cartões de crédito exigiu a decisão de crédito automatizada e realçou a utilidade de métodos de EC. As quedas nas taxas de inadimplência observadas comprovaram suas vantagens.

Em meados da década de 70, o Congresso Americano tornou ilegal a discriminação de crédito, a menos que houvesse justificativa estatística para ela, o que desencadeou a adoção de EC. Nos anos 80, o método foi expandido para outros produtos, como crédito pessoal e empréstimo imobiliário. Foram introduzidos métodos estatísticos mais avançados, como a regressão logística. Nos anos 90, a expansão do *marketing* e das ferramentas computacionais

possibilitou o uso de EC para direcionamento de campanhas e se iniciaram os primeiros testes com técnicas de inteligência artificial, tais como Sistemas Especialistas e redes neurais. Thomas (2002) afirma que a EC é uma ferramenta que permite reconhecer diferentes grupos que compõem uma população, quando não é possível identificar as características que os separam, mas apenas correlatas.

A atual ênfase é em avaliar como a empresa pode maximizar o lucro com um cliente. Procura-se estimar a resposta (a probabilidade de um cliente reagir positivamente a uma mala direta sobre um novo produto), utilização (a chance de um cliente usar o produto), retenção (a probabilidade de um cliente continuar utilizando o produto após o término do período da oferta introdutória), atrito (a possibilidade de o cliente migrar para outro credor) e gerenciamento de dívida (a chance de sucesso de diversas abordagens à prevenção de inadimplência no princípio dos problemas com o cliente).

Vasconcellos (2002) diz que o termo EC é um termo que descreve métodos estatísticos que classificam candidatos ao crédito em grupos de risco.

Segundo Pereira et al (2002), a EC é como o sistema de pontuação dos dados demográficos e cadastrais de um proponente ao crédito. Com base nesse sistema, as propostas são aprovadas, rejeitadas, recomendadas ou não recomendadas, dependendo do ponto de corte do modelo implementado e da alçada de decisão de crédito.

O sistema avalia as características do cliente e, em alguns casos da proposta, atribuindo um determinado peso a cada característica, variando conforme seu valor. Estas podem ser o histórico de pagamentos de dívidas, número e tipo de contas, ações de cobrança, renda,

imóveis próprios, permanência em empregos (no caso de pessoas). Alguns fatores têm efeito positivo e outros negativos sobre o escore. Atraso no pagamento, suspensão de limites de crédito e falência afetam negativamente, enquanto que posse de casa própria, longa permanência em empregos e baixo endividamento contribuem de forma positiva. Os pesos são obtidos a partir de estatísticas, e normalmente expressam uma percentagem. Assim, por exemplo, se 35% dos clientes “bons” pertencem a uma certa faixa etária, a um proponente com idade correspondente é atribuído o peso 0,35. Os pesos (coeficientes) são multiplicados pelos valores das respectivas variáveis e somados aos outros valores assim calculados, obtendo-se um valor, denominado escore.

O escore é o indicativo para a decisão de conceder, ou não, o crédito. Para tomar tal decisão, o escore é comparado com um valor previamente estabelecido, chamado ponto de corte. É na obtenção deste último que reside a maior parte dos problemas enfrentados pelos profissionais envolvidos. O ponto de corte deve ser confiável a ponto de evitar perdas para a empresa, tanto pela aceitação, errada, de clientes que venham a se tornar inadimplentes, quanto pela rejeição, igualmente errada, de clientes que se revelariam adimplentes.

O amplo uso de sistemas de EC é devido a estes basearem suas informações em dados reais e em estatísticas, que normalmente são mais seguros do que métodos subjetivos ou de julgamento: o processo torna-se supostamente mais explícito, eficiente e transparente.

Os modelos de escoragem [...] apresentam aplicações em todas as fases do ciclo de crédito. Permitindo que a decisão seja tomada de forma objetiva, padronizada e imparcial, o que não é garantido pela análise julgamental. [...] No início do relacionamento, algumas de suas aplicações principais incluem a prévia mensuração do risco do proponente e a atribuição de linhas diferenciadas em função do perfil do proponente. Já no gerenciamento do portfólio, aplicam-se com destaque à manutenção das linhas concedidas, concessão de linhas adicionais, *cross-sell* e precificação diferenciada, além de ações preventivas. Na fase de cobrança, define a severidade da ação a ser tomada. Portanto o leque de aplicações possíveis é vasto, permitindo o constante monitoramento do comportamento do cliente e mudanças de rota, caso necessário. Isto é fundamental, pois o risco do tomador de crédito não

termina com a aprovação da operação, mas sim após cumprimento do contrato e de todas as suas obrigações perante o credor. (LOURENÇO, 2005, p.2).

Além de reduzir a subjetividade na decisão de concessão de crédito, o sistema reduz também o tempo de processamento desta, liberando o decisor para outras tarefas, como a avaliação de outras características não computadas pelo sistema, a captação de novos clientes e a recuperação de créditos problemáticos concedidos.

2.4.3. Desenvolvimento de um sistema de escoragem de crédito

O desenvolvimento de um sistema de EC requer as seguintes premissas, dentre outras:

- Participação de analistas de crédito e de informação da instituição financeira: estes analistas conhecem as características do produto e do mercado analisados, as informações disponíveis e onde encontrá-las;
- Utilização de dados e informações relativos aos produtos e mercados nos quais a instituição atua. A utilização de modelos externos requer testes rigorosos para verificar o grau de aderência de tais modelos às operações da instituição;
- Dependência entre a qualidade das informações e a correta seleção de variáveis cadastrais utilizadas. São recomendáveis processos de auditoria para verificar a exatidão das informações e a exatidão do modelo. É comum a ocorrência de problemas como a falta de uma base de dados confiável e também com questões legais, que muitas vezes impedem o adequado levantamento das informações.

Guimarães (2000) aponta a dinâmica das características da população como um dos fatores que podem comprometer o desempenho do sistema: “um atributo indicador de situação financeira estável em determinada época pode não ser significativo em outra”.

Os fatores considerados críticos para a implantação, e também para o sucesso, do sistema são: a cultura da empresa, a definição da política de score, os procedimentos operacionais vigentes e a qualidade dos sistemas de gerenciamento de informações. (GUIMARÃES, 2000, p. 11).

Quanto às vantagens de um método quantitativo de análise de propostas em substituição a critérios subjetivos (julgamentais), Lawrence (1984) aponta as seguintes:

- Apóia o processo decisório de deferimento de propostas de crédito;
- Auxilia na detecção de tendências econômicas;
- Permite a implementação eficiente de mudanças na política de crédito.

Sicsú (1998) expõe etapas que constituem uma metodologia para o Desenvolvimento e Validação de um Sistema de EC, que serão discutidas nos tópicos seguintes deste capítulo.

2.4.3.1. Planejamento e definições

➤ Quanto às finalidades de uso

O objetivo pode ser a aprovação ou a pré-aprovação do crédito. Sistemas para a pré-aprovação limitam-se às informações históricas cadastrais e, geralmente, não incluem características da operação, como valor e número de parcelas da operação em questão. Nestes casos o score é uma informação adicional para o analista de crédito responsável pela decisão. Este procedimento é recomendável para operações com montantes significativos para a instituição, que exigem análises mais complexas.

➤ Produtos e mercados

Quanto mais abrangente o modelo, menor será a sua eficiência. Deve-se definir para quais produtos de crédito e para quais mercados o modelo será construído. Modelos analisando vários produtos ou mercados acabam desprezando suas características específicas na formação do cálculo do score. Se uma instituição possui dois produtos de crédito distintos oferecidos ao proponente, ela não poderá utilizar o mesmo modelo de EC para os dois produtos, pois ambos possuem características diferentes e seus clientes terão, conseqüentemente, comportamentos diferentes.

➤ **Horizonte de previsão**

De maneira análoga à abrangência dos produtos e mercados considerados, quanto maior o horizonte de previsão do sistema, menor será a sua eficiência. Esta definição é importante para a definição do período e tamanho da amostra de dados para alimentar o modelo.

➤ **Tipos de clientes**

Deve-se distinguir se o sistema vai operar com dados de novos clientes ou de clientes que já se relacionam com a instituição. No último caso, informações relativas às outras operações do cliente, como atrasos nos pagamentos, têm maior valor preditivo para a decisão de crédito do que as variáveis cadastrais.

➤ **Conceito de inadimplência**

O risco de inadimplência da operação de crédito é utilizado para calcular a rentabilidade esperada da operação e evitar perdas. Deve-se chegar a um consenso entre os analistas sobre o que caracteriza os clientes em “bons clientes” ou “maus clientes”, os inadimplentes. O objetivo é definir exatamente como as operações de crédito seriam classificadas como desejáveis ou indesejáveis para a instituição.

Instituições conservadoras poderiam considerar “bons” os clientes que nunca tiveram atrasos em um período determinado de observação. Outras, menos conservadoras, admitiriam atrasos entre 30 e 60 dias. Os critérios variam de acordo com o tipo de produto e com a cultura da empresa.

Os “maus clientes” teriam inadimplência avançada e poderiam ser definidos, por exemplo, como os que estariam com mais de sessenta dias de atraso, que tenham participado de programas de renegociação de dívida ou que tenham exigido atenção especial legal ou gerencial.

Normalmente a base de dados também contém registros de clientes que tiveram seus créditos recusados pela instituição.

➤ **Classificações intermediárias**

Os estudos de EC costumam se limitar, como ponto de partida, à classificação das operações de crédito em duas categorias (“boas” ou “más”). Pode ser interessante classificar as operações em diferentes categorias intermediárias em função de critérios pré-definidos e classificar as novas propostas em função dessas classes. Entretanto, isto aumenta a complexidade do modelo e da metodologia estatística necessária.

2.4.3.2. Identificação das variáveis potenciais

As variáveis que serão utilizadas devem satisfazer as seguintes condições:

- Possuir validade estatística;
- Não transgredir nenhuma lei vigente;

- Devem, sempre que possível, ser quantitativas ou em condições de serem tratadas como tal.

O processo deve ser interativo, com a equipe do projeto e os analistas fornecendo conhecimento técnico e de domínio para a identificação e escolha das variáveis. O desempenho do modelo depende principalmente da qualidade da informação e da habilidade em classificar as variáveis.

➤ **Caracterização do proponente e da operação**

Procura-se calcular o escore de uma operação através das variáveis que caracterizam o proponente e, quando possível, a operação. São analisados:

- Dados cadastrais;
- Histórico de operações realizadas pelo proponente;
- Índices de liquidez, rentabilidade, etc., calculados a partir dos demonstrativos financeiros quando se tratar de pessoa jurídica e suas coligadas, e informações negativas tais como protestos, cheques devolvidos, concordatas, entre outros.

Quanto à operação são consideradas variáveis sobre o produto e ao tipo de financiamento, como as abaixo para o caso de um financiamento de automóvel, por exemplo são consideradas as seguintes condições, entre outras:

- Tipo de veículo
- Ano de fabricação
- Percentual de entrada
- Comprometimento da renda do proponente

➤ **Identificação e definição das variáveis**

A equação discriminante é montada a partir da amostra de créditos selecionados na empresa usando as variáveis relevantes. São relevantes as variáveis que contêm informação capaz de discriminar os grupos.

Os analistas de crédito e os profissionais da informação devem identificar as variáveis potenciais consideradas para a construção do modelo. É natural escolher inicialmente um conjunto muito amplo de dados cadastrais, quando alguns dos analistas objetivam captar características que consideram importantes, mas que posteriormente irão se mostrar irrelevantes.

Num segundo momento a lista de variáveis potenciais deve ser reduzida baseada na experiência dos envolvidos no projeto, para eliminar a dimensionalidade do problema e viabilizar a aquisição dos dados.

A definição das variáveis deve ser muito clara para todos os analistas. Os metadados, como dicionário de dados e modelo conceitual corporativo, devem informar precisamente a relação entre o atributo e a sua codificação para estabelecer contexto e linguagem padrão entre os envolvidos no projeto.

2.4.3.3. Planejamento amostral e coleta de dados

Esta é uma das etapas mais importantes no desenvolvimento de um modelo. O modelo será a imagem da sua experiência retratada na amostra. Presume-se que o comportamento passado se repetirá no futuro. Para tanto, os créditos e os proponentes analisados para a criação do modelo devem ser os mais próximos possíveis aos créditos e proponentes futuramente avaliados pelo modelo.

➤ **Quanto à base de dados**

A base de dados utilizada para seleccionar a amostra contém registros de clientes cujos créditos foram negados ou concedidos em um período de observação. Se o período de observação for muito antigo, as amostras podem não representar bem a realidade atual.

➤ **Seleção e o dimensionamento da amostra**

A amostra deve ser seleccionada de forma aleatória. São compiladas as informações disponíveis sobre os proponentes e o crédito. Assim como no caso de operações muito antigas, dados sobre operações muito recentes não são inseridos nas amostras.

São seleccionados dados sobre clientes que pagaram seus débitos regularmente, sobre clientes que se tornaram inadimplentes durante o período de concessão e, quando disponível, deve-se também tratar e analisar dados sobre clientes que foram rejeitados no acesso ao crédito. Estes podem ser incluídos como “maus clientes” assumindo que a decisão do analista de crédito que recusou a proposta foi perfeita. Outra estratégia é incluí-los no estudo apenas em uma segunda iteração, aplicando sobre ela a fórmula obtida pelo modelo e classificando-os como “pseudobons” ou “pseudomaus”. Nesta segunda iteração os “pseudoclientes” e os “pseudobons” assim como os “maus clientes” e os “falsos clientes” seriam agrupados para recalcular a fórmula do modelo baseada nestas novas amostras. Geralmente o número de “pseudobons” é pequeno, a menos que a análise tradicional de aprovação da instituição seja de baixa qualidade.

O tamanho da amostra depende de vários fatores, como o número e o tipo de variáveis que serão analisadas. Normalmente a amostra é dividida em dois grupos: um para cálculo do modelo e um para verificação do mesmo.

Apesar da eficiência do uso de amostragens, é prática usual no desenvolvimento de modelos de EC a seleção de censos, ou seja, todos os clientes durante o período de observação. Isto implica em maior demanda por recursos técnicos e computacionais.

➤ **Quanto aos aspectos operacionais**

Normalmente as informações estão dispersas em vários sistemas e arquivos espalhados pela instituição. Isto requer uma iniciativa para reunir, tratar e validar os dados utilizados na criação do modelo e consome grande parte do esforço humano e computacional de todo o projeto.

Os Metadados como os presentes em um Modelo de Dados Corporativo, com todas as entidades de negócio documentadas e um rígido processo de cadastro e validação das informações dos clientes são os maiores facilitadores para este processo, mas são raramente implementados pelas instituições.

2.4.3.4. Determinação da fórmula de escoragem

A primeira tarefa para a determinação da fórmula de escoragem é a análise e o tratamento das variáveis potenciais escolhidas pelos analistas. Além dos aspectos técnicos, esta análise permite a familiarização com os dados e uma maior sensibilidade para o desenvolvimento do sistema.

➤ **Análise e o tratamento das variáveis**

As variáveis potenciais escolhidas anteriormente são analisadas com três objetivos:

- Identificar eventuais inconsistências. Por exemplo, um indivíduo com vinte e dois anos de idade dificilmente teria mais de dez anos de emprego;
- Análise de informações indisponíveis. A interpretação e o tratamento para ausência de valores podem ser diferenciados entre as variáveis do modelo;
- Detectar a presença de valores discrepantes que, mesmo sendo reais, poderiam comprometer a estimativa dos pesos das variáveis.

Técnicas estatísticas são freqüentemente utilizadas para selecionar, filtrar e agrupar as variáveis. Um exemplo de técnica empregada é a análise bivariada, que é uma análise exploratória da positividade, negatividade ou neutralidade de uma variável (ex.: “casado”, “solteiro”, “divorciado” ou “viúvo”), quanto a sua tendência a discriminar a variável alvo (“bom” ou “mau” cliente). É usada para agrupar duas ou mais características com tendências semelhantes em uma só variável, aumentando sua significância.

➤ **Determinação da fórmula de escoragem**

Segundo Lewis (1992), a EC é uma técnica estatística de pontuação dos proponentes de crédito, tendo por base as informações cadastrais ou de comportamento do cliente e visando à segregação de características que permitam distinguir os bons dos maus riscos de crédito.

As técnicas estatísticas como a análise discriminante, a regressão linear e a logística, entre outras, podem ser utilizadas para determinar os pesos das variáveis. Vários softwares baseados em cálculos estatísticos disponíveis no mercado realizam esta tarefa e são capazes de selecionar as variáveis que participam da fórmula do modelo. Modelos preliminares podem

ser desenvolvidos até mesmo através de planilhas eletrônicas. Diferentes critérios e métodos devem ser experimentados para se chegar até a fórmula final.

Pode-se aplicar a Análise Discriminante, uma técnica multivariada usada na resolução de problemas que envolvem a separação de conjuntos distintos de objetos, ou observações, e a alocação de novos objetos, ou observações em um grupo específico. A Análise Discriminante integra o conjunto de técnicas usadas no reconhecimento de padrões.

A figura 10 representa dois grupos que se interceptam em dois eixos que representam valores comuns de duas de suas variáveis. Os grupos podem ser distintos projetando os seus membros na equação discriminante.

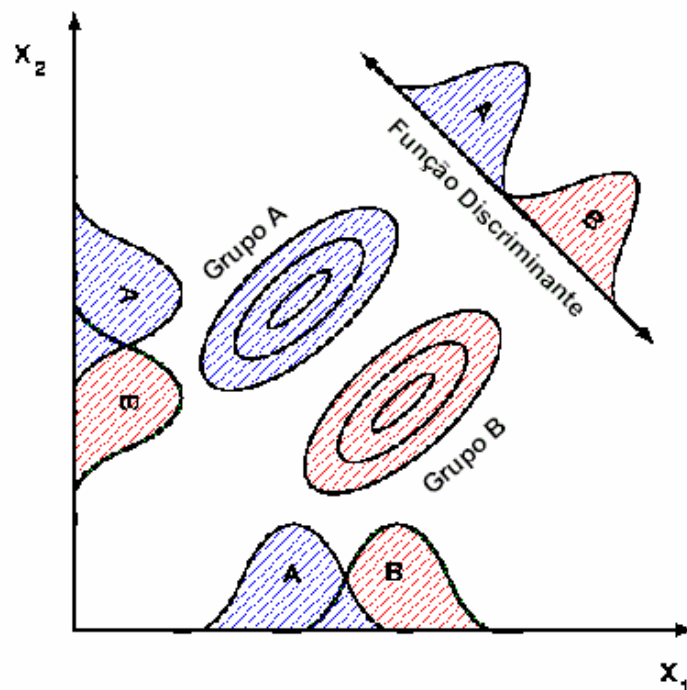


Figura 10 – Representação de uma análise discriminante

Fonte: Davis (1986)

Um dos objetivos da Análise Discriminante é determinar a que grupo, dentre dois ou mais definidos *a priori*, pertence um novo elemento, com base em características observadas para o mesmo, evitando a superposição observada nos eixos da figura 10. Os dados dos elementos de cada um dos grupos são coletados e procura-se derivar uma função linear para melhor discriminar os grupos entre si. O resultado almejado é a obtenção de um conjunto único de coeficientes para cada uma das variáveis independentes e que classifique, com a máxima precisão, cada elemento observado em um dos grupos previamente definidos.

De maneira análoga, a análise de regressão logística utiliza a base de dados para prever a inadimplência, assumindo que a probabilidade de inadimplência segue uma distribuição logística, com resultado binomial, 0 ou 1, e não uma distribuição normal como é preconizado pela análise multivariada. Vários estudos foram realizados sobre a aplicação de ambos os métodos, e ambos produziram classificações semelhantes em termos de classificação de “bons” e “maus” clientes. A figura 11 mostra exemplos de curvas características da distribuição normal e logística.

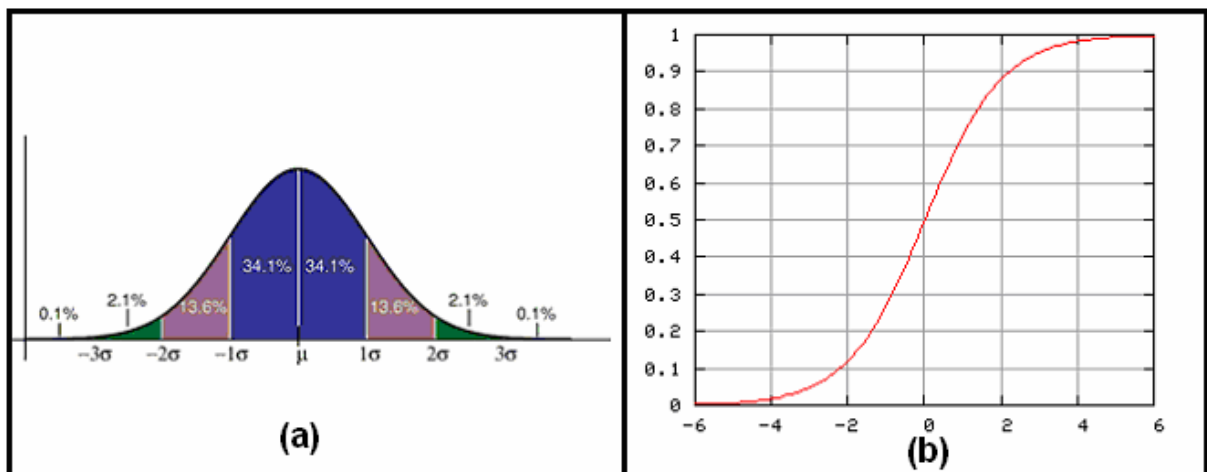


Figura 11 – Curvas características da distribuição normal (a) e logística (b)

Fonte: Wikipedia contributors (2006c e 2006b)

O detalhamento das técnicas estatísticas para a criação do score não é objetivo deste trabalho, assim como também não se objetiva aqui o detalhamento dos algoritmos existentes para MD. A intenção é listar a contribuição de ambos nas respectivas etapas para descoberta de conhecimento.

Vários analistas experientes com a análise tradicional de crédito sentem desconforto com a ausência de algumas variáveis na fórmula do modelo ou pela presença de algumas outras, algumas vezes com elevado peso, que anteriormente não mereciam tanta atenção. A não explicitação de alguma variável na fórmula não significa que ela não é ou deixou de ser importante: significa que outras variáveis já incluídas na fórmula contêm direta ou indiretamente a informação representada por esta variável. Os analistas de crédito podem experimentar trocar algumas das variáveis na criação do modelo e, se isso não aumenta a eficiência do mesmo, pode pelo menos chegar a uma fórmula equivalente e melhor aceita pelos analistas de crédito.

2.4.3.5. Determinação do ponto de corte

O ponto de corte representa a pontuação que divide a população entre aprovados e rejeitados. Proponentes com pontuação superior ao ponto de corte seriam aprovados e os outros seriam rejeitados. As propostas submetidas à avaliação têm seu score calculado a partir dos dados do cliente e, em alguns casos, da operação.

Saunders (1999) afirma que o score obtido da equação do modelo de EC pode ser interpretado como a probabilidade de inadimplência, ao se comparar a pontuação de um crédito qualquer com determinada pontuação estabelecida como ponto de corte ou pontuação mínima aceitável. Neste sentido o score também pode ser utilizado para a classificação dos

créditos entre bons e maus, desejáveis e rejeitados conforme seja superior ou inferior àquela pontuação.

O ponto de corte determina a agressividade da instituição frente ao mercado. Instituições mais conservadoras trabalham com pontos de cortes mais altos, com menores perdas potenciais, menores taxas de aprovação e, conseqüentemente, menor risco. O ponto de corte permite controlar quantidade (volume) e qualidade (inadimplência e lucratividade) do estoque de operações da instituição.

O uso de EC pode incorrer em dois tipos de erros:

- ERRO I: recusar uma operação que seria lucrativa;
- ERRO II: aprovar uma operação que será problemática.

Existem três métodos para estabelecimento do ponto de corte:

- Por taxa de aprovação;
- Por taxa de “maus”;
- Por rentabilidade da carteira.

O primeiro método limita a possibilidade de incorrer no ERRO I. É o método mais simples e conservador. A relação entre o percentual de clientes aprovados e um valor mínimo de escore faz com que a proporção entre “bons” e “maus” clientes aumente, elevando a lucratividade da carteira.

O segundo método também não é muito sofisticado e fixa a taxa de maus clientes aceitável (ERRO II). Este método nem sempre significa alta rentabilidade, já que esta depende de fatores como receita e perdas por conta.

O terceiro método é o mais sofisticado e permite maximizar a relação risco x lucro. Para calcular a rentabilidade da carteira são utilizados dados corretos sobre as operações dos clientes. As seguintes informações podem ser utilizadas para calcular a rentabilidade da carteira de clientes:

- Receitas geradas pelas contas “boas” e “más” (até mesmo um “mau” cliente pode gerar receitas, através do pagamento de juros antes da inadimplência completa);
- Perdas de contas más;
- Despesas do produto estudado;
- *Spreads* praticados.

O objetivo é calcular a rentabilidade de cada escore, classificando-os em rentáveis e não rentáveis. O ponto de corte pode então ser fixado no primeiro valor de escore que possuir resultado positivo. O quadro 1 contém dados hipotéticos para auxílio na determinação do ponto de corte:

Faixas de Pontuação	Boas Clientes				Más clientes				Total de Clientes				Taxa de aprovação	Taxa de más		Margem de contribuição
	Na faixa		Acumulado		Na faixa		Acumulado		Na faixa		Acumulado			Faixa	Total	
	\$	%	\$	%	\$	%	\$	%	\$	%	\$	%		%	%	
151-155	243	1,77%	243	1,77%	53	5,44%	53	5,44%	296	2,01%	296	2,01%	100,00%	17,91%	6,40%	(57.005,25)
156-160	98	0,71%	341	0,71%	21	2,15%	74	7,59%	119	0,81%	415	2,82%	97,99%	17,65%	6,37%	(31.612,04)
161-165	200	1,46%	541	1,46%	36	3,69%	110	11,28%	236	1,61%	651	4,43%	97,18%	15,25%	6,31%	(41.833,49)
166-170	363	2,64%	904	2,64%	48	4,92%	158	16,21%	411	2,80%	1062	7,22%	95,57%	11,68%	6,16%	(21.630,50)
171-175	642	4,68%	1546	4,68%	67	6,87%	225	23,08%	709	4,82%	1771	12,04%	92,78%	9,45%	5,99%	(27.509,24)
176-180	1142	8,32%	2688	8,32%	138	14,15%	363	37,23%	1280	8,71%	3051	20,75%	87,96%	10,78%	5,80%	(51.799,51)
181-185	1728	12,59%	4416	12,59%	153	15,63%	516	52,92%	1881	12,79%	4932	33,54%	79,25%	8,13%	5,25%	56.234,97
186-190	2138	15,57%	6554	15,57%	155	15,90%	671	68,82%	2293	15,59%	7225	49,14%	66,46%	6,76%	4,70%	153.676,02
191-195	2212	16,11%	8766	16,11%	133	13,64%	804	82,46%	2345	15,95%	9570	65,08%	50,86%	5,67%	4,06%	231.882,18
196-200	2065	15,04%	10831	15,04%	86	8,82%	890	91,28%	2151	14,63%	11721	79,71%	34,92%	4,00%	3,33%	322.544,33
201-105	1508	10,98%	12339	10,98%	52	5,33%	942	96,62%	1560	10,61%	13281	90,32%	20,29%	3,33%	2,85%	298.290,70
206-210	1390	10,12%	13729	10,12%	33	3,38%	975	100,00%	1423	9,68%	14704	100,00%	9,68%	2,32%	2,32%	345.699,86
Total	13729	100,00%	13729	100,00%	975	100,00%	975	100,00%	14704	100,00%	14704	100,00%		6,63%		

Quadro 1 – Exemplo de escore para definição do ponto de corte.

Fonte: Herszkowicz (1998).

No quadro 1, a coluna da taxa de aprovação equivale a todos os clientes existentes menos os situados nas faixas acima do ponto de corte estipulado. Segundo Herszkowicz (1998), se o

ponto de corte for estabelecido em 176 pontos, 87,95% dos clientes seriam aprovados; para atingir uma taxa de maus máxima de 5%, o ponto de corte seria estabelecido em 186; baseado na rentabilidade o ponto de corte poderia ser fixado em 181, que é a primeira faixa com margem de contribuição positiva.

➤ **Faixas de escores**

As faixas de escores constituem classes de risco. Agrupando os escores em três classes, com alto escore (e baixo risco), médio escore e baixo escore (e alto risco), o sistema pode automaticamente rejeitar propostas para a terceira classe e aprovar propostas para a primeira classe. As propostas da segunda faixa de escore poderiam ser enviadas para um comitê para análise mais criteriosa, o que é recomendável durante a implantação e estabilização de um Sistema de EC.

A tendência é trabalhar com um mínimo de dez faixas de escore para melhor classificação dos proponentes. É necessário avaliar e estabelecer os limites de risco dentro de cada faixa e, quando possível, estabelecer os limites pela rentabilidade esperada de cada faixa.

2.4.3.6. Determinação das regras de decisão

Algumas regras existentes na instituição são aplicadas na decisão de crédito mesmo após a definição do escore e em qual classe de risco um proponente se encaixa. Algumas restrições, como idade mínima do proponente, podem vetar uma operação mesmo se o proponente estiver classificado em uma baixa faixa de escore.

Estas restrições são filtros não necessariamente baseados em critérios estatísticos, mas baseados em normas adotadas pela instituição. Alguns deles se devem à experiência da

instituição com casos anteriores. A instituição deve constantemente avaliar a pertinências destes filtros para que a perpetuação dos mesmos não restrinja a realização de bons negócios. O relaxamento destes filtros pode ser avaliado e testado no novo sistema, visando melhorar o desempenho do processo.

2.4.3.7. Validação do sistema de escoragem de crédito

O sistema deve ser testado para verificar sua eficiência. Amostras de testes das operações concluídas devem ser utilizadas para classificação entre “boas” ou “más”. É recomendável analisar seu acerto para segmentos específicos de clientes ou operações, para verificar a correlação dos escores com certas variáveis e analisar eventuais resultados aberrantes. Estes podem sugerir a necessidade de criar fórmulas distintas para segmentos onde o modelo não teve desempenho desejado.

O uso de casos atípicos durante a validação deve ser bem interpretado. A resposta do sistema para estes casos pode não ser o valor real destas operações, pois o modelo foi desenvolvido aplicando técnicas estatísticas em uma massa de operações, e não analisando casos isolados.

2.5. Síntese do referencial teórico

Esta pesquisa teve seu referencial teórico concentrado nas metodologias para DCBD com MD segundo Fayyad et al (1996) e também nas metodologias para EC, expostas por Sicsú (1998). Ambas serviram base para a criação do modelo de apoio à decisão de crédito, considerando a descoberta de conhecimento e o ciclo de vida da informação.

O quadro 2 relaciona as duas metodologias, traçando um paralelo entre suas etapas.

Desenvolvimento de um Sistema de EC por Sicsú (1998)	DCBD por Fayyad; Piatetsky-Shapiro; Smyth (1996)
Planejamento e Definições	Identificação do Domínio
	Escolha da Função de MD
Identificação das Variáveis potenciais	Seleção e Limpeza dos Dados
Planejamento Amostral e Coleta de Dados	Redução do volume de Dados
Determinação da Fórmula de Escoragem	Escolha do algoritmo de MD
Determinação do Ponto de Corte	Busca por padrões
Validação do Sistema de EC	Avaliação do Conhecimento
Determinação das Regras de Decisão	Refinamento do Conhecimento

Quadro 2 – Relação entre as metodologias para DCBD, MD e EC

Ambas metodologias baseiam-se no trabalho sobre os dados registrados sobre os clientes e suas operações, extraíndo informações e conhecimento útil para a instituição. Em ambas metodologias os dados analisados devem ser selecionados e tratados para a obtenção de amostras válidas. Os modelos resultantes são validados através da comparação das regras descobertas com amostras de validação.

As metodologias possuem caráter descritivo e preditivo: o escore criado, por si só, representa uma descrição das operações e serve como base para a previsão da capacidade de pagamento dos proponentes e do resultado da instituição. Da mesma forma, com o emprego de métodos de MD, como a árvore de decisão, os dados registrados são processados para extração e

apresentação ¹¹ (caráter descritivo) do conhecimento requisitado, que consiste no relacionamento da variável dependente com as outras variáveis analisadas. Este conhecimento de fácil compreensão pode ser confrontado com os dados de um proponente para prever seu comportamento.

O conhecimento obtido nas duas metodologias significa um incremento no conhecimento do domínio, subsidiando a avaliação e eventual alteração das regras organizacionais para decisão.

Enquanto a metodologia para EC, que é considerada por diversos autores como a melhor alternativa para apoiar a decisão de concessão de crédito, define etapas específicas para atuar neste domínio, a metodologia para DCBD corresponde ao estudo de processos de armazenamento, extração, compartilhamento e aplicação de conhecimento de maneira mais ampla.

Existem algumas metodologias para o desenvolvimento de sistemas de apoio à decisão. Alguns aspectos são comuns a todas as metodologias. Deve ser criado um grupo de profissionais que têm por missão o acompanhamento da evolução do sistema. [...], tal sistema deverá ter a sua base de conhecimento avaliada periodicamente, bem como o seu desempenho, para adequação quanto às necessidades de recursos de *hardware* adicionais, ou mesmo para análise de sua eficácia. (MENDES, 1997, p.45)

Conclui-se então que o EC é um caso específico da DCBD, aplicada ao domínio financeiro, especificamente ao problema de se conceder crédito minimizando o risco de inadimplência e maximizando as receitas. Nesse domínio específico, a identificação das variáveis depende da experiência dos especialistas de crédito, a determinação do ponto de corte depende da análise do negócio (podendo estabelecer metas sobre rentabilidade, taxa de aprovação ou composição

¹¹ O resultado final apresentado difere entre as diferentes técnicas de MD. A árvore de decisão apresenta um resultado de fácil compreensão para o usuário final, enquanto técnicas como as redes neurais resultam em um modelo mais complexo de difícil visualização para o analista de negócio.

da carteira), com objetivo de classificar os proponentes. O uso de MD apoiada nos conceitos e definições da EC pode auxiliar no processo de decisão de crédito.

3. Criação do modelo de mineração de dados

3.1. Caracterização do ambiente pesquisado

A pesquisa é relativa a uma empresa do setor financeiro, com sede em Belo Horizonte (MG). Os dados analisados são relativos ao seu produto denominado “Crédito Pessoal”, que é um empréstimo garantido por cheques a ser pago de 3 (três) a 12 (doze) prestações. Foram realizadas entrevistas com um analista de negócios da instituição, para construção do conhecimento de domínio necessário. Como existe uma padronização no processo de aprovação da instituição, as indicações deste analista designado foram consideradas representativas deste processo.

Foi declarado que a política de crédito da instituição define limites de empréstimos entre R\$ 300,00 a R\$ 5.000,00 (trezentos reais a cinco mil reais) para o produto Crédito Pessoal. Foi exposto que o processo de aprovação de proponentes se restringe à satisfação de alguns pré-requisitos fixos, entre os quais (nem todos serão registrados nesta pesquisa por questões de confidencialidade):

- Consultas ao SPC e SERASA;
- Idade mínima de 20 (vinte) anos;
- Posse de conta corrente bancária com idade mínima de 6 (seis) meses;
- Permanência no emprego atual por no mínimo 1 (um) ano;
- Renda igual ou maior ao salário mínimo¹²;
- Comprometimento máximo de 25 % da renda com a soma do valor da prestação do empréstimo e da prestação de aluguel;

¹² O valor do salário mínimo era de R\$ 300,00 (trezentos reais) no período de coleta de dados para análise.

Conforme pesquisado sobre a política da maioria das instituições financeiras no Brasil, a empresa analisada também se apóia basicamente em consultas aos BIC, acrescentando ao processo a verificação de uma lista fixa, composta por pré-requisitos. Estes foram determinados por alguns analistas da instituição baseados nas suas experiências profissionais em pesquisas de mercado.

O processo muito simplificado não identifica os diferentes grupos de proponentes pelas suas características. A instituição possui então um cadastro de clientes com operações aprovadas e um cadastro de clientes recusados. A análise contínua do comportamento dos clientes, baseada em seus atributos (variáveis para análise), seria indicada para classificar e prever o relacionamento da instituição com os proponentes, conforme se pretende verificar nesta pesquisa.

3.2. Desenvolvimento do modelo

As etapas presentes no quadro 2, que compara os modelos para EC e MD, foram realizadas. A MD foi utilizada como ferramenta, direcionada para a extração do conhecimento buscado pelo modelo de EC pesquisado.

Quanto ao planejamento e definições, o estudo foi voltado para a extração do conhecimento para pré-aprovação de crédito, limitada à análise das informações cadastrais do cliente para a previsão de seu comportamento, não considerando os dados da operação requisitada. O produto em questão é o Crédito Pessoal, um empréstimo de quantias entre R\$ 300,00 a

R\$ 5.000,00, garantido por cheques com prazo de pagamento entre 3 e 12 prestações. O período de análise foi estabelecido entre julho de 2003 e julho de 2005. Os analistas da instituição consideraram este um bom período para análise, pois apresentou condições políticas e financeiras relativamente estáveis. Considera-se que a previsão do comportamento é válida para um período de até 12 meses.

Foram analisados os dados cadastrais de uma população de 15375 clientes que contrataram Crédito Pessoal no período que se inicia julho de 2003 e se encerra em julho de 2005.

O saldo das operações consideradas na análise é de R\$ 23.371.087,46. A maior parte (86,13%) dos empréstimos analisados situa-se abaixo de R\$ 2000,00 sendo que o valor médio das operações é de R\$ 1.182,09 e o prazo médio corresponde a 8,47 prestações. O montante de R\$ 20.518.291,82 é referente a contratos liquidados (87,88%) enquanto R\$ 2.852.795,64 são referentes a contratos ativos.

Para avaliar o risco de inadimplência, os indivíduos são diferenciados em grupos de acordo com o seu risco de crédito: é necessário chegar ao conceito de “bons clientes” e “maus clientes”. Foram realizadas entrevistas com um analista de negócios da instituição e diversas análises quanto aos saldos das aplicações registradas para chegar até a definição.

As faixas para reclassificação de risco de operações, presentes na Resolução do Conselho Monetário Nacional (CMN) 2682, de 21.12.1999, foram inicialmente adotadas pra maior padronização na definição das faixas de inadimplência dos clientes. Segundo esta resolução, os créditos são classificados em vários níveis – decrescendo de AA até H – de acordo com a percepção do risco de crédito do tomador, das garantias oferecidas e da modalidade

operacional. Uma variável-chave para esta classificação é o tempo de atraso das operações. Em cada um desses níveis é exigido um percentual de provisão¹³ que varia de 0% a 100% sobre o total dos créditos registrados naquela faixa. As classes de risco estão presentes no quadro 3:

Atrasos	Classificação	Provisão [%]
Nenhum dia	Risco AA	0,0
Até 14 dias	Risco A	0,5
Entre 15 e 30 dias	Risco B	1,0
Entre 31 e 60 dias	Risco C	3,0
Entre 61 e 90 dias	Risco D	10,0
Entre 91 e 120 dias	Risco E	30,0
Entre 121 e 150 dias	Risco F	50,0
Entre 150 e 180 dias	Risco G	70,0
Mais de 180 dias	Risco H	100,0

Quadro 3 – Classes de risco para provisionamento segundo a CMN 2682.

Fonte: Banco Central do Brasil (1999).

As instituições financeiras freqüentemente buscam a classificação de seus clientes nestas faixas de risco. A classificação é realizada de acordo com o maior atraso registrado pelo cliente no histórico de suas operações.

Definidas as faixas iniciais, o montante ativo e o montante liquidado da carteira foram calculados para cada uma delas, visando determinar o comportamento da carteira, conforme representado no quadro 4.

¹³ Esta pesquisa se restringe aos dados cadastrais dos clientes para a construção do modelo, portanto os valores das operações propostas e quaisquer cálculos de provisão estão fora de seu escopo.

Faixa de risco	Montante ativo	Montante Liquidado	Operações
AA	R\$ 0,00	R\$ 2.544.003,52	2529
A	R\$ 101.589,48	R\$ 6.978.853,45	5730
B	R\$ 110.717,70	R\$ 3.209.870,44	2861
C	R\$ 109.150,36	R\$ 2.824.285,01	2563
D	R\$ 74.112,80	R\$ 1.086.392,52	1009
E	R\$ 101.397,78	R\$ 702.666,24	666
F	R\$ 136.024,76	R\$ 511.264,99	533
G	R\$ 177.138,45	R\$ 437.268,50	470
H	R\$ 2.042.664,31	R\$ 2.223.687,15	3410

Quadro 4 – Distribuição da carteira em relação às faixas de risco

A figura 12 mostra a média dos saldos ativos, liquidados e o valor médio de juros captado em cada faixa de risco, considerando uma taxa de 6,46%.

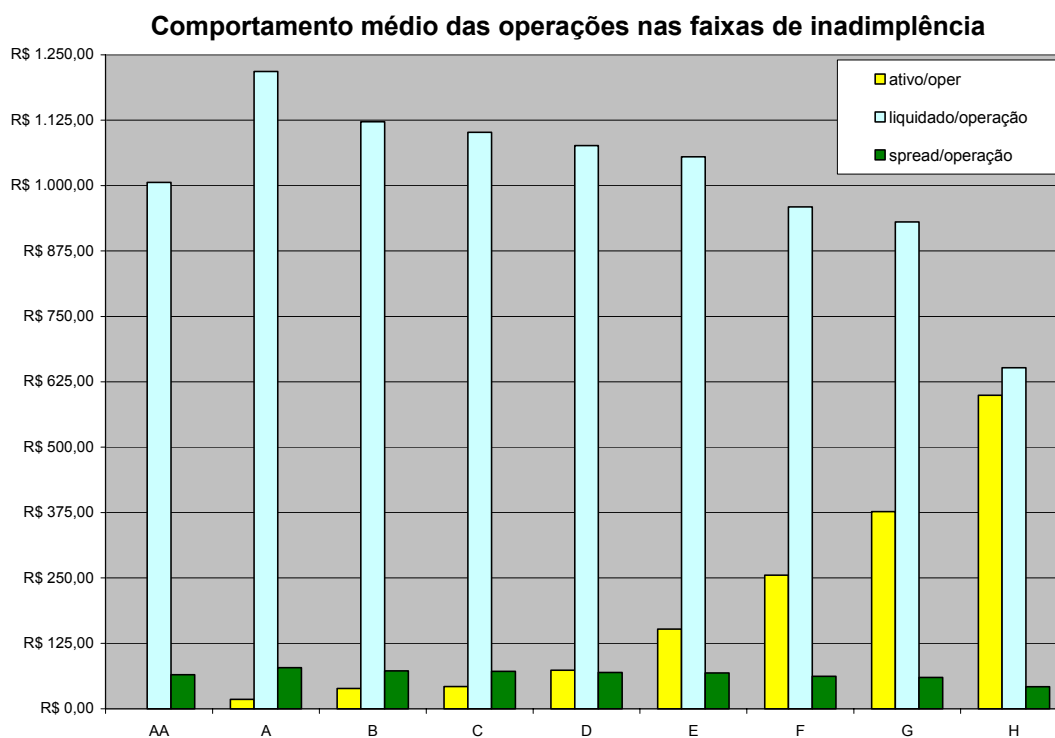


Figura 12 – Comportamento médio das operações nas faixas de inadimplência

O saldo médio das operações ativas aumenta nas faixas de risco mais elevado, enquanto o valor médio liquidado diminui. Esta informação reforça a expectativa de maior risco com clientes em faixas mais altas de máximo atraso observado. Buscou-se determinar qual era o atraso máximo observado, em dias, até o qual o cliente seria classificado como “bom cliente” e acima do qual o cliente seria classificado como “mau cliente”.

Da análise das informações levantadas obteve-se que a divisão dos juros captados em cada faixa pelo respectivo saldo em operações ativas, resulta em um indicador indireto de liquidez e qualidade da carteira *versus* a inadimplência que foi considerado satisfatório pelo analista da instituição. A distribuição da taxa juros/ativo está apresentada na figura 13.

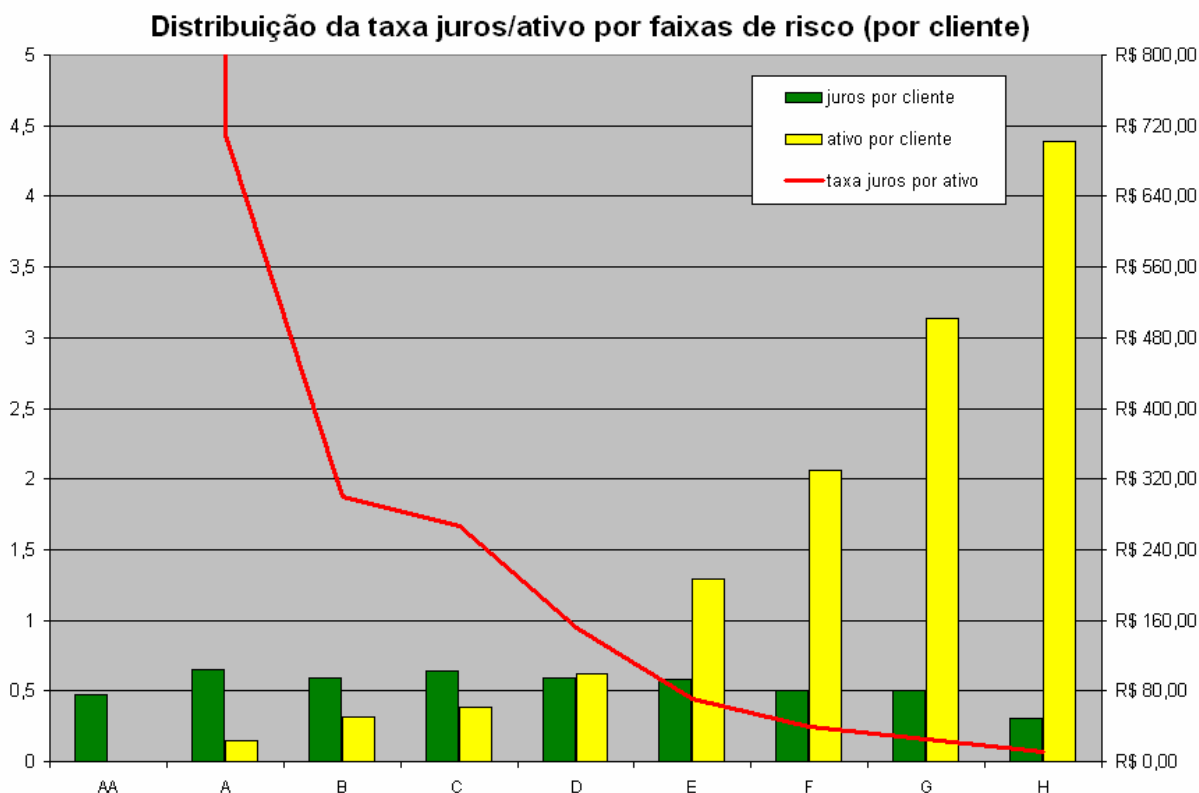


Figura 13 – Distribuição da taxa juros/ativo por faixas de risco

Observou-se que até a faixa de risco C - clientes com até 60 dias de atraso – o indicador obtido é superior a uma unidade, valor alvo definido pelos analistas da instituição para indicar boa liquidez na faixa.

Para refinar o conhecimento obtido acima, foram projetados os valores deste indicador para cada dia de atraso observado e optou-se por criar uma linha de tendência utilizando média móvel em cinco períodos, de acordo com a figura 14:

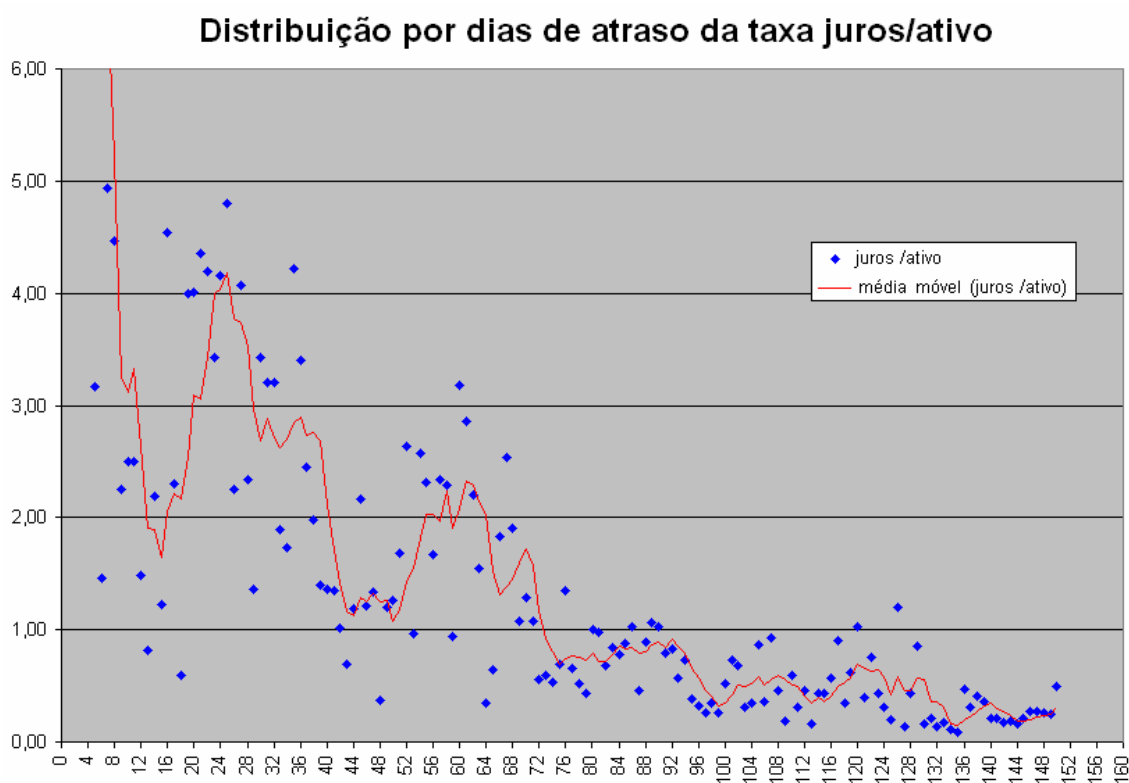


Figura 14 – Distribuição diária da taxa juros/ativo

O gráfico na figura 14 mostra que a linha de tendência escolhida projeta o atraso máximo com o indicador juros/ativo até 72 dias. Este é um novo conhecimento de domínio, que é

incorporado ao processo completo de descoberta do risco de crédito definindo “bons clientes” os que apresentam atraso máximo registrado em 72 dias. Este critério está apresentado, resumidamente, no quadro 5.

Tipo de Cliente	Tempo de Atraso em Qualquer Prestação
Cliente bom	Atraso até 72 dias
Cliente mau	Atraso maior que 72 dias

Quadro 5 – Critério para a classificação dos clientes

Para se selecionar os bons e os maus clientes, portanto, seguem-se dois passos: primeiro, verifica-se no histórico do cliente o número de dias do seu maior atraso em todos os contratos do período analisado e, segundo, agrupam-se os clientes entre “bons” e “maus” conforme seu histórico de maior atraso tenha sido inferior a 72 dias ou superior a esta marca, respectivamente. Assim, por exemplo, se um cliente contratou 5 (cinco) operações no período analisado e pagou em dia todas as prestações à exceção da última parcela do último contrato, na qual ficou inadimplente por 80 (oitenta) dias, este cliente será incluído no grupo dos maus clientes, pois apresentou atraso superior a 72 dias no período analisado.

A distribuição dos valores na carteira de acordo com o critério “cliente bom” ou “cliente mau” é apresentada no quadro 6:

	“Bons” Clientes	“Maus” Clientes	Total
Número de clientes	10801	4574	15375
Contratos ativos	R\$ 349.551,42	R\$ 2.503.244,22	R\$ 2.852.795,64
Contratos Liquidados	R\$ 16.057.550,37	R\$ 4.460.741,45	R\$ 20.518.291,82
Carteira total	R\$ 16.407.101,79	R\$ 6.963.985,67	R\$ 23.371.087,46

Quadro 6 – Distribuição dos empréstimos por tipo de cliente

Relacionando o quadro 6 com as informações de negócio fornecidas pelos analistas, é possível visualizar o nível de inadimplência da carteira, cerca de 30% do valor dos créditos, dada a proporção de maus clientes em relação ao total de saldos (R\$ 6.963.985,67/ R\$ 23.371.087,46). A proporção entre os maus clientes e o número total de clientes é de 42%, conforme apresentado no gráfico da figura 15.

Carteira de operações por tipo de cliente

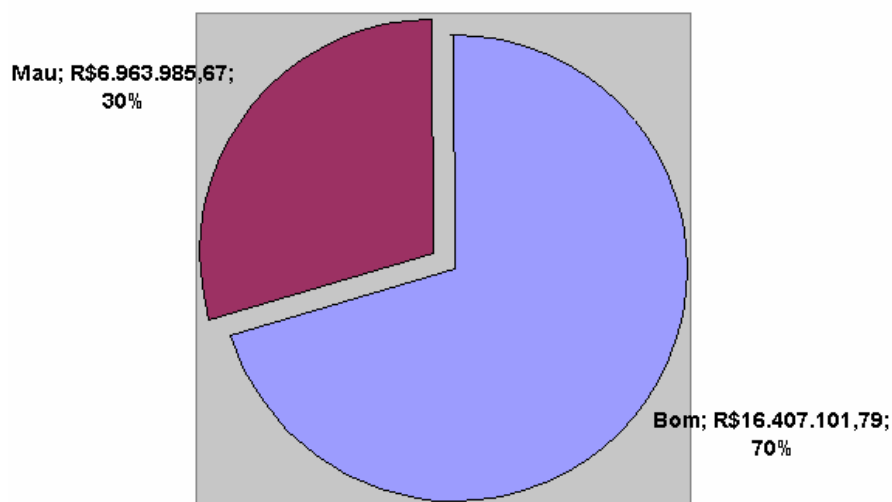


Figura 15 – Carteira de operações por tipo de cliente

Os 15375 clientes dividem-se na seguinte proporção em relação à variável sexo: são 8.481 (55,16%) mulheres e 6894 homens (44,84%), o que está disposto no quadro 7.

Sexo	Bom	% Bom	Mau	%Mau	Total	% Total
Feminino	6242	73.60%	2239	26.40%	8481	55.16%
Masculino	4559	66.13%	2335	33.87%	6894	44.84%
Total	10801	70.25%	4574	29.75%	15375	15375

Quadro 7 – Distribuição dos clientes por sexo

A idade média dos mutuários é de 45 anos, o que significa um quadro de clientes com idade mediana. A renda média destes clientes é de R\$ 949,49. É grande o número de clientes que declara não possuir dependentes menores de 18 anos, 13.887 (90,03%) observações, embora a proporção de casados, desquitados e viúvos na população alcance 70.54% casos, conforme o exposto na figura 16.

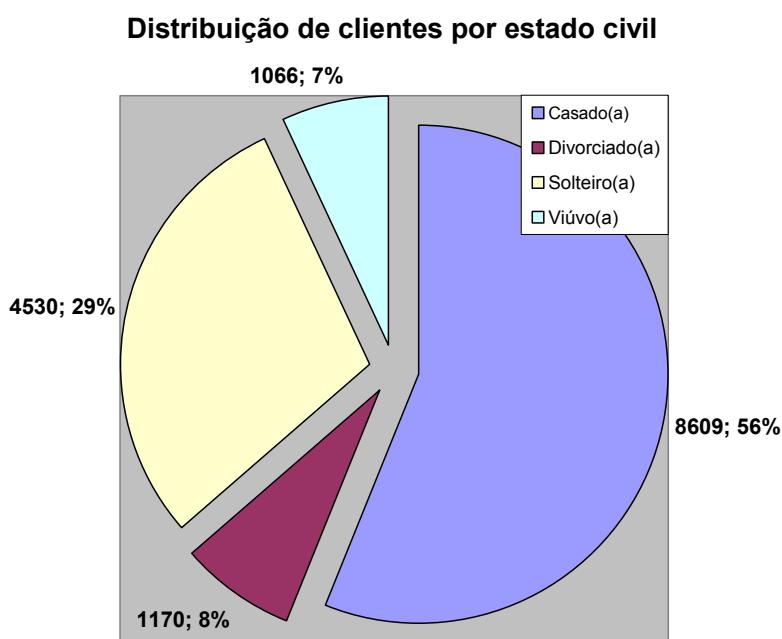


Figura 16 – Distribuição de clientes por estado civil

A instituição disponibilizou uma cópia da base e do seu dicionário de dados, contendo as suas entidades, os seus atributos e o significado de suas instâncias.

Embora o número total de mutuários registrados na instituição seja bem maior, a amostra foi restrita apenas aos clientes com créditos aprovados e com no mínimo 6 (seis) parcelas dentro do período entre julho de 2003 e julho de 2005, para que seus dados tenham

representatividade para a correta construção do modelo. Como os dados dos clientes reprovados não participam da sua construção, eles são usados em uma etapa posterior: a especulação final sobre a resposta do modelo em relação a esta população. Esta seria uma especulação do comportamento destes clientes no caso dos seus créditos terem sido aprovados. Os dados estão registrados em uma base de dados única, utilizada pelos sistemas que atendem aos operadores e analistas de crédito da instituição.

Foi arbitrado que todos os 15375 cadastros válidos seriam utilizados para criação e validação do modelo. Os registros foram divididos aleatoriamente na proporção de 69% para criação do modelo e 31% para validação posterior¹⁴.

O *software Microsoft Analysis Services 2000* foi a ferramenta escolhida para construção do modelo do MD. Esta escolha se deu porque esta é uma ferramenta de fácil acesso e uso simplificado para reprodução dos procedimentos adotados.

3.2.1. Construção da árvore de decisão

Foram criadas consultas dimensionais a um cubo OLAP para proporcionar acesso rápido às informações presentes na base de dados antes da construção da árvore de decisão. A execução de consultas dinâmicas pelos analistas apoiou o processo de seleção de variáveis para o modelo.

¹⁴ As referências sobre análise multivariada indicam que o pesquisador deveria usar de 50% a 70% para a construção do modelo. O método usado para distribuição dos clientes entre amostra para construção ou para a validação do modelo foi baseado no dígito verificador do CPF: os dígitos 0, 7 e 6 correspondem à amostra de validação do modelo. Os restantes correspondem à amostra de construção do modelo.

A figura 17 mostra o aplicativo para criação de cubos OLAP com as suas dimensões de análise e um exemplo de consulta que pode ser submetida.

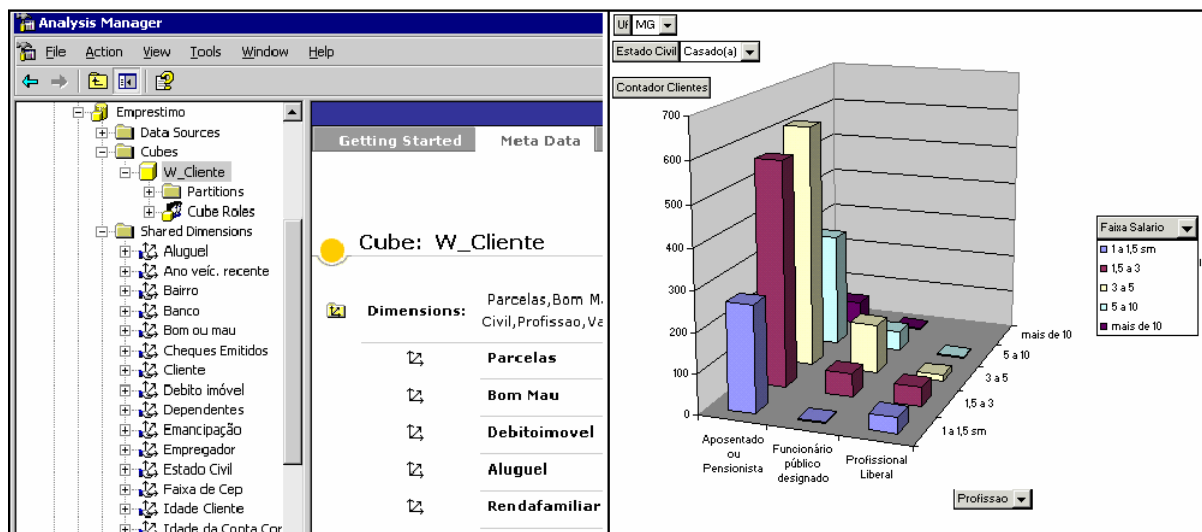


Figura 17 – Cubo OLAP criado e exemplo de consulta disponível

Com o apoio das informações extraídas do cubo OLAP e do seu conhecimento de domínio, o analista de negócio selecionou os seguintes atributos como variáveis de entrada para o modelo: Celular / Cidades / Estado / Estado Civil / Idade do cliente / Nível de instrução / Número de dependentes / Número de veículos / Profissão / Região / Renda Familiar / Salário / Sexo do Cliente / Tempo de emprego / Tipo de residência / Veículos.

O *Analysis Services* foi também utilizado para construção da árvore de decisão. O seu algoritmo de árvore de decisão realiza a regressão e a classificação usando os atributos informados. As variáveis são informadas na ferramenta, especificando quais são variáveis de entrada - as variáveis potenciais citadas anteriormente - e qual é o atributo a ser previsto: a classificação do cliente entre “bom” e “mau”. O algoritmo de árvore de decisão da ferramenta escolhida avalia os valores de cada uma das variáveis de entrada, calculando faixas de valores

necessárias (para variáveis numéricas como salário e valor do aluguel) , quais os pesos e a hierarquia das variáveis selecionadas e quais variáveis não influenciam o comportamento da variável a ser prevista. As variáveis de entrada foram fornecidas conforme a figura 18.

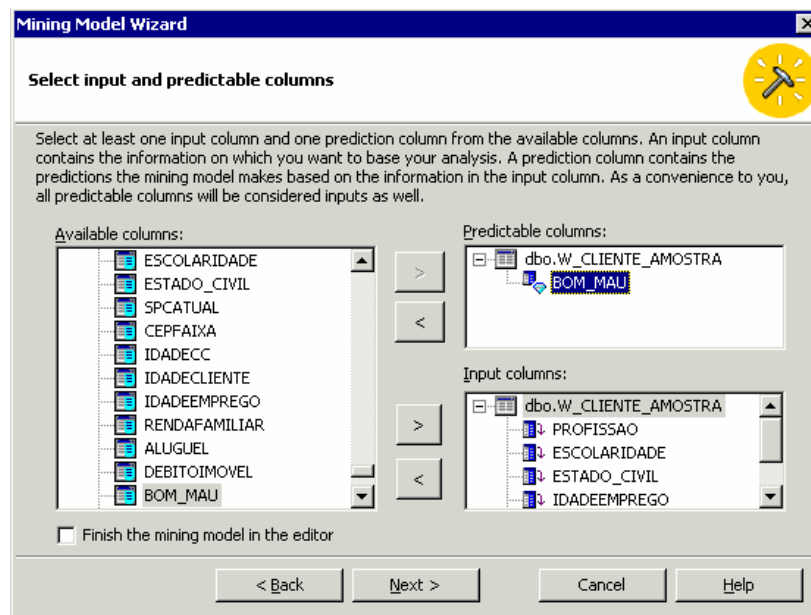


Figura 18 – Interface para criação da árvore de decisão

A Ferramenta cria o modelo de árvore de decisão e o processa, o que consiste em realizar o “treinamento” do modelo. Isto significa que o algoritmo insere os registros de dados no modelo, cria os nós da árvore de acordo com o peso dos relacionamentos de seus atributos e calcula as probabilidades para a variável de predição em cada nó. Isto é feito a partir dos registros de dados contidos na amostra. A figura 19 mostra os comandos de acesso a dados utilizados na criação do modelo.

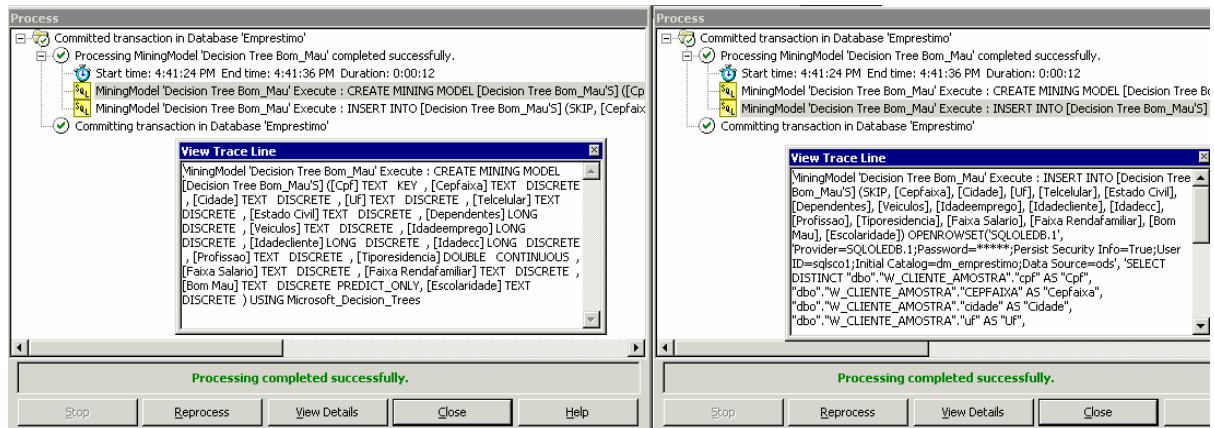


Figura 19 – Comandos utilizados para na criação e no treinamento do modelo

A árvore de decisão resultante pode ser visualizada através de criação, conforme a figura 20:

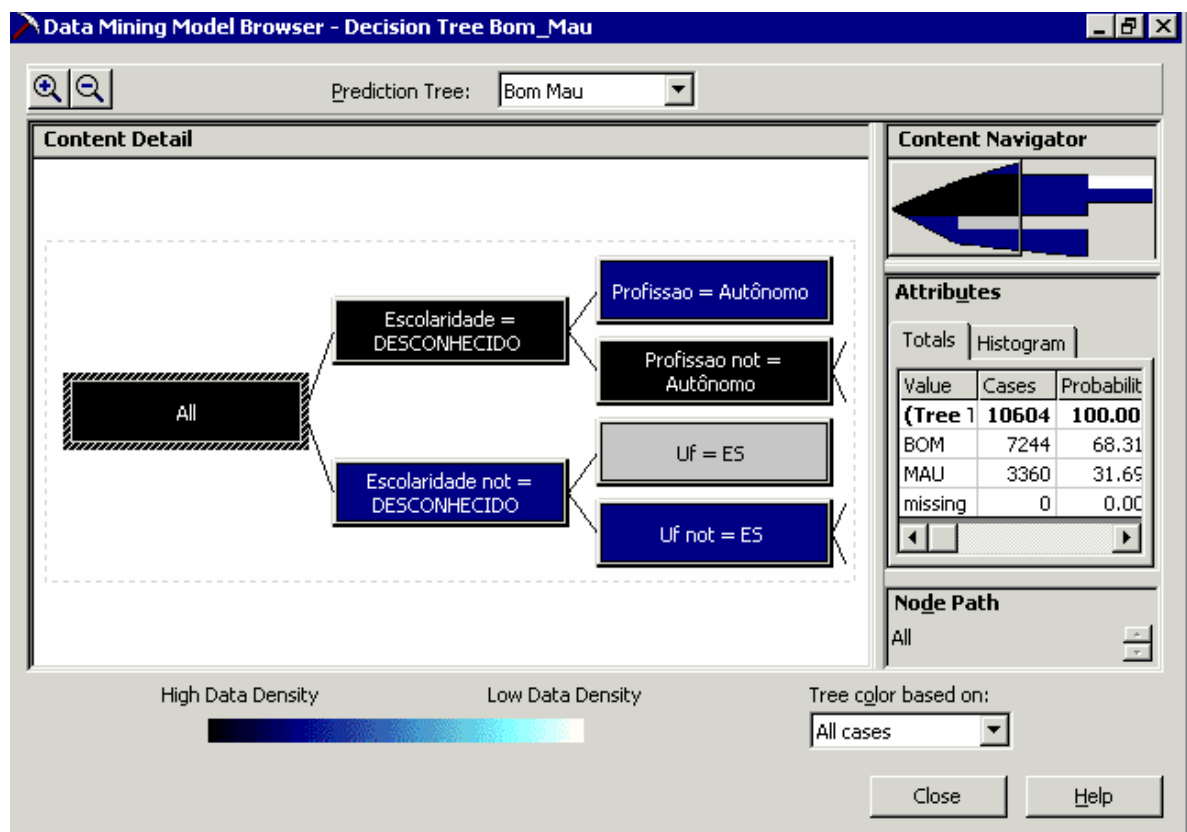


Figura 20 – Visualização da árvore de decisão

Na figura 20, a seleção está no primeiro nó da árvore, com todos os registros. Na porção superior à direita está o mapa para localização em árvores com muitos ramos. À direita no meio da tela está o indicador de proporção entre “bons” e “maus” clientes para cada ramo selecionado. Tem-se então que da amostra completa 31,69% dos clientes são “bons clientes”. Caminhando pela árvore é possível encontrar os nós onde existem diferenças acentuadas entre estas classificações. A criação de grupos hierarquicamente classificados de acordo com seus riscos de inadimplência em função das suas variáveis caracteriza o conhecimento extraído do modelo. Então aí se completa o ciclo de vida da informação, assim composto:

- **Dado** bruto contido nos registros de empréstimos e variáveis dos clientes;
- **Informação** extraída nas consultas dinâmicas ao cubo OLAP;
- **Conhecimento** obtido pela metodologia e pelo algoritmo de MD, realizando a classificação dos eventos na amostra de acordo com as variáveis informadas.

3.2.2. Predição da inadimplência dos proponentes

A partir do conhecimento representado na árvore de decisão foi possível realizar a predição do comportamento dos proponentes.

O processamento da árvore de decisão cria uma tabela cujo conteúdo expressa em valores percentuais as ocorrências da variável “bom” ou “mau” de acordo com as ocorrências das variáveis de entrada que tiveram pesos suficientes para serem expressas pelo algoritmo aplicado.

O processo de predição pela ferramenta adotada é uma simples consulta de união entre os registros no modelo “treinado” e os registros dos proponentes. A predição se dá pelo simples posicionamento dos novos registros nas posições da árvore de decisão.

O *Microsoft SQL Server 2000* possui um componente de extração e transformação de dados: o *Data Transformation Services* (DTS). O DTS trabalha com o conceito de pacotes de transformação, como o exposto na figura 21.

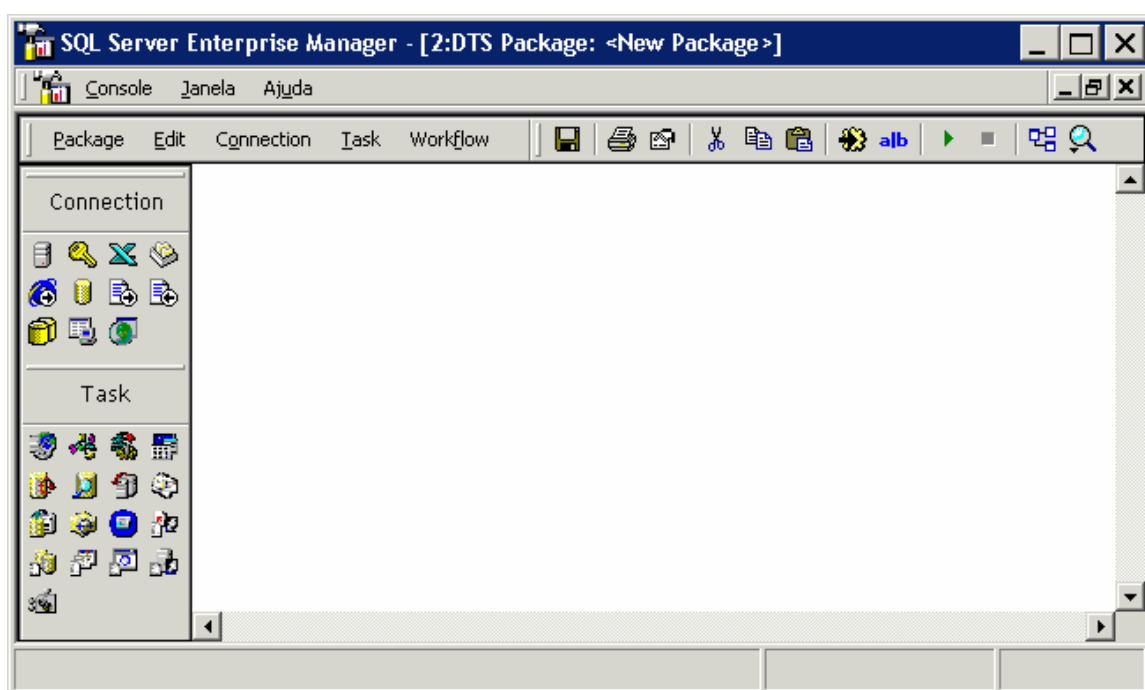


Figura 21 – Ferramenta para criação de pacotes DTS

Na figura 21 pode-se ver o grupo *Connection* (acima à esquerda), onde o analista informa conexões de entrada e saída, e o grupo *Task* (abaixo à esquerda), onde o analista seleciona e informa os parâmetros para as tarefas de transformação. Existem diversos tipos de conexões, como para acessos a arquivos texto, arquivos HTML, planilhas Excel, entre outros. Existem também diversos tipos de tarefas, como para execução de cópias de registros com alterações

em atributos, execução de comando SQL, processamento de cubos OLAP, envio de correio eletrônico e mineração de dados, entre outros.

O DTS possui um mecanismo para comparação dos registros dos proponentes com a árvore de decisão através de uma de suas tarefas de transformação de dados. Foi criado um pacote DTS com a tarefa de mineração de dados, como na figura 22.

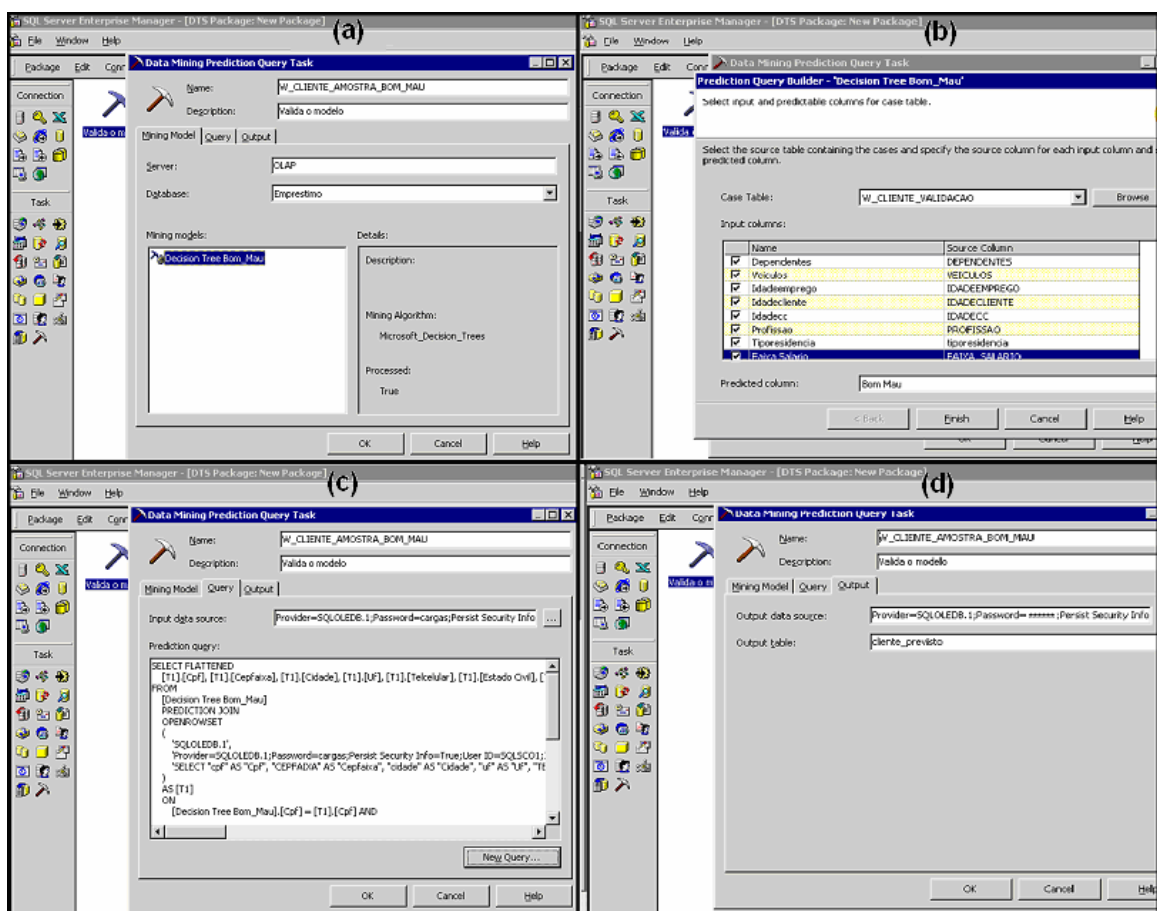
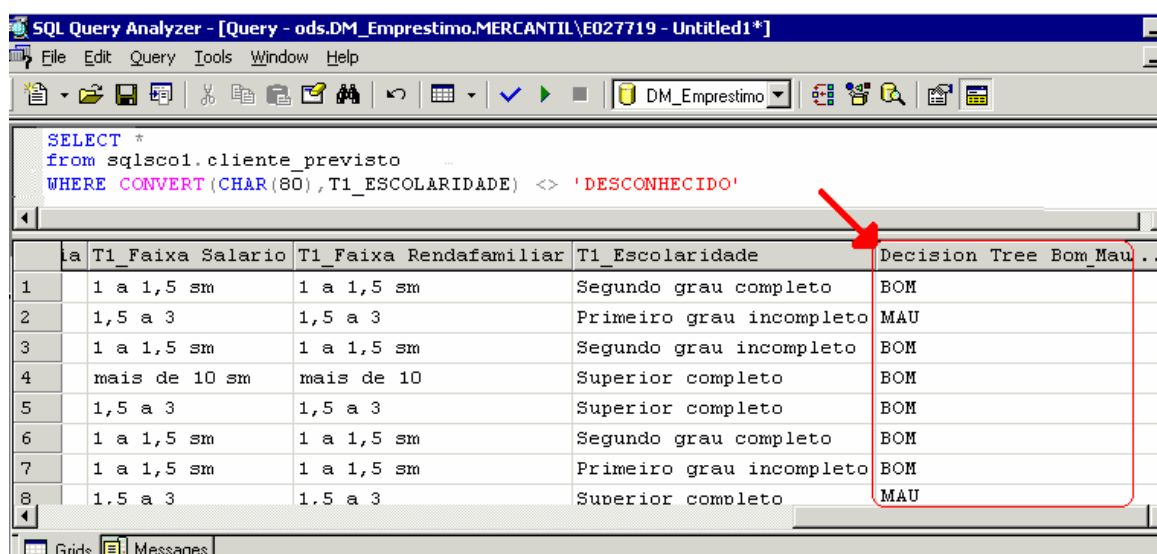


Figura 22 – Pacote DTS criado para a predição

Na figura 22 (a) está a identificação do servidor, da base de dados e da árvore de decisão escolhida. Na figura 22 (b) é informada qual é a *case table*, i.e., a tabela que contém os registros dos proponentes a serem previstos e é também informado o relacionamento entre

suas variáveis com as variáveis do modelo. A consulta de união e predição resultante aparece na figura 22 (c). É interessante notar na sua quinta linha a cláusula *PREDICTION JOIN*, usada neste tipo de consulta. Na figura 22 (d) está a indicação para a criação da tabela de saída, que contém uma cópia dos registros da *case table*, acrescidos do valor do atributo previsto indicado pelo conhecimento descoberto presente na árvore de decisão. Neste caso a resposta da variável prevista é binária (“bom” ou “mau”), de acordo com os percentuais verificados nos nós onde os registros de cada proponente se posicionam.

A figura 23 mostra o resultado da consulta à tabela de saída. A última coluna contém o valor da variável de saída.



SQL Query Analyzer - [Query - ods.DM_Emprestimo.MERCANTIL\E027719 - Untitled1*]

```
SELECT *
from sqlsc01.cliente_previsto
WHERE CONVERT (CHAR (80), T1_ESCOLARIDADE) <> 'DESCONHECIDO'
```

	T1_Faixa Salario	T1_Faixa Rendafamiliar	T1_Escolaridade	Decision Tree Bom_Mau ..
1	1 a 1,5 sm	1 a 1,5 sm	Segundo grau completo	BOM
2	1,5 a 3	1,5 a 3	Primeiro grau incompleto	MAU
3	1 a 1,5 sm	1 a 1,5 sm	Segundo grau incompleto	BOM
4	mais de 10 sm	mais de 10	Superior completo	BOM
5	1,5 a 3	1,5 a 3	Superior completo	BOM
6	1 a 1,5 sm	1 a 1,5 sm	Segundo grau completo	BOM
7	1 a 1,5 sm	1 a 1,5 sm	Primeiro grau incompleto	BOM
8	1.5 a 3	1.5 a 3	Superior completo	MAU

Figura 23 – Consulta à tabela com o atributo previsto

A taxa de classificação correta foi de 74%. O acerto na previsão de bons e maus clientes foi de 89% e 46%, respectivamente (erros tipo I e tipo II), o que foi considerado satisfatório. A identificação de 46% dos maus clientes atualmente aprovados pela instituição representaria

uma melhoria significativa na qualidade da carteira da organização. Marques (2002) afirma que outros trabalhos da espécie atingiram resultados similares.

O pacote DTS pode ser executado em tempo real pelo sistema de empréstimo ou pode ser agendado para execução em lotes. Assim a previsão obtida pela MD pode retornar ao ambiente operacional da instituição com o aumento do conhecimento, como representado na figura 24.

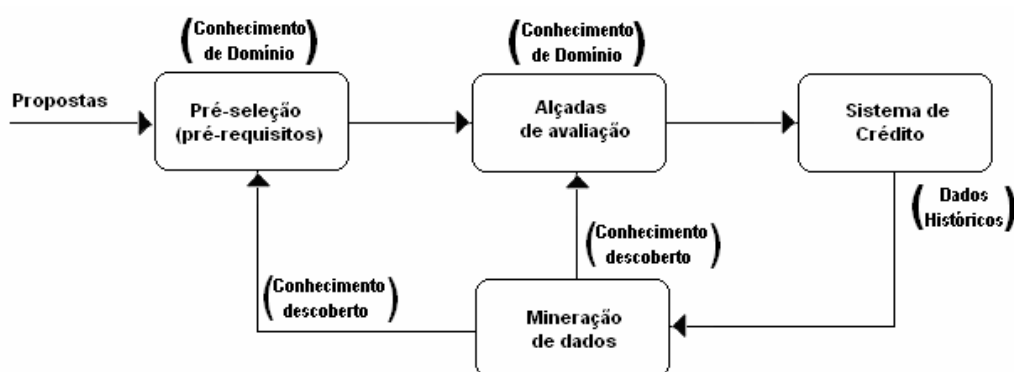


Figura 24 – Fluxo operacional com a adição do conhecimento descoberto

Os dados históricos representam a memória organizacional e a MD representa o processo de aprendizagem. O conhecimento descoberto e exposto pelo modelo de MD também significa um acréscimo no conhecimento de domínio. Os analistas podem avaliar a estrutura da árvore de decisão para avaliação de regras operacionais, como os atuais pré-requisitos para concessão de crédito.

O quadro 8 representa o resumo das etapas e definições para a implementação da MD sobre os dados pesquisados.

Definições para a criação do modelo de MD apoiado no modelo para Desenvolvimento de um Sistema de EC por Sicsú (1998)	
Planejamento e Definições	Pré-aprovação de crédito Empréstimo Pessoal 3 a 12 meses Máximo de 12 parcelas mensais Bom cliente com atraso até 72 dias Mau cliente com atraso superior a 72 dias
Identificação das Variáveis Potenciais	Celular Cidade Estado Estado Civil Idade do cliente Idade do emprego atual Nível de instrução Número de dependentes Profissão Região Renda Familiar Salário Sexo do Cliente Tempo de Emprego Tipo de Residência Veículos
Planejamento Amostral e Coleta de Dados	Período de 24 meses Apenas clientes com no mínimo 6 prestações no período Aproximadamente 70% para amostra de desenvolvimento Aproximadamente 30% para amostra de validação
Determinação da Fórmula de Escoragem	Ferramenta de árvore de decisão
Determinação do Ponto de Corte	Por percentual de bons clientes: os clientes que se posicionarem em nós com maioria simples de bons clientes serão automaticamente recomendados como bons clientes.
Validação do Sistema de EC	Testes com a amostra de validação
Determinação das Regras de Decisão	O conhecimento extraído é analisado pelos analistas de crédito para a contínua avaliação dos critérios de rentabilidade e de classificação dos clientes.

Quadro 8 – Resumo das definições para a implementação da MD

4. Conclusões

O processo decisório relativo à concessão de crédito é totalmente baseada em informações e em conhecimento. O aperfeiçoamento dos processos de apoio à decisão de crédito se beneficia do desenvolvimento e aplicação de processos de conhecimento nas instituições financeiras. A predição do risco de inadimplência possibilita redução dos custos, maximiza as receitas e proporciona maior disponibilidade de recursos para operações regulares.

Entretanto, por vezes as instituições não analisam seus registros de dados, que compõem parte da “memória organizacional”. Nela poderiam ser extraídas informações regionais e conhecimentos específicos sobre a prática de cada instituição.

Muitas instituições se limitam a consultas aos BIC que, pela dificuldade de gerenciar as informações regionais e específicas de cada instituição, não provêm ainda no Brasil o serviço de análise de crédito com detalhamento de informações necessário à tomada de decisão em análise de crédito, e se limitam a manter listas de crédito negativo.

As instituições costumam também definir regras para a seleção dos proponentes. Estas regras são impostas devido à incapacidade dos analistas em avaliar o risco exato de cada proponente, o que seria a solução ótima para a decisão de crédito. Desta forma, os analistas de crédito estabelecem critérios genéricos mínimos para a aprovação das propostas, estabelecendo o que seria a solução satisfatória para esta necessidade de informação. A instituição deve constantemente avaliar a pertinências destas regras operacionais para que a perpetuação das mesmas não restrinja a realização de bons negócios.

Ignorar o conhecimento potencial que pode ser extraído das bases de dados que contêm o registro de sua experiência é perder a chance de aprender sobre as melhores regras e critérios de decisão. Como a análise de grandes bases de dados utilizando métodos manuais é inviável, a MD sobre o histórico das operações é uma boa alternativa para a extração deste conhecimento organizacional.

No transcorrer deste trabalho, procurou-se compreender e estabelecer a relação entre os conceitos de DBCB, MD e EC, visando elucidar a contribuição destes ao apoio à decisão de crédito. A opção foi pela DCBD através de um modelo preditivo usando MD inspirado nos conceitos de EC, que é considerada por diversos autores como a melhor alternativa para apoiar a decisão de concessão de crédito, definindo regras e etapas específicas para atuar neste domínio.

O estudo dos conceitos de EC guiou a compreensão dos conceitos deste domínio específico de conhecimento, estabelecendo um suporte para a definição das etapas e definições realizadas na criação do modelo de MD. A apresentação dos resultados na árvore de decisão é diferente do resultado apresentado pela EC, que apresenta uma tabela de pontuações representando o risco de inadimplência em cada faixa. Entretanto, o conhecimento extraído do modelo de MD foi considerado satisfatório tanto para a descrição das operações analisadas através da visualização da árvore de decisão quanto para a previsão do seu risco de crédito através da metodologia escolhida. Foi comprovado que o conhecimento extraído significa um incremento no conhecimento do domínio, subsidiando a avaliação e eventual alteração das regras organizacionais para decisão.

Conforme afirmação anterior, conclui-se que a DCBD corresponde ao estudo de processos de armazenamento, extração, compartilhamento e aplicação de conhecimento de maneira mais ampla, o que leva à conclusão de que o EC é um caso específico da DCBD, aplicada ao domínio financeiro, especificamente ao problema de se conceder crédito minimizando o risco de inadimplência e maximizando as receitas. Nesse domínio específico, a identificação das variáveis depende da experiência dos especialistas de crédito, a determinação do ponto de corte depende da análise do negócio (podendo estabelecer metas sobre rentabilidade, taxa de aprovação ou composição da carteira), com objetivo de classificar os proponentes. Conclui-se também que o uso de MD, que se concentra nos processos de extração e aplicação da DCBD, apoiado nos conceitos e definições para EC pode auxiliar no processo de decisão de crédito. A Tecnologia da Informação pode apoiar este processo simplificando várias das tarefas como a seleção das variáveis de entrada, a extração dos padrões e a verificação da performance do modelo.

Foi concluído que a metodologia utilizada e a tecnologia que a apoiou caracterizam um processo de conhecimento. Neste processo os dados históricos representam a memória organizacional e a MD representa o processo de aprendizagem. O conhecimento descoberto e exposto pelo modelo de MD significa um acréscimo no conhecimento de domínio. Os analistas de crédito podem avaliar a estrutura da árvore de decisão para avaliação de regras operacionais, como os atuais pré-requisitos para concessão de crédito.

Conforme afirmação anterior, a qualidade do modelo depende da qualidade da base de dados. Quanto melhor a qualidade dos dados registrados, melhores as possibilidades do modelo atingir altas taxas de acerto.

Os sistemas especialistas e a mineração de dados são hoje uma realidade. Eles respondem a questões, fazem recomendações, e geralmente auxiliam o usuário orientando-o no processo de tomada de decisão. O conhecimento descoberto será mais bem utilizado por usuários familiarizados com o conhecimento do domínio trabalhado, pois estes usuários possuem o conhecimento tácito necessário para avaliar as respostas do sistema.

O caráter preditivo da MD é apenas uma indicação baseada nos dados históricos, e não substitui a intuição, o talento e a criatividade humana. Algumas situações podem ser reconhecidas mais rapidamente pelo julgamento humano do que pelo processo de DCBD. Por exemplo, a febre aftosa provocou, recentemente, no sul do país a demissão de milhares de funcionários em municípios com economias baseadas nas receitas de seus frigoríficos. Um analista bem informado conseguiria prever o aumento do risco de inadimplência naqueles municípios antes que o fenômeno começasse a adquirir relevância nos modelos de EC ou de MD, que se baseiam na análise do histórico de operações da instituição.

Uma das propostas de estudos futuros para continuidade do trabalho, é a avaliação do grau de acerto do modelo desenvolvido em função da qualidade dos dados analisados. Atividades como auditoria nas informações registradas no sistema origem, recadastramento visando corrigir dados errôneos e preencher atributos ausentes poderiam aumentar a qualidade dos dados na base origem. A sugestão é avaliar e procurando equacionar a relação entre a qualidade de dados no sistema origem e a eficiência do modelo de MD.

5. Referências Bibliográficas

- AGRAWAL, R; SRIKANT, R. **Fast Algorithms for Mining Association Rules**. IBM Almaden Research Center, 1994
Disponível em: <http://i2pi.com/PAPERS/analysis/agrawal94fast.pdf> . Acesso em: 01 jun. 2005.
- BANCO CENTRAL DO BRASIL, **RESOLUCAO No.2682**, 31 dez. 1999
Disponível em http://www.auditoriainterna.com.br/resol_2682.htm. Acesso em: 10 set. 2005.
- BARRETO, A. **A Eficiência Econômica e a Viabilidade de Produtos e Serviços de Informação**. Ciência da Informação - Vol 25, número 3, 1996.
- BORKO, H. **Information science: What is it?** American Documentation, v.19, n. 1, p.3-5; jan. 1968.
- CHAPMAN, P. et al. **The CRISP-DM Process Model**. 1999.
Disponível em: <http://www.crisp-dm.org>. Acesso em: 20 mai. 2005.
- CHEN, S Y; LIU, X, **The contribution of data mining to information science**. Journal of Information Science, 30 (6), pp.550-558, 2004.
- CHOO, C. W. **A organização do conhecimento**. São Paulo: Editora SENAC, 2003.
- DAVENPORT, T. H. **Ecologia da informação: por que só a tecnologia não basta para o sucesso na era da informação**. São Paulo : Futura, 1998.
- DAVENPORT, T. H; PRUSAK, L. **Conhecimento Empresarial: como as organizações gerenciam seu capital intelectual**. Rio de Janeiro: Campus, 1998.
- DAVIS, J. C. **Statistics and data analysis in geology**. New York. John Wiley and Sons, Inc.1986
- DUARTE, Â J. M. A. et al, **Reformas Microeconômicas e Crescimento de longo prazo**. Ministério da Fazenda Brasília, dez. de 2004. Disponível em http://www.fazenda.gov.br/spe/Texto_Versao%20Final%205.pdf .Acesso em: 02 jun. 2005.
- FAYYAD, U; et al. **From Data Mining to Knowledge Discovery in Databases**. In: FAYYAD, U. M. (Ed.). Knowledge Discovery and Data Mining. Menlo Park, California: AAAI Press: The MIT Press, p.01-54, 1996.
- FERREIRA, L Á. **O CRM sob a ótica da gestão de conhecimento: uma avaliação de soluções**. Dissertação-UFMG – Escola de Ciência da Informação, 2003.
- FRAWLEY, et al., **Knowledge Discovery in Databases: An Overview**. Ai Magazine, 1992 .Disponível em: <http://citeseer.ist.psu.edu/cache/papers/cs/12267/http://zSzzSzwwww.kdnuggets.comzSzgpubszSzaimag-kdd-overview-1992.pdf/frawley92knowledge.pdf> . Acesso em: 02 jun. 2004.

FURNIVAL, A C., **Delineando as limitações: sistemas especialistas e conhecimento tácito**. Ciência da Informação - Vol 24, número 2, 1995.

GALINDO & MILLER, **Can Credit Registries Reduce Credit Constraints? Empirical Evidence on the Role of Credit Registries in Firm Investment Decisions**. Preparado para o Seminário "Towards Competitiveness: The Institutional Path", Santiago, Março de 2001. Disponível em: http://www.aaep.org.ar/espa/anales/pdf_01/inv_galindo_miller.pdf. Acesso em: 06 jun. 2005,

GONZÁLEZ DE GÓMEZ, M.N. **Para uma reflexão epistemológica acerca da Ciência da Informação**. Perspect. cienc. inf., Belo Horizonte, v. 6, n. 1, p. 5 - 18, jan./jun., 2001

GUIMARÃES, I A. **Construção e avaliação de uma regra de reconhecimento e classificação de clientes de uma instituição financeira com base na análise multivariada**. Curitiba, PR. 2000. Dissertação Disponível em: http://www.cpgmne.ufpr.br/dissertacoes/D020_Inacio_Andruski_Guimaraes08122000.pdf. Acesso em: 10 de set. 2005

HERSZKOWICZ, F, **Credit Scoring: A Aplicação de Métodos Estatísticos na Avaliação do Risco de Crédito**. Trabalho de conclusão de curso, FEA - Universidade de São Paulo, 1998.

INMON, W.H.: **Building the Data Warehouse**. New York, Wiley, 1992,

KIM, D H., **The link between Individual and Organizational Learning**. Sloan Management Review, 1993

LAWRENCE, D B. **O Negócio de Crédito ao Consumidor – Risco e Recompensa**. São Paulo. Bandeirante S.A. Gráfica e Editora 1984.

LEWIS, E. **An Introduction to Credit scoring**. San Rafael, California. Fair Isaac and Co. Inc., 1992.

LOURENÇO, F C. L. **Vantagens do uso de métodos quantitativos no ciclo do crédito**. Revista Business – Equifax do Brasil, Abril de 2005. Disponível em http://www.equifax.com.br/rev_bus/05_abr/pag_pvi.asp. Acesso em: 05 de jun. 2005.

MARIO, P C. **Modelos de credit scoring: ferramentas para a decisão de crédito**. XIV Congresso Brasileiro de Contabilidade. Goiânia, GO. 2000. Disponível em: <http://www.milenio.com.br/siqueira/Tr114.htm>. Acesso em: 10 jun. 2005.

MARQUES, L F.B. **Gerenciamento do Risco de Crédito: cálculo do risco de crédito para a carteira de um banco de varejo**. PPGA - UFRGS, 2002. Dissertação disponível em: http://volpi.ea.ufrgs.br/teses_e_dissertacoes/td/000564.pdf. Acesso em: 07 jul. 2005

MENDES, R D. **Inteligência artificial: sistemas especialistas no gerenciamento da informação**. Ci. Inf., Brasília, v. 26, n. 1, p. 39-45, jan./abr. 1997.

MINTZBERG, H et al. **The Structure of "Unstructured" Decision Processes** *Administrative Science Quarterly*, Vol. 21, No. 2, p. 246-275, jun., 1976

Microsoft SQL Server 2005 Books Online Disponível em:
<http://msdn2.microsoft.com/en-us/library/ms174879.aspx> . Acesso em 10 nov 2006

MIRANDA, Roberto C. R., **O uso da informação na formulação de ações estratégicas pelas empresas**. Ci. Inf., Brasília, v. 28, n. 3, p. 286-292, set/dez, 1999.

NONAKA, I. E TAKEUCHI, H. **Criação do conhecimento na empresa**. Rio de Janeiro: CAMPUS, 1997

PEREIRA, Ana P. F., et al. **Uso do Credit Score na Análise de Crédito de Pessoa Física**. Niterói, RJ. 2002. Disponível em <http://www.latec.uff.br/anais/Artigos/38.pdf> . Acesso em: 10 jun; 2005.

PINHEIRO, A. C.; MOURA, A. **Segmentação e uso de informações nos mercados de crédito brasileiros**. Textos para discussão do BNDES. Rio de Janeiro, Fevereiro de 2001. Disponível em <http://www.finame.com.br/conhecimento/Td-88.pdf> . Acesso em: 02 jun; 2005.

Projeto Tamanduá - Manual de Ajuda. DCC/UFMG, Belo Horizonte, ago. 2006. Disponível em <http://tamandua.speed.dcc.ufmg.br> . Acesso em: 12 set. 2006.

REZENDE, S.O et al. **Mineração de dados. Em Sistemas Inteligentes: Fundamentos e Aplicações**, Volume 1, pp. 307-335, Barueri, SP. Editora Manole. 2003

SARACEVIC, T. **Ciência da Informação: Origem, evolução e relações**, Perspec. Ci. Inf., Belo Horizonte, v.1, n1. p41-62, jan/jun. 1996.

SAUNDERS, A. **Credit Risk Measurement: New approaches to Value at Risk and Other Paradigms**. New York: John Wiley & Sons, Inc. , 1999.

SENGER, L J; CALDAS JÚNIOR, J. **Análise de risco de crédito utilizando redes neurais artificiais**; Revista do CCEI / Universidade da Região da Campanha. v.5 n.8 URCAMP, 2001.

SCHRINCKEL, W. **Análise de Crédito: concessão e gerência de empréstimos**. São Paulo. Atlas, 1997.

SICSÚ, A L. **DESENVOLVIMENTO DE UM SISTEMA DE CREDIT SCORING**. São Paulo. Tecnologia de Crédito, jan/mar 1998.

SIMON, H A., **Administrative Behavior: a Study of Decision-Making Processes in Administrative Organization**. New York: Free Press, 1976.

SIRIHAL, A B.; LOURENÇO, C A. **INFORMAÇÃO E CONHECIMENTO: aspectos filosóficos e informacionais**. INFORMAÇÃO & SOCIEDADE: ESTUDOS, v.12 n.1, 2002.

SLOVIC, P. et al. **Behavioral Decision Theory**, Oregon: Annual Reviews Inc.1977.

THOMAS, L C. **Credit Scoring and Its Applications**. Society for Industrial & Applied, 2002.

VALENTIM, M.L.P., **Inteligência Competitiva em Organizações: dado, informação e conhecimento**. DataGramaZero - Revista de Ciência da Informação - v.3 n.4, ago., 2002.

VASCONCELLOS, M.S. **Proposta de método para análise de concessões de crédito para pessoas físicas**. S.P. 2002. Dissertação disponível em:
<http://www.teses.usp.br/teses/disponiveis/12/12138/tde-18112004-125442/>. Acesso em: 16 jun. 2005.

Wikipedia contributors. **Decision tree learning**. Wikipedia, The Free Encyclopedia. 2006a. Disponível em: http://en.wikipedia.org/w/index.php?title=Decision_tree_learning&oldid=86085103. Acesso em: 10 jun. 2006.

Wikipedia contributors. **Logit**. Wikipedia, The Free Encyclopedia, 2006b. Disponível em: <http://en.wikipedia.org/w/index.php?title=Logit&oldid=79109933>. Acesso em: 10 jun. 2006.

Wikipedia contributors. **Normal distribution**. Wikipedia, The Free Encyclopedia, 2006c. Disponível em: http://en.wikipedia.org/w/index.php?title=Normal_distribution&oldid=86853936. Acesso em 10 jun. 2006

WERSIG, G. **Information science: the study of postmodern knowledge usage**. Information Processing & Management, v.29, n2, 1993.