CRISTIANE KARCHER

REDES BAYESIANAS APLICADAS À ANÁLISE DO RISCO DE CRÉDITO

São Paulo
2009

CRISTIANE KARCHER

REDES BAYESIANAS APLICADAS À ANÁLISE DO RISCO DE CRÉDITO

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Engenharia

Área de Concentração:
Engenharia Elétrica - Sistemas Eletrônicos

Orientador: Prof. Livre-Docente Flavio Almeida de Magalhães Cipparrone

São Paulo
2009

Este exemplar foi revisado e alterado em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 28 de março de 2009.

Assinatura do autor _____

Assinatura do orientador _____

FICHA CATALOGRÁFICA

Karcher, Cristiane
Redes bayesianas aplicadas à análise do risco de crédito / C. Karcher. -- ed.rev. --São Paulo, 2009.
103 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos.

1. Crédito 2. Estatística para inteligência artificial 3. Modelos lineares generalizados 4. Inferência estatística I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Sistemas Eletrônicos II. t.

DEDICATÓRIA

Dedico esse trabalho ao
Daniel pelo amor, compreensão e
incentivo ao longo de anos.

AGRADECIMENTOS

Ao professor Dr. Flávio Almeida de Magalhães s Cipparrone, pela orientação e oportunidade de crescimento acadêmico.

Ao professor Dr. Afonso de Campos Pinto pela orientação, apoio e incentivo durante a preparação deste trabalho.

À minha querida irmã Viviane Karcher e aos amigos Paulo do Canto Hubert Jr, Cláudio de Nardi Queiroz e Pedro Savadovsky pelas contribuições a este trabalho.

À empresa MAPS Soluções & Serviços pelo conhecimento, incentivo e compreensão durante a preparação desse trabalho.

À meus pais e minha avó Esther pelo incentivo e pelas lições de vida.

EPÍGRAFE

Se quisermos progredir, não devemos repetir a história, mas fazer uma história nova.

(Mahatma Ghandi)

RESUMO

Modelos de *Credit Scoring* são utilizados para estimar a probabilidade de um cliente proponente ao crédito se tornar inadimplente, em determinado período, baseadas em suas informações pessoais e financeiras. Neste trabalho, a técnica proposta em *Credit Scoring* é Redes Bayesianas (RB) e seus resultados foram comparados aos da Regressão Logística. As RB avaliadas foram as *Bayesian Network Classifiers,* conhecidas como Classificadores Bayesianos, com seguintes tipos de estrutura: *Naive Bayes*, *Tree Augmented Naive Bayes (*TAN) e *General Bayesian Network* (GBN). As estruturas das RB foram obtidas por Aprendizado de Estrutura a partir de uma base de dados real. Os desempenhos dos modelos foram avaliados e comparados através das taxas de acerto obtidas da Matriz de Confusão, da estatística Kolmogorov-Smirnov e coeficiente Gini. As amostras de desenvolvimento e de validação foram obtidas por *Cross-Validation* com 10 partições. A análise dos modelos ajustados mostrou que as RB e a Regressão Logística apresentaram desempenho similar, em relação a estatística Kolmogorov-Smirnov e ao coeficiente Gini. O Classificador TAN foi escolhido como o melhor modelo, pois apresentou o melhor desempenho nas previsões dos clientes “maus” pagadores e permitiu uma análise dos efeitos de interação entre variáveis.

Palavras-chave: Redes Bayesianas, Risco de Crédito, Regressão Logística.

ABSTRACT

Credit Scoring Models are used to estimate the insolvency probability of a customer, in a period, based on their personal and financial information. In this text, the proposed model for Credit Scoring is Bayesian Networks (BN) and its results were compared to Logistic Regression. The BN evaluated were the Bayesian Networks Classifiers, with structures of type: *Naive Bayes*, Tree Augmented *Naive Bayes* (TAN) and General Bayesian Network (GBN). The RB structures were developed using a Structure Learning technique from a real database. The models performance were evaluated and compared through the hit rates observed in Confusion Matrix, Kolmogorov-Smirnov statistic and Gini coefficient. The development and validation samples were obtained using a Cross-Validation criteria with 10-fold. The analysis showed that the fitted BN models have the same performance as the Logistic Regression Models, evaluating the Kolmogorov-Smirnov statistic and Gini coefficient. The TAN Classifier was selected as the best BN model, because it performed better in prediction of “bad” customers and allowed an interaction effects analysis between variables.

Keywords: Bayesian Networks, Credit Risk, Logistic Regression.

LISTA DE ILUSTRACÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIC	Medida *Akaike's Information Criterion*
BAYES	Medida *Bayesian Dirichlet*
GBN	Classificador Bayesiano *General Bayesian Network*
KS	Estatística *Kolmogorov-Smirnov*
MDL	Medida *Minimal Description Length*
RB	Rede(s) Bayesiana(s)
ROC	*Receiver Operating Characteristic*
TAN	Classificador Bayesiano *Tree Augmented Naive Bayes*

SUMÁRIO

1. INTRODUÇÃO

Neste trabalho é proposta a aplicação de Redes Bayesianas (RB) na construção de modelos de *Credit Scoring* e suas aplicações serão comparadas com a Regressão Logística, que é a técnica mais aplicada atualmente em *Credit Scoring* (ROSA, 2000). As RB avaliadas serão as *Bayesian Network Classifiers* (FRIEDMAN et al., 1997), chamadas de Classificadores Bayesianos, que são RB aplicadas em problemas de classificação de dados.

Os modelos de *Credit Scoring* são utilizados para estimar a probabilidade de um cliente proponente ao crédito se tornar inadimplente, em determinado período, dadas suas informações pessoais e financeiras que possam influenciar na capacidade do cliente em pagar a dívida. Esta probabilidade estimada, chamada de *score* com valores entre 0 e 100, é uma estimativa do risco de inadimplência de um cliente em determinado período.

No processo de concessão de crédito, quando um novo cliente solicita um crédito, o mesmo fornece suas informações cadastrais e financeiras que, juntamente às variáveis da operação, são utilizadas para lhe gerar um *score* de 0 a 100 pontos. Este *score* poderá, então, ser utilizado na decisão de conceder ou não o crédito ao cliente, a partir de um ponto de corte, acima do qual o pedido do cliente será aceito. O ponto de corte é definido a partir da análise dos erros de rejeitar um cliente "bom" pagador e de aceitar um cliente "mau" pagador, em determinado período (erros do tipo I e II, respectivamente), e também a partir da análise da rentabilidade esperada do cliente. Atualmente são utilizadas entre três e cinco faixas de *score* para a classificação dos clientes (entre três e cinco), principalmente em função do requerimento imposto pela Resolução 2682 (BANCO CENTRAL, 1999), exigindo que os clientes tenham uma classificação de crédito com diferentes níveis de provisão.

Os modelos de *Credit Scoring* são utilizados no início do relacionamento com o cliente, período em que a Instituição mensura previamente o risco do proponente e atribui a ele ou não linhas diferenciadas em função do seu perfil. Os modelos de *Credit Scoring* começaram a ser utilizados principalmente no segmento varejista do

crédito ao consumidor, que tem como características o grande volume de transações, o baixo valor unitário por transação, spread de taxa de juros elevado e a necessidade de velocidade na decisão (LOURENÇO, 2005).

Atualmente os *Credit Scoring* são considerados ferramentas importantes para pré-qualificar os tomadores de crédito e auxiliar os gestores a tomar decisões de risco mais adequadas ao negócio. O uso destes modelos permite que a decisão sobre a concessão ou não do crédito seja tomada de forma objetiva, padronizada e imparcial, o que não é garantido na análise julgamental. Isto possibilita que o cliente seja tratado de forma personalizada, independente do canal de atendimento.

Existem dois tipos de modelos de mensuração utilizados para estimar a probabilidade de um cliente se tornar inadimplente, são eles (SECURATO, 2002):

- *Credit Scoring* – obtido a partir das informações cadastrais fornecidas pelos clientes tais como: tipo de residência, nível de renda, idade, ocupação, grau de instrução, relacionamento com instituições financeiras, consultas aos bureaus[1] de crédito.

- *Behavioural Scoring* – é um sistema de pontuação com base em análise comportamental e utiliza as informações que a empresa já possui sobre o cliente na renovação, manutenção ou concessão de uma nova linha de crédito. Pode incluir informações relacionadas aos hábitos de consumo, de pagamento, comprometimento de renda etc.

Neste trabalho serão empregados somente os modelos de *Credit Scoring* mencionados anteriormente. Nos modelos de *Credit Scoring* as variáveis preditoras são as informações pessoais e financeiras fornecidas pelos clientes na proposta de crédito e a variável resposta é a classificação do cliente de acordo com seu risco de inadimplência (cliente “bom” pagador ou “mau” pagador). No decorrer do texto, a variável resposta também pode ser chamada de *classe* e as variáveis preditoras ou explicativas também podem ser chamadas de *atributos*.

Na Estatística e a Inteligência Artificial existem diversas técnicas utilizadas em *Credit Scoring* tais como: Árvores de Classificação, Redes Neurais, Análise do Discriminante Linear, Regressão Logística. No entanto, o uso das Redes Neurais ainda é restrito, apesar de ser uma ferramenta poderosa de reconhecimento de padrões, devido a sua natureza de "caixa preta", pois não se conhece as relações de dependência entre as variáveis do modelo e nem a contribuição de cada variável.

RB são propostas para modelos de *Credit Scoring*, pois se tratam de modelos probabilísticos nos quais são conhecidas as relações entre as variáveis do domínio, ao contrário de Redes Neurais. As *RB* utilizadas em problemas de classificação de dados são chamadas de Classificadores Bayesianos e têm como objetivo prever a classe de objetos que não foram classificados como, por exemplo, classificar um novo cliente como "bom" ou "mau" pagador, de acordo com a observação de suas variáveis preditoras.

A implantação do Plano Real e o fim do período inflacionário contribuíram para o reaquecimento da economia e o crescimento da demanda por crédito no Brasil. Nos últimos anos, as operações de crédito do sistema financeiro apresentaram crescimento expressivo observado pelo aumento da relação do volume total dos empréstimos privados e o Produto Interno Bruto (PIB), que passou de 26,2% em dezembro de 2003 para 33,7% em novembro de 2006 e para 34,6% em fevereiro de 2007 (BANCO CENTRAL, 2007). Para 2008, a Febraban (Federação Brasileira dos Bancos) espera que a proporção chegue a 38% (MARCHESINI, 2007).

No Brasil, a concessão de crédito é uma atividade financeira que vem crescendo nos últimos anos no Brasil. Os fatores favoráveis para o aumento das concessões são: condição de mercado, maior demanda, crescimento da economia, crescimento de renda e nível menor de inadimplência (SIQUEIRA, 2007). Atualmente, o crédito já corresponde à metade do lucro dos bancos, sendo superiores aos ganhos com títulos do governo e tarifas (PAIVA, 2007).

A avaliação do risco de crédito tem sido bastante debatida em 2007 e 2008 devido à crise financeira mundial, iniciada em março de 2007 nos Estados Unidos

[1] *Bureaus* de crédito são informações de mercado a respeito do risco de crédito de um cliente.

com a crise no crédito imobiliário para o segmento de clientes *subprime* (de segunda linha). O segmento de crédito *subprime* é o dos clientes com renda muito baixa, por vezes com histórico de inadimplência e com dificuldade de comprovar renda. Como os empréstimos a clientes *subprime* têm uma qualidade mais baixa, por terem maior risco de não serem pagos, eles oferecem uma taxa de retorno mais alta, a fim de compensar esse risco assumido pelos credores.

Em busca de rendimentos maiores, gestores de fundos e bancos compravam esses títulos *subprime* das instituições que fizeram o primeiro empréstimo, o que permitia que uma nova quantia em dinheiro fosse emprestada, antes mesmo de o primeiro empréstimo ser pago. Também interessado em lucrar, um segundo gestor também poderia comprar o título adquirido pelo primeiro, e assim por diante, gerando uma cadeia de venda de títulos. Porém, se a ponta (o tomador) não consegue pagar sua dívida inicial, ele dá início a um ciclo de não-recebimento por parte dos compradores dos títulos. O resultado: todo o mercado passa a ter medo de emprestar e comprar os *subprime*, o que termina por gerar uma crise de liquidez (retração de crédito). Nesta crise financeira mundial, o medo é que com menos crédito disponível, caia o consumo e diminua o crescimento das economias (FOLHA ON LINE, 2007).

Um dos primeiros reflexos da crise *subprime* foi, em setembro de 2007, quando três fundos do banco francês BNP Paribas tiveram suas negociações suspensas por não ser possível avaliá-los com precisão, devido aos problemas no mercado *subprime* americano. Depois desta medida, o mercado imobiliário passou a reagir em pânico e algumas das principais empresas de financiamento imobiliário passaram a sofrer os efeitos da retração. A *American Home Mortgage* (AHM), uma das 10 maiores empresas do setor de crédito imobiliário e hipotecas dos EUA, pediu concordata. Entre as vítimas mais recentes da crise, estão as duas maiores empresas hipotecárias americanas, a *Fannie Mae* e a *Freddie Mac*, que possuem quase a metade dos US$ 12 trilhões em empréstimos para a habitação nos EUA e, em setembro de 2008, tiveram uma ajuda de até US$ 200 bilhões. Menos sorte teve o banco *Lehman Brothers*, que não teve ajuda do governo dos EUA, como a que foi destinada às duas hipotecárias, e pediu concordata. Como medida emergencial para evitar uma desaceleração ainda maior da economia, já que 70% do PIB americano é

movido pelo consumo, o presidente americano George W. Bush sancionou em fevereiro de 2008 um pacote de estímulo que incluiu o envio de cheques de restituição de impostos a milhões de norte-americanos. Em setembro de 2008, com o agravamento o governo dos EUA lançou um pacote no valor de 600 bilhões de dólares de estímulo à economia e diversos países da Europa adotaram medidas similares para tentar salvar seus sistemas financeiros (FOLHA ON LINE, 2008).

Estes fatos observados na economia mundial alertam para a necessidade de uma gestão eficiente e responsável do risco de crédito pelas Instituições que concedem crédito. Para isso, as Instituições adotam processos de concessão de crédito baseados em modelos estatísticos para mensuração e gestão do risco de inadimplência de suas carteiras de crédito.

A decisão sobre a concessão ou não de um produto de crédito a um cliente é fundamental para o resultado financeiro da Instituição, já que o lucro dos credores está diretamente associado à proporção de clientes aprovados e ao percentual de clientes que pagam as dívidas contraídas. Atualmente, na crise financeira mundial, podem-se observar os reflexos de uma gestão do risco de crédito com altos níveis de inadimplência assumidos por diversas Instituições Financeiras ao redor do mundo.

Este trabalho é organizado em seis capítulos: Introdução, Revisão Bibliográfica, Fundamentação Teórica, Metodologia, Resultados e Conclusão. No segundo capítulo há uma revisão bibliográfica de modelos de *Credit Scoring*, *RB* e Classificadores Bayesianos, além de serem descritos artigos de aplicações de *RB* aplicadas na análise do Risco de Crédito. No terceiro capítulo será apresentada a fundamentação teórica de *RB*, Classificadores Bayesianos, Regressão Logística Múltipla e Medidas de Avaliação dos Modelos de Classificação. A Metodologia empregada será descrita no quarto capítulo e está divida em: amostra *German Credit*, categorização das variáveis preditoras, seleção de variáveis, construção das amostras de desenvolvimento e de validação e softwares utilizados. No quinto capítulo serão apresentados e comparados os resultados das aplicações dos Classificadores Bayesianos e da Regressão Logística em modelos de *Credit Scoring*. Finalmente, no sexto capítulo há a conclusão deste estudo e serão propostos trabalhos futuros.

2. REVISÃO BIBLIOGRÁFICA

Neste capítulo serão descritos alguns estudos sobre aplicações de modelos de *Credit Scoring*. Posteriormente, apresentaremos alguns estudos teóricos e de aplicações de *RB* em problemas de classificação de dados, que é a técnica proposta para modelos de *Credit Scoring*.

2.1 Modelos de *Credit Scoring*

Os modelos de *Credit Scoring* são utilizados para estimar a probabilidade de um cliente proponente ao crédito se tornar inadimplente, em determinado período, dadas suas informações pessoais e financeiras que possam influenciar na capacidade do cliente em pagar a dívida. Esta probabilidade atribuída a cada novo cliente é chamada de *score*, assumindo valores entre 0 e 100, e é considerada uma estimativa do risco de inadimplência do cliente, em determinado período. Assim, o *score* do cliente pode ser utilizado na decisão de conceder ou não o crédito, a partir de um ponto de corte acima do qual o pedido do cliente será aceito.

Com isso, do ponto de vista de modelagem estatística, o problema de concessão de crédito por uma Instituição a um cliente é um problema prático de classificação. Inúmeras técnicas já foram aplicadas em *Credit Scoring* tais como: Regressão Linear, Análise Discriminante, Regressão Logística, Redes Neurais, Algoritmos Genéticos, Árvores de Decisão. Na literatura científica também existem diversos estudos comparativos das aplicações de diferentes técnicas e alguns serão descritos a seguir.

A Análise Discriminante Linear foi um dos primeiros modelos de *Credit Scoring*. Eisenbeis (1978) discute diversos problemas em aplicar Análise Discriminante Linear em *Credit Scoring.* Eisenbeis (1978) discutiu que um ponto desfavorável ao uso destes modelos está no fato das matrizes de variâncias e covariância das classes “bom” e “mau” provavelmente não serem iguais. Além isso, outro ponto desfavorável

é o fato das variáveis explicativas não apresentarem normalidade multivariada, por serem predominantemente categóricas.

A Regressão Linear Múltipla é outra técnica utilizada na formulação de modelos de *Credit Scoring* com resposta do tipo binária (“bom” ou ”mau”). Hand (2001) discutiu que em dados de *Credit Scoring* estes modelos apresentam problemas de heterocedasticidade. Mas, a principal limitação apontada foi a de que os valores estimados para a variável de resposta não pertencem ao intervalo [0,1], podendo assumir valores negativos e até mesmo maiores que um, o que não é uma resposta esperada.

Rosa (2000) apresentou a uma aplicação de Regressão Logística no problema de concessão de crédito em um produto de financiamento de veículos, comparada com aplicações de modelos baseados em árvores de decisão. Em seu trabalho, Rosa (2000) concluiu que as ferramentas baseadas em árvore de decisão classificaram os clientes de forma um pouco mais precisa, em relação às taxas de acerto nas previsões dos clientes “bons” e “maus” pagadores. No entanto, a Regressão Logística, que apresentou bons resultados também, possui a vantagem de ser um modelo de fácil compreensão e interpretação dos parâmetros. A Regressão Logística também tem a vantagem de produzir como resultado uma probabilidade, o que permite a ordenação dos clientes quanto ao risco de inadimplência.

Arminger, Enache e Bonne (1997) comparam aplicações de Regressão Logística, Árvore de Classificação e um tipo de Rede Neural chamada *Feedforward Network*. Através da avaliação da proporção de classificações corretas, o estudo concluiu que o modelo de Regressão Logística apresentou desempenho melhor do que os modelos de Árvore de Classificação e de Redes Neurais, sendo que os dois últimos modelos apresentaram resultados equivalentes. Os autores também propõem um procedimento combinado dos três modelos utilizando seus valores previstos e observados. Este procedimento apresentou resultados superiores aos obtidos nos modelos de Árvore de Classificação e de Redes Neurais, porém inferiores ao modelo de Regressão Logística.

West (2000) fez um estudo comparativo da aplicação em *Credit Scoring* de diversos tipos de Redes Neurais e diversas técnicas como: Regressão Logística, Análise Discriminante Linear e Árvores de Decisão. O estudo sugeriu que os modelos de Redes Neurais apresentaram acurácia maior se comparados aos demais modelos aplicados, mas necessitam de um conhecimento maior para a construção da topologia e para realizar o treinamento da rede. O estudo também sugeriu que a Regressão Logística é uma boa alternativa aos modelos de Redes Neurais. West (2000) também mostra que os modelos de Regressão Logística apresentaram acurácia maior do que os modelos de Análise Discriminante Linear nos dados analisados e nesta base de dados os modelos de Árvore de Decisão não apresentou resultados satisfatórios.

Modelos de *Credit Scoring*, quando são aplicados em bases de dados diferentes, podem apresentar resultados distintos devido às características da base de dados empregada, tais como: a representatividade da amostra em relação à população alvo, o número de observações disponível, além de poderem apresentar características particulares à população alvo. Por isso, é recomendável que os modelos sejam comparados em uma mesma base de dados. No entanto, se os modelos são aplicados em bases de dados diferentes é possível que alguns de seus resultados sejam distintos, como observado nos modelos de Árvores de Decisão e de Redes Neurais em West (2000), Arminger, Enache e Bonne (1997) e Rosa (2000).

Hand e Henley (1997) elucidaram diversos cuidados para a aplicação de modelos de *Credit Scoring*. Um problema de grande relevância em *Credit Scoring*, apontado no estudo, é o do viés na amostra utilizada na construção dos modelos. Este problema ocorre, pois somente as propostas de crédito que foram aceitas são utilizadas nos modelos de *Credit Scoring*, o que torna a amostra de treinamento viesada porque esta não contém toda a população de clientes.

Os métodos que procuram corrigir este viés amostral são conhecidos como Inferência dos Rejeitados e consistem em inferir qual seria o comportamento dos indivíduos rejeitados caso eles tivessem sido aprovados. Hand e Henley (1997) também discutiram que a mudança na população alvo dos modelos degrada o seu

desempenho e ocorre devido a pressões econômicas e mudanças no ambiente competitivo. Por isso, periodicamente novos modelos devem ser construídos.

Rosa (2000) e Hand e Henley (1997) também descreveram os cuidados na definição da variável resposta em modelos de *Credit Scoring*. A definição de um cliente “bom” ou “mau” ou “indeterminado” depende da Instituição considerar este cliente lucrativo ou não. Por exemplo, um cliente é considerado “bom” se não apresentou atraso em seus pagamentos e com isso ele pode trazer lucro ao credor. Um cliente é considerado “mau” se apresentou, por exemplo, atraso de mais do que três meses e com isso ele não é lucrativo ao credor. Por fim, um cliente é “indeterminado” se puder ou não ser lucrativo ao credor. Apesar de um cliente poder ser classificado em três classes (“bom”, “mau” ou “indeterminado”), nos modelos de *Credit Scoring* são utilizadas somente as classes “bom” e “mau” da variável resposta, ou seja, variável resposta binária.

2.2 Redes Bayesianas e Classificadores Bayesianos

Redes Bayesianas (RB) são grafos acíclicos e direcionados que permitem a representação da distribuição de probabilidades conjunta de um conjunto de variáveis aleatórias. Cada vértice do grafo representa uma variável aleatória e as arestas representam as dependências diretas entre variáveis. Uma RB possui a seguinte premissa de independência condicional: cada variável é independente das variáveis que não são suas descendentes no grafo, dada a observação de seus pais.

Em RB, estas premissas de independência são exploradas para reduzir o número de parâmetros necessários para caracterizar uma distribuição de probabilidades, e para calcular de forma eficiente as probabilidades a posteriori dadas evidências. Os parâmetros de uma RB são armazenados em tabelas de probabilidades condicionais de cada variável dado seus pais. A distribuição conjunta da RB é determinada unicamente pelas distribuições condicionais de cada variável da RB dado seus pais, pela Regra da Cadeia, que é definida posteriormente na Seção 3.1.

A estrutura de uma RB, ou topologia do grafo, pode ser definida manualmente com os relacionamentos entre variáveis sendo definidos por especialistas ou pode ser aprendida a partir de bases de dados utilizando algoritmos de aprendizado de estrutura. Os parâmetros de uma RB podem ser obtidos a partir do conhecimento de probabilidades por especialistas, do aprendizado a partir de bases de dados ou pela combinação de ambas as abordagens (NEAPOLITAN, 2004).

O aprendizado em RB também tem sido bastante estudado por diversos autores como Neapolitan (2004), Buntine (1996) e Heckerman (1995).

RB aplicadas em problemas de classificação de dados são chamadas de Classificadores Bayesianos. Estes modelos têm como objetivo descrever e distinguir classes e também prever a classe de objetos que não foram classificados.

Neste trabalho serão descritos os seguintes Classificadores Bayesianos: *Naive Bayes*, *Tree Augmented Naive Bayes (*TAN) e *General Bayesian Network* (GBN).

Os Classificadores Bayesianos mais simples são conhecidos como *Naive Bayes* (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997) e possuem a hipótese que todos os atributos são independentes dado à classe. Friedman e Goldszmidt (1996) compararam a aplicação do Classificador *Naive Bayes* com o Classificador GBN (*RB* Irrestritas), com aprendizado de estrutura utilizando a medida MDL, descrito Seção 3.1.2. O estudo concluiu que os Classificadores GBN apresentaram desempenho significativamente superior ao Classificador *Naive Bayes*, mas apresentaram desempenho pobre em bases de dados com mais de 15 atributos. Este fraco desempenho observado nos Classificadores GBN com mais do que 15 atributos deve-se ao grande número de parâmetros destes Classificadores. Esta baixa assertividade e baixo poder discriminante observados em alguns aprendizados dos Classificadores GBN podem ocorrer devido ao grande número de dependências avaliadas em conjuntos de dados com mais do que 15 atributos.

Os Classificadores *Naive Bayes,* na presença de variáveis altamente correlacionadas (redundantes), podem ampliar desnecessariamente o peso da evidência destes atributos sobre a classe, o que pode prejudicar a assertividade das

classificações. Com isso, Langley e Sage (1994) mostraram que a seleção de variáveis preditoras (atributos) através dos métodos *forward* e *backward* melhorou a acurácia do Classificador *Naive Bayes* em muitos casos.

Friedman, Geiger e Goldszmidt (1997), propuseram o Classificador *Tree Augmented Naive Bayes (*TAN) como uma extensão ao Classificador *Naive Bayes,* permitindo a análise de interações entre variáveis preditoras desde que a estrutura representada por estas variáveis seja a estrutura de uma árvore. Portanto, o Classificador TAN proposto encontra a relação entre atributos restrita ao espaço de estruturas do tipo árvores e esta busca pode ser feita em tempo polinomial (CHOW; LIU, 1968).

Um problema que pode ocorrer nas aplicações dos Classificadores bayesianos, principalmente nos Classificadores BAN e GBN, é o *overfitting* (superajuste) (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997) (CHENG; GREINER, 1999, 2001). Este problema é decorrente do grande número de parâmetros que a rede bayesiana construída pode apresentar e pode degradar o desempenho do Classificador. Para o problema de *overfitting.*

Friedman et al (1997) propuseram o uso da medida MDL no aprendizado de estrutura do Classificador Bayesiano GBN, pois a medida MDL é capaz de regular a complexidade da rede bayesiana pela penalização daquelas que contenham muitos parâmetros, o que pode ajudar a evitar o problema de *overfitting*.

Para contornar o problema de *overfitting*, *C*heng e Greiner (2001) propuseram que no Classificador GBN fossem selecionadas as variáveis preditoras do *Markov Blanket* da variável resposta (classe). A escolha do subconjunto de variáveis do *Markov Blanket* da variável resposta, para compor o Classificador GBN, é um procedimento natural de seleção de atributos, pois as variáveis do *Markov Blanket* da classe “protegem” a variável resposta da influência de qualquer outra variável de fora do seu *Markov Blanket*.

A construção de Classificadores a partir de bases de dados de instâncias (observações) pré-classificadas é um problema muito estudado na área Aprendizado

de Máquina (*Machine Learning*) (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997), que é a área que combina Estatística com Inteligência Artificial (WITTEN; FRANK, 2005).

RB têm apresentado inúmeras aplicações acadêmicas e na indústria. Na área financeira, Poku (2005) e Guidici (2004) constroem modelos causais utilizando RB para a mensuração e gestão de Riscos Operacionais Financeiros. Poku (2005) construiu modelos de RB combinando o conhecimento de especialistas com as informações de séries históricas de perdas operacionais. Para ações de Marketing, Baesens et al. (2004) utilizaram RB para classificar clientes quanto ao seu potencial de gasto e oferece a eles novos produtos ou vantagens. Especificamente neste estudo, foram utilizados Classificadores bayesianos, RB aplicadas em problemas de classificação, para prever o aumento ou diminuição do gasto futuro de um cliente baseado nas em suas informações iniciais de compras. As aplicações de RB na análise de Risco de Crédito são descritas na Seção 2.3, a seguir.

2.3 Redes Bayesianas aplicadas à análise do Risco de Crédito

Serão descritos, a seguir, dois estudos de aplicações dos Classificadores bayesianos em modelos de *Credit Scoring*, o de Baesens et al. (2002) e o de Chang et al. (2000). Algumas das metodologias de construção dos Classificadores bayesianos empregadas neste trabalho se basearam nas metodologias empregadas nestes dois artigos. As metodologias comuns e as contribuições do presente estudo, em relação aos artigos de Baesens et al. (2002) e Chang et al. (2000), serão descritas a seguir, após uma breve introdução a estes artigos.

O artigo de Baesens et al. (2002) descreveu a aplicação dos Classificadores bayesianos *Naive Bayes*, TAN e GBN em modelos de *Credit Scoring*. O aprendizado de estrutura empregado na construção dos Classificadores GBN utilizou a simulação *Markov Chain Monte Carlo* (MCMC) e a seleção de variáveis destes Classificadores foi feita utilizando o *Markov Blanket* da variável resposta (Cliente “bom” ou “mau”). As principais conclusões do estudo foram que os Classificadores GBN apresentaram um

bom desempenho em *Credit Scoring* e que a seleção de variáveis pelo *Markov Blanket* da variável resposta resultou em modelos mais parcimoniosos e poderosos.

Chang et.al. (2000) descreveram a teoria de construção de escores a partir de Classificadores Bayesianos. No artigo, foi construída uma RB aplicando aprendizado de estrutura com busca da estrutura pela adição e remoção de arestas até que não haja melhoria na razão de verossimilhanças (*log-likelihood ratio*). As variáveis preditoras originais (todas categóricas) foram convertidas em variáveis *dummy* (variáveis binárias) e foram selecionadas pelo *Markov Blanket* da variável resposta. Além disso, foram obtidos os cliques da variável resposta, que são subconjuntos de variáveis condicionalmente independentes, dada a observação da variável resposta dentro do seu *Markov Blanket* e, que formaram subconjuntos de variáveis interpretáveis para o processo de concessão de crédito. Além disso, as RB aplicadas também foram comparadas ao modelo de Regressão Logística Múltipla com seleção de variáveis por *Forward Stepwise*.

As metodologias em comum entre este trabalho e os trabalhos de Baesens et al. (2002) e Chang et al. (2000) são: construção das amostras de desenvolvimento e de validação por *Cross-Validation* com 10 partições (*10-fold*), seleção de variáveis nos Classificadores GBN pelo *Markov Blanket* da variável resposta, conversão das variáveis categóricas originais em *dummies* (variáveis binárias) e comparação do desempenho dos Classificadores Bayesianos com o da Regressão Logística.

A contribuição deste estudo está na aplicação mais abrangente dos Classificadores Bayesianos, incluindo a aplicação de aprendizados de estrutura com diferentes medidas de avaliação da estrutura, tais como: AIC, MDL, Bayes, Bdeu e Entropia. Além disso, também serão avaliados procedimentos de seleção de variáveis baseados na filtragem pelo ganho de informação e pelo método *Wrapper.* Adicionalmente, também será abordada a categorização de variáveis contínuas e agrupamento de níveis das variáveis categóricas com muitos níveis, baseada na análise bivariada do risco relativo e da medida WOE (*Weights of Evidence*).

3. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são descritos os principais conceitos da teoria de Redes Bayesianas (RB) e uma introdução à Inferência e ao Aprendizado em RB. O conhecimento do ferramental teórico de RB é importante para o entendimento dos Classificadores Bayesianos, que na mais são do que RB aplicadas em problemas de classificação de dados. Os Classificadores Bayesianos empregados serão: *Naive Bayes*, *Tree Augmented Naive Bayes* (TAN) e *General Bayesian Network* (GBN), descritos na Seção 3.2. Na Seção 3.3 será apresentada parte da teoria de Regressão Logística Múltipla, que é a técnica comparada aos Classificadores Bayesianos nas aplicações em *Credit Scoring*. Por fim, na Seção 3.4, serão apresentadas as medidas utilizadas para avaliar e comparar os modelos de classificação, que são: as taxas de acerto obtidas da Matriz de Confusão, estatística Kolmogorov-Smirnov (KS) e coeficiente Gini.

3.1 Redes Bayesianas

Por definição, uma Rede Bayesiana (RB) é composta dos seguintes elementos (JENSEN, 2001):

i. Um conjunto de variáveis e um conjunto de arestas direcionadas entre as variáveis.
ii. Cada variável tem estados finitos e mutuamente exclusivos.
iii. As variáveis e as arestas direcionadas representam um grafo acíclico direcionado.
iv. Cada variável *A*, com pais $B_1, B_2, ..., B_n$, possui uma tabela de probabilidades condicionais, $P(A|B_1, B_2, ..., B)$, associada.

Em outras palavras, RB são grafos acíclicos e direcionados que permitem a representação da distribuição conjunta de probabilidades de um conjunto de

variáveis aleatórias. Cada vértice do grafo representa uma variável aleatória e as arestas representam dependências entre variáveis. Em uma RB, se há uma aresta direcionada entre os nós A e B, então dizemos que A é pai de B e B é filho de A. Neste trabalho serão tratadas somente as RB com variáveis discretas, mas a teoria a respeito do tratamento de variáveis contínuas em RB pode ser encontrada em Neapolitan (2004).

RB com variáveis discretas satisfazem a condição de Markov (NEAPOLITAN, 2004), que é dada por: cada variável da RB é condicionalmente independente do conjunto de todos os seus não-descendentes dado o conjunto de todos os seus pais. Em uma RB, a distribuição conjunta de probabilidades de um conjunto de variáveis discretas $\{X_1, X_2, ..., X_n\}$ é dada pela Regra da Cadeia,

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P\left(X_i \middle| Pa_i\right) \quad (1)$$

Os parâmetros de uma RB são definidos como,

$$\Theta_i = P\left(X_i \middle| Pa_i\right) \qquad i = 1, ..., n \quad (2)$$

em que, Θ_i é uma tabela de probabilidades condicionais de X_i dado seus pais Pa_i.

Com isso, o conjunto de parâmetros de uma RB é dado por $\Theta_s = \{\Theta_1, \Theta_2, ..., \Theta_n\}$ e são todas as tabelas de probabilidades condicionais da RB com variáveis discretas $\{X_1, X_2, ..., X_n\}$.

Um importante aspecto de uma RB é a sua estrutura (topologia do grafo), que permite a representação de complexas relações entre variáveis de forma gráfica e intuitiva. A estrutura gráfica de uma RB facilita o entendimento das relações entre variáveis do seu domínio, além de permitir o uso combinado de informações obtidas do conhecimento de especialistas com dados históricos para obter a distribuição conjunta de probabilidades da rede.

A estrutura de uma RB pode ser determinada manualmente, com apoio de especialistas, ou pode ser aprendida a partir de bases de dados utilizando algoritmos

de aprendizado de estrutura. Os parâmetros de uma RB podem ser obtidos através da elucidação[2] de probabilidades por especialistas, através do aprendizado a partir de bases de dados ou através da combinação de ambas as abordagens.

Na Figura 1 há um exemplo de RB, que utiliza variáveis discretas, aplicada no diagnóstico de doenças. As variáveis desta RB são {Idade (I), Profissão (P), Clima (C), Doença (D), Sintomas (S)}.

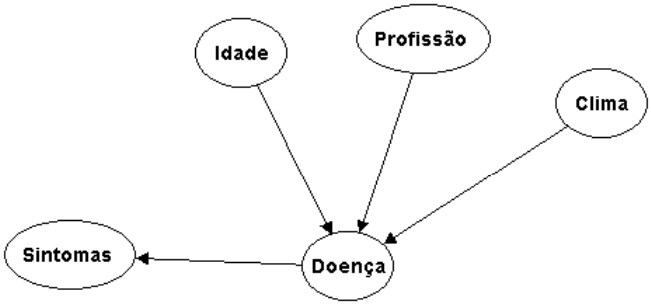

Figura 1 - Exemplo de Rede Bayesiana aplicada em diagnóstico médico.

Na RB da Figura 1, o nó “Sintomas” tem o nó “Doença” como pai e três ancestrais “Idade”, “Profissão” e “Clima”. Através das premissas de independência condicional, podemos dizer que “Sintomas” é dependente de “Idade”, “Profissão” e “Clima” indiretamente através de sua influência sobre “Doença”. Tomando o nó “Clima”, que não possui pai, então podemos dizer que “Clima” é independente de “Profissão” e “Idade”. Aplicando a Regra da Cadeia, equação (1), a distribuição conjunta de probabilidades desta RB é dada por,

$$P(Idade, Profissão, Clima, Doença, Sintomas) = P(Idade).P(Profissão).P(Clima).P(Doença \mid Idade, Profissão, Clima).P(Sintomas \mid Doença) \quad (3)$$

Utilizando somente as letras iniciais de cada variável a equação (3) pode ser reescrita como,

$$P(I,P,C,D,S) = P(I) \,.\, P(P) \,.\, P(C) \,.\, P(D \mid I,P,C) \,.\, P(S \mid D) \quad (4)$$

Pela Regra da Cadeia, as tabelas de probabilidades condicionais de cada variável da RB precisam ser especificadas para que se obtenha a distribuição de

[2] Elucidação é o procedimento de obtenção de distribuições a partir do conhecimento de especialistas.

probabilidades conjunta da RB. Com isso, na RB da Figura 1, é necessário especificar as tabelas *P(Idade)*, *P(Profissão)*, *P(Clima)*, *P(Doença|Idade,Profissão,Clima)* e *P(Sintomas|Doença)* para determinarmos a distribuição conjunta de probabilidade *P(Idade, Profissão, Clima, Doença, Sintomas)*.

Na Tabela 1, temos as tabelas *P(Idade)* e *P(Sintomas|Doença)*. As probabilidades apresentadas nestas tabelas também podem ser chamadas de parâmetros, como definimos previamente em (1). Note que a variável “Idade”, que é uma variável contínua, foi discretizada (ou categorizada) para criar uma variável discreta binária.

Tabela 1 - Probabilidades Condicionais de "Doença" dado "Sintomas" e distribuição de "Idade".

Idade < 45	0.46
Idade ≥ 45	0.54

Sintomas	Doença		
	Úlcera no Estômago	Infarto	Nenhuma
Dor de Estômago	0.8	0.05	0.05
Dor no Peito	0.15	0.90	0.10
Nenhuma	0.05	0.05	0.85

Se no exemplo da Tabela 1, se não fossem utilizadas as suposições de RB e a Regra da Cadeia, ao invés de 5 tabelas seria necessário definir uma grande tabela de probabilidades para obtermos a distribuição conjunta das 5 variáveis. Com isso, a RB fornece uma maneira de simplificar a representação de uma distribuição conjunta de probabilidades.

RB também podem ser utilizadas para calcular novas probabilidades (a posteriori) a partir de informações (evidências) sobre uma ou mais variáveis da rede. Em uma RB, um subconjunto de variáveis E com valores conhecidos, $E=e$, em uma dada situação, é conhecido como conjunto de evidência, ou simplesmente evidência, por exemplo, $E=\{X_2=x_2,X_6=x_6\}$. Também podemos dizer que uma variável está instanciada (evidência forte) se conhecemos o estado desta variável.

Um conceito importante em RB é o de d-separação. Segundo Jensen (2001), dizemos que dois vértices distintos X e Y estão d-separados em uma RB se, para

todos os caminhos entre *X* e *Y* existe um vértice intermediário *V* (distinto de *X* e *Y*) tal que a conexão entre *X* e *Y* através de *V*:

• é serial ou divergente e *V* recebeu uma evidência ou;

• é convergente e nem *V* nem algum de seus descendentes receberam uma evidência.

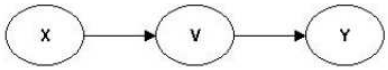

Figura 2 - Conexão Serial: *X* e *Y* estão d-separados se *V* recebeu uma evidência.

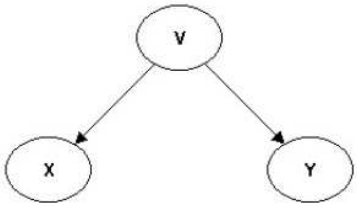

Figura 3 - Conexão Divergente: X e Y estão d-separados se V recebeu uma evidência.

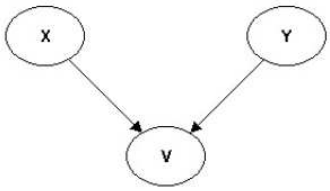

Figura 4 - Conexão Convergente: X e Y estão d-separados se nem V nem seus descendentes recebeu uma evidência.

Em RB se dois vértices quaisquer estão d-separados então eles são condicionalmente independentes. Através do conceito de d-separação é possível identificar a condição de independência condicional entre variáveis em uma RB.

Outro conceito importante em RB é o de *Markov Blanket*. O *Markov Blanket* de uma variável *X* é o conjunto das variáveis que são pais de *X*, filhos de *X* e as variáveis que compartilham um filho com *X* (JENSEN, 2002). Com isso, se todas as variáveis do *Markov Blanket* de *X* possuem evidências, então *X* está d-separado de todas as outras variáveis da RB e, conseqüentemente, *X* é condicionalmente independente de todas as outras variáveis da rede, dado seu *Markov Blanket*. No exemplo da Figura 5, o *Markov Blanket* de I é {C, E, K, L, H}.

Em uma RB, qualquer variável é influenciada diretamente somente pelas variáveis que compõe o seu *Markov Blanket*. Com isso, o conceito de *Markov Blanket* pode ser utilizado para seleção de variáveis em RB, como descreveremos na Seção 3.2.4.

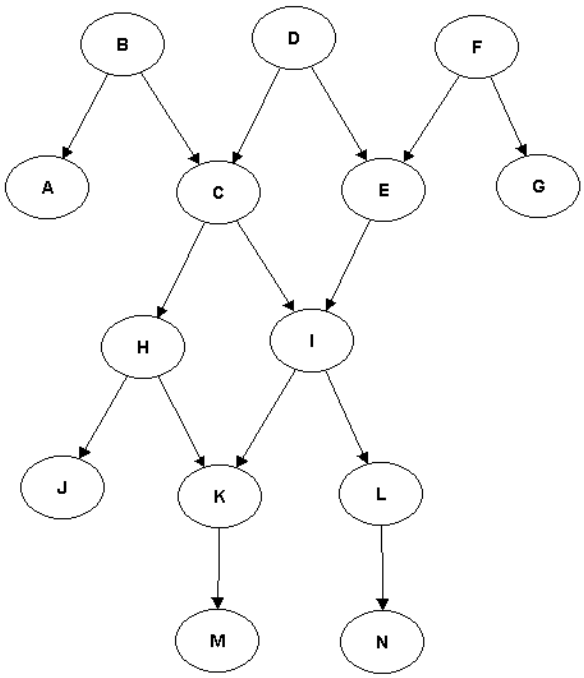

Figura 5 - Outro exemplo de Rede Bayesiana

3.1.1 Inferência em Redes Bayesianas

A Inferência em RB é o processo de atualização das probabilidades a posteriori de variáveis dado uma evidência fornecida. Em RB, a evidência pode ser definida para qualquer subconjunto de nós e a probabilidade a posteriori pode ser calculada para qualquer outro subconjunto de nós.

Como uma RB contém a distribuição de probabilidade conjunta de todas as variáveis do seu domínio, então é possível obter a distribuição de probabilidades de qualquer variável do seu domínio a partir da Regra da Probabilidade Total (Apêndice A). No exemplo de RB da Figura 1, a distribuição da variável “Doença” (D), pela Regra da Probabilidade Total, é dada (inferida) por,

$$P(D=d)=\sum_{i}\sum_{p}\sum_{c}\sum_{s}P(I=i,P=p,C=c,D=d,S=s) \qquad \forall d \quad (5)$$

Pela Regra da Cadeia, a equação (5) pode ser escrita como,

$$P(D=d)=\sum_i\sum_p\sum_c\sum_s P(I=i).P(P=p)P(C=c).P(D=d\,|\,I=i,P=p,C=c)P(S=s\,|\,D=d)\quad\forall d \quad (6)$$

em que, *i*, *p*, *c*, *d* e *s* representam, respectivamente, cada estado das variáveis "Idade", "Profissão", "Clima", "Doença", "Sintoma".

Segundo Zhang e Poole (1996), Inferência em uma RB se refere ao processo de cálculo da probabilidade a posteriori $P(X\,|\,Y=Y_o)$ de um conjunto de variáveis X depois de obter algumas observações (evidência) $Y=Y_0$. Aqui *Y* é uma lista de variáveis observadas e Y_0 é a lista de valores observados. Pelo Teorema de Bayes, $P(X\,|\,Y=Y_o)$ é dada por,

$$P(X\,|\,Y=Y_o)=\frac{P(X,Y=Y_o)}{P(Y=Y_o)} \quad (7)$$

Com isso, $P(X\,|\,Y=Y_o)$ *é* obtida a partir da distribuição marginal $P(X,Y)$, que por sua vez é calculada a partir da distribuição conjunta $P(X_1,X_2,...,X_n)$ pela soma de probabilidades de todas as variáveis fora do domínio de $X\cup Y$ uma por uma. No entanto, isso não é viável, pois esta soma fora do domínio de $X\cup Y$ requer um número exponencial de adições (ZHANG; POOLE, 1996).

Para exemplificar o procedimento de Inferência utilizaremos o exemplo de RB da Figura 1. Dada uma evidência *E=e={Idade='<45'; Sintoma='Dor de estômago'}*, queremos obter distribuição a posteriori da variável "Doença" (D) que é,

$$P(D=d'\,|\,E=e)=\frac{P(D=d',E=e)}{P(E=e)}=\frac{P(D=d',I='<45',S='\text{Dor de estômago}')}{P(I='<45',S='\text{Dor de estômago}')}\quad\forall d' \quad (8)$$

que é igual a,

$$P(D=d'\,|\,E=e)=\frac{\sum_p\sum_c P(I='<45',P=p,C=c,D=d',S='\text{Dor de estômago}')}{\sum_p\sum_c\sum_d P(I='<45',P=p,C=c,D=d,S='\text{Dor de estômago}')}\quad\forall d' \quad (9)$$

Aplicando a Regra da Cadeia o numerador da equação (9) pode ser escrito como,

$$\sum_p\sum_c P(I='<45').P(P=p)P(C=c).P(D=d'\,|\,I='<45',P=p,C=c)P(S='\text{Dor de estômago}'\,|\,D=d') \quad (10)$$

O denominador da equação (9) pode ser escrito como,

$$\sum_{p}\sum_{c}\sum_{d} P(I='<45').P(P=p)P(C=c).P(D=d\,|\,I='<45', P=p, C=c)P(S='\text{Dor de estômago}'\,|\,D=d) \quad (11)$$

em que *p*, *c*, *d* representam, respectivamente, cada estado das variáveis “Idade”, “Profissão”, “Clima” e “Doença”.

Para diminuir o tempo necessário para o cálculo das probabilidades a posteriori, diversos algoritmos de Inferência têm sido propostos. Existem dois tipos de algoritmos de Inferência em RB: os exatos e aproximados. Entre os algoritmos exatos destaca-se o algoritmo *Junction Tree* proposto por Jensen (1991), que é baseado em Teoria dos Grafos. Os algoritmos de Inferência exatos podem não ser eficientes em redes com um grande número de nós e arestas, pois o problema de Inferência é *NP-hard* (ZHANG; POOLE, 1996). Para estes casos, podem ser utilizados algoritmos aproximados, baseados em simulação estocástica, tais como: *Forward Sampling*, *Likelihood Weighting*, *Gibbs Sampling*, *Metropolis-Hasting* (NEAPOLITAN, 2004).

Apesar da complexidade inerente do procedimento de Inferência, sistemas que possuem RB como base de conhecimento têm se mostrado muito eficientes e têm sido muito difundidos ganhando importância inclusive em áreas comerciais. Os sistemas *Hugin*, *Netica* e *JavaBayes* são exemplos destes sistemas.

3.1.2 Aprendizado em Redes Bayesianas

A estrutura e os parâmetros de uma RB podem ser obtidos de duas maneiras: a partir de informações de especialistas ou aprendizado a partir de uma base de dados. Também é possível a combinação das duas alternativas para o aprendizado de parâmetros. A utilização de informações de especialistas pode ser muito trabalhosa, principalmente para determinar os parâmetros, pois é necessário obter um grande número de probabilidades. O aprendizado a partir de uma base de dados requer, além da base de dados em si, um algoritmo de aprendizado de parâmetros e de estrutura.

Dado uma base de dados de treinamento *D* com observações independentes de um conjunto de variáveis discretas *X* e alguma informação a priori ξ (obtida a partir de informações de especialistas), o problema de aprendizado em RB consiste em encontrar a estrutura S^h e parâmetros $\boldsymbol{\Theta}_s$ que melhor expliquem os dados contidos em *D*.

Existe uma variedade de ferramentas de aprendizado em RB, para estruturas conhecidas e desconhecidas, para bases de dados completas e incompletas. O caso em que a estrutura da RB é conhecida é o mais simples, pois é necessário aprender somente as tabelas de probabilidade condicionais (parâmetros) da RB. O caso em que a estrutura da RB é desconhecida é mais complexo, pois é necessário aprender a estrutura e posteriormente as tabelas de probabilidade condicionais da RB.

Conhecida a estrutura S^h de uma RB com parâmetros independentes $\Theta_s = \{\Theta_1, \Theta_2, ..., \Theta_n\}$, em que Θ_i são as tabelas de probabilidades $P(X_i \mid Pa_i, \Theta_i, S^h)$ e, dado uma base de dados completa *D* de exemplos independentes de um conjunto de variáveis discretas $\{X_1, X_2, ..., X_n\}$, o problema de aprendizado de parâmetros se resume a calcular a distribuição a posteriori $P(\Theta_S \mid D, S^h)$ que é dada por,

$$P(\Theta_S \mid D, S^h) = \prod_{i=1}^{n} P(\theta_i \mid D, S^h) \tag{12}$$

A obtenção dos parâmetros a partir de bases de dados pode ser feita através da simples contagem de freqüências (NEAPOLITAN, 2004) ou a partir da abordagem combinada de dados observados em *D* com alguma informação a priori ξ de especialistas. Esta abordagem combinada é baseada em distribuições de Dirichlet (NEAPOLITAN, 2004). Além disso, se os dados observados em *D* estiverem incompletos, são utilizados algoritmos EM (*Expectation Maximization*) (HECKERMAN, 1995).

Para o problema de aprendizado de estrutura, consideraremos um conjunto finito *S* de possíveis estruturas de uma RB. Cada estrutura $S^h \in S$ pode representar a distribuição conjunta de probabilidades do conjunto de variáveis discretas $X = \{X_1, X_2, ..., X_n\}$. Dada uma base de dados completa *D*, a tarefa do aprendizado

de estrutura está em obter a distribuição a posteriori $P(S^h|D)$, que pelo teorema de Bayes é dada por,

$$P(S^h|D) = \frac{P(S^h)P(D|S^h)}{P(D)} \tag{13}$$

A distribuição $P(S^h)$ é chamada a priori de cada possível estrutura S^h, $P(D|S^h)$ é chamada verossimilhança marginal e $P(D)$ é uma constante de normalização.

Para o aprendizado de estrutura em RB serão apresentadas duas abordagens. A primeira abordagem propõe algoritmos de aprendizado de estrutura chamados *CI-based* (*Conditional Independence-based*) e se baseia na análise de dependência entre os nós. Nestes algoritmos as relações de dependência entre variáveis são avaliadas através de testes de independência condicional, como qui-quadrado ou informação mútua, e são criadas arestas para as dependências mais relevantes indicadas por estes testes. Estes algoritmos utilizam o conceito de d-separação, ou seja, no conceito de que a estrutura de uma RB armazena todas as relações de independência condicional entre nós (CHENG; GREINER, 1999, 2001).

Os testes realizados nos algoritmos *CI-based* consistem em avaliar quais dois nós x_i e x_j são condicionalmente independentes, dado um conjunto de nós *c*. Isso é feito, por exemplo, avaliando se a informação mútua condicional dos nós é menor do que um valor *e*. A informação mútua condicional é calculada por (MADDEN, 2003),

$$I(x_i, x_j|c) = \sum_{X_i, X_j, C} P(X_i, X_j, C) \ln\left(\frac{P(X_i, X_j|C)}{P(X_i|C)P(X_j|C)}\right) \tag{14}$$

A segunda abordagem propõe algoritmos de aprendizado de estrutura, chamados *Score-based,* que consistem em introduzir uma medida (*score*), para avaliar o quanto cada possível estrutura S^h explica dos dados *D,* e um método de busca de uma estrutura, entre as possíveis S^h, com o mais alto valor para esta medida (HECKERMAN, 1995).

As medidas utilizadas para avaliar o quanto cada possível estrutura S^h explica dos dados de *D* descritas pela literatura são: Entropia, AIC, MDL (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997), Bayes e BDeu (HECKERMAN, 1995).

Seja S^h uma possível estrutura de uma Rede Bayesiana. Dada uma base de dados de treinamento *D*, definiremos N_{ijk} como o número de observações em *D* tal que $X_i=k$ e $Pa_i=j$ com i=1,..., n; j=1,...,q_i e k=1,..., r_i, e *N* é o número total de observações. O valor q_i é definido como o número de pais de X_i e r_i é definido como o número de estados da variável X_i. As medidas utilizadas nos algoritmos *Score-based*, que chamaremos de *Score*(S^h,D), são definidas como,

- *Entropia*:

$$Score_{Entropia}(S^h,D) = -\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}.\ln\left(\frac{N_{ijk}}{N_{ij}}\right) \quad (15)$$

- *Akaike's Information Criterion* (AIC):

$$Score_{AIC}(S^h,D) = \left|S^h\right| - \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\ln\left(\frac{N_{ijk}}{N_{ij}}\right) \Rightarrow$$

$$Score_{AIC}(S^h,D) = \left|S^h\right| + Score_{Entropia}\left(S^h, D\right) \quad (16)$$

- *Minimal Description Length* (MDL):

$$Score_{MDL}(S^h,D) = \frac{1}{2}\left|S^h\right|\ln N - \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\ln\left(\frac{N_{ijk}}{N_{ij}}\right) \Rightarrow$$

$$Score_{MDL}(S^h,D) = \frac{1}{2}\left|S^h\right|\ln N + Score_{Entropia}\left(S^h, D\right) \quad (17)$$

em que, $\left|S^h\right|$ é o número de parâmetros da estrutura S^h.

- *Bayesian Dirichlet*, que será chamado *Bayes*:

$$Score_{Bayes}(S^h,D) = P\left(S^h, D\middle|\xi\right)$$

Pelo Teorema de Bayes,

$$Score_{Bayes}(S^h,D) = P(S^h|\xi)P(D|S^h,\xi)$$

Heckerman (1995) calcula $P(D|S^h,\xi)$ utilizando distribuições de Dirichlet e obtém o seguinte resultado,

$$Score_{Bayes}(S^h,D) = P(S^h|\xi)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})} \quad (18)$$

em que, $\Gamma(.)$ é a função gamma e N'_{ijk} são parâmetros da distribuição Dirichlet que satisfazem $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$.

Os valores de N'_{ijk} são obtidos a partir dos parâmetros da RB determinados com auxílio de especialistas. O valor $N'_{ijk} = 1$ indica que não há informação de especialistas (não informativo). $P(S^h|\xi)$ é a distribuição a priori dada a estrutura S^h e também é obtida com apoio de especialistas ou pode ter distribuição uniforme caso não haja informações de especialistas.

- *Bayesian Dirichlet Equivalent*, que será chamado *BDeu*: Heckerman (1995) descreve a medida $Score_{BDeu}(S^h,D)$ como a aplicação de $N'_{i\,jk} = 1/(r_i.q_i)$ em $Score_{Bayes}(S^h,D)$, dado pela equação (18). Sendo que, a expressão $N'_{i\,jk} = 1/(r_i.q_i)$ resulta em $N'_{i\,j} = 1/q_i$.

O problema de otimização para busca de uma estrutura S^h que produza um alto valor para uma medida é *NP-hard* (CHICKERING; GEIGER; HECKERMAN, 1994) e o número de possíveis estruturas de uma RB cresce exponencialmente com o número de variáveis.

Ambas as abordagens de aprendizado de estrutura têm suas vantagens e desvantagens, geralmente os algoritmos de busca apresentam resultados mais rápidos, mas o método de busca pode não encontrar a melhor solução.

3.2 Classificação Bayesiana

Redes Bayesianas (RB) podem ser utilizadas em problemas de classificação de uma maneira clara e direta e as *RB* utilizadas em problemas de classificação de dados são chamadas de Classificadores Bayesianos. Nos Classificadores bayesianos com variáveis discretas $\{A_1, A_2, ..., A_n, C\}$, uma delas, *C*, é a variável classe (variável resposta) e as demais, $\{A_1, A_2, ..., A_n\}$, são os atributos (variáveis preditoras). Neste trabalho serão descritos os seguintes Classificadores bayesianos: *Naive Bayes*, *Tree Augmented Naive Bayes (*TAN) e *General Bayesian Network* (GBN).

3.2.1 Classificador *Naive Bayes*

Os Classificadores bayesianos mais simples conhecidos são os chamados *Naive Bayes*. Os Classificadores *Naive Bayes* partem da hipótese que todos os atributos são independentes, dado a variável classe, e sua representação gráfica é dada na Figura 6.

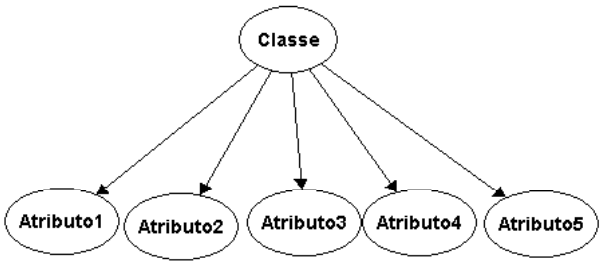

Figura 6 - Estrutura do Classificador *Naive Bayes* com 5 atributos e uma classe

Sob a hipótese de independência condicional entre atributos dada a classe, aplicando a Regra da Cadeia é obtida a distribuição conjunta de probabilidades do Classificador *Naive Bayes* dada por,

$$P(A_1, ..., A_n, C) = P(C).\prod_{i=1}^{n} P(A_i | C) \qquad (19)$$

Em um Classificador bayesiano, com atributos discretos e classe C, assumindo valores {0,1}, a probabilidade de classificarmos um novo caso, $\{A_1 = a_1,..., A_n = a_n\}$, em *C=1* é,

$$P(C = 1 \mid A_1 = a_1,..., A_n = a_n) = \frac{P(C = 1).P(A_1 = a_1,..., A_n = a_n \mid C = 1)}{P(A_1 = a_1,..., A_n = a_n)} \quad (20)$$

E a probabilidade de classificarmos um novo caso em *C=0* é,

$$P(C = 0 \mid A_1 = a_1,..., A_n = a_n) = \frac{P(C = 0).P(A_1 = a_1,..., A_n = a_n \mid C = 0)}{P(A_1 = a_1,..., A_n = a_n)} \quad (21)$$

Com isso, uma nova observação (caso), $\{A_1 = a_1,...A_n = a_n\}$, é classificada na classe *C=1* segundo o seguinte critério:

$$\frac{P(C = 1 \mid A_1 = a_1,..., A_n = a_n)}{P(C = 0 \mid A_1 = a_1,..., A_n = a_n)} \geq 1 \quad (22)$$

O critério descrito em (22) que pode ser escrito como,

$$\frac{P(C = 1)}{P(C = 0)} . \frac{P(A_1 = a_1,..., A_n = a_n \mid C = 1)}{P(A_1 = a_1,..., A_n = a_n \mid C = 0)} \geq 1 \quad (23)$$

No caso do Classificador bayesiano *Naive Bayes*, um novo caso $\{A_1 = a_1,...A_n = a_n\}$ é classificado em *C=1* segundo o seguinte critério:

$$\frac{P(C = 1)}{P(C = 0)} . \prod_{i=1}^{n} \frac{P(A_i = a_i \mid C = 1)}{P(A_i = a_i \mid C = 0)} \geq 1 \quad (24)$$

O Classificador *Naive Bayes* é conhecido por sua simplicidade e eficiência, pois apresentam estrutura fixa e parâmetros ajustáveis. Embora sua suposição de independência seja problemática, pois esta hipótese raramente se verifica no mundo real, os Classificadores *Naive Bayes* têm apresentado um bom desempenho em um grande número de aplicações, especialmente naquelas em que as variáveis preditoras não são fortemente correlacionadas (CHENG; GREINER, 2001).

3.2.2 Classificador *Tree Augmented Naive Bayes* (TAN)

O Classificador bayesiano TAN é uma extensão do *Naive Bayes*, pois permite o relaxamento da hipótese de independência condicional entre atributos dado a classe.

O Classificador TAN foi proposto por Friedman e Goldszmidt (1997) e possibilita representar dependências entre pares de atributos. No Classificador TAN a dependência entre atributos deve ser representada pela estrutura de uma árvore, ou seja, cada atributo deve ter no máximo um pai, fora a classe. Como pode ser observado na Figura 7.

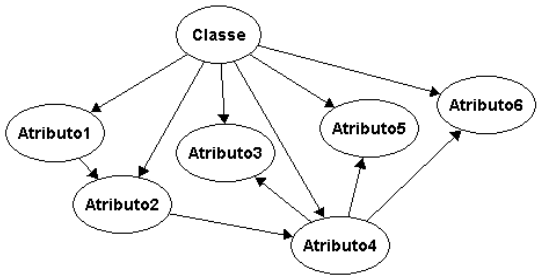

Figura 7 - Estrutura de um Classificador TAN com seis atributos e uma classe

O Classificador TAN utiliza a propriedade de que a busca da melhor estrutura, restrita ao espaço de estruturas do tipo árvore, é feita em tempo polinomial (CHOW. LIU, 1968). Nesta busca uma vez que cada atributo pode ter no máximo um "pai", é necessário encontrar atributo com maior dependência condicional dado à classe.

Como muitas relações de dependência entre as variáveis não podem ser representadas nem mesmo por estruturas tipo TAN, é necessária a construção de modelos mais complexos que permitam que cada nó da rede (exceto a classe) tenha um número arbitrário de pais.

3.2.3 Classificador *General Bayesian Network* (GBN)

Um Classificador bayesiano GBN é uma RB Irrestrita utilizada em problemas de classificação. diferente dos Classificadores bayesianos *Naive Bayes*, TAN e BAN, que tratam a variável classe como um nó especial pai de todos os atributos, o Classificador GBN trata o nó classe como um nó que não necessariamente é pai de todos os atributos. A Figura 8 mostra um exemplo de Classificador GBN.

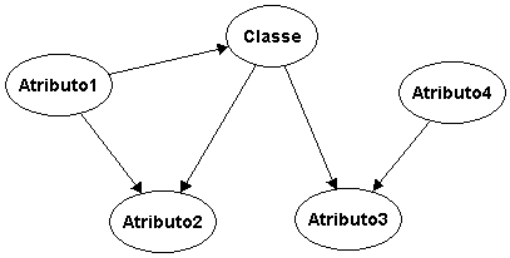

Figura 8 - Estrutura de um Classificador GBN com quatro atributos e uma classe

A construção dos os Classificadores GBN pode ser feita utilizando o aprendizado de estrutura descritos na Seção 3.1.2.

Um problema que pode ocorrer em aplicações dos Classificadores GBN é o *overfitting* (superajuste). *Overfitting* é um fenômeno que ocorre quando um modelo se ajusta demais aos dados de treinamento e não pode ser generalizado para a população inteira. Normalmente, um modelo com problema de *overfitting* não apresenta um bom desempenho fora dos dados de treinamento.

O estudo de Cheng e Greiner (2001) indica que o problema de *overfitting* no Classificador GBN é decorrente do grande número de parâmetros que a RB construída pode apresentar e pode degradar o desempenho do Classificador. Para contornar esse problema, o estudo sugere a seleção das variáveis do *Markov Blanket* da classe para compor o Classificador GBN.

Em uma RB, qualquer variável é influenciada somente pelas variáveis que compõe o seu *Markov Blanket*. Com isso, nos Classificadores GBN, o subconjunto de atributos contidos no *Markov Blanket* da variável classe é um procedimento natural de seleção de variáveis.

Os estudos de Baesens et al. (2002) (2004) concluem, através de aplicações práticas, que o uso do conceito de *Markov Blanket* para seleção de variáveis no Classificador GBN resulta em modelos parcimoniosos e poderosos.

Para o problema de *overfitting*, Friedman, Geiger e Goldszmidt (1997) propõem o uso da medida MDL no aprendizado de estrutura do Classificador GBN, pois a medida MDL é capaz de regular a complexidade da RB pela penalização daquelas que contenham muitos parâmetros, o que ajuda a evitar o problema de *overfitting*.

3.3 Regressão Logística

A análise de Regressão Logística Múltipla (HOSMER; LEMESHOW, 1989) para uma resposta binária é a técnica mais utilizada no desenvolvimento de modelos de *Credit Scoring* (ROSA, 2000).

A Regressão Logística múltipla pode ser escrita como um caso particular dos Modelos Lineares Generalizados (MCCULLAGH; NELDER, 1989) (PAULA, 2004), com função de ligação logito e variável resposta Y_i com distribuição Bernoulli com probabilidade de sucesso (média) π_i.

Seja $Y_i \in \{0,1\}$ a variável resposta para o cliente i (0 = “o i-ésimo cliente é ‘mau’ pagador”, 1= ”o i-ésimo cliente é ‘bom’ pagador”), o modelo de Regressão Logística pode ser escrito como,

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_o + \beta_1 x_{i1} + ... + \beta_p x_{ip} = x_i^T \beta \quad \text{ou} \quad \pi_i = \frac{\exp\left(x_i^T \beta\right)}{1+\exp\left(x_i^T \beta\right)} \qquad (25)$$

em que, π_i é a probabilidade do cliente i ser “bom” pagador, $x_i = (1, x_{i1}, ..., x_{ip})^T$ é o vetor de variáveis preditoras do cliente i e $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ é o vetor dos parâmetros (coeficientes) do modelo.

Como a variável resposta Y_i tem distribuição Bernoulli com probabilidade de sucesso π_i, então:

- $E(Y_i = 1 \mid x_1,...., x_p) = P(Y_i = 1 \mid x_1,...., x_p) = \pi_i$, que é a probabilidade de que o cliente seja 'bom' pagador dado as variáveis preditoras.
- $E(Y_i = 0 \mid x_1,...., x_p) = P(Y_i = 0 \mid x_1,...., x_p) = 1 - \pi_i$, que é a probabilidade de que o cliente seja 'mau' pagador dado as variáveis preditoras.

A distribuição variável resposta Y_i para cada observação da amostra de clientes é dada por,

$$P(Y_i = y_i) = \pi_i^{\,y_i} (1 - \pi_i)^{1-y_i} \qquad \textit{para i=1,...,n} \tag{26}$$

O método da Máxima Verossimilhança será utilizado para estimar os parâmetros $\beta_0, \beta_1,..., \beta_p$ do modelo de Regressão Logística múltipla. Para isso, inicialmente escreveremos a função de Verossimilhança (BOLFARINE; SANDOVAL, 2001) da variável resposta Y_i em todas as observações da amostra $y=(y_1,y_2,...,y_n)$ sob o modelo de Regressão Logística como,

$$L(y_1, y_2,..., y_n, \beta) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} \pi_i^{\,y_i} (1 - \pi_i)^{1-y_i} \tag{27}$$

Por conveniência de cálculos, trabalharemos com a log-verossimilhança que é dada por,

$$\ln\big(L(y_1, y_2,..., y_n, \beta)\big) = \ln\left(\prod_{i=1}^{n} P(Y_i = y_i)\right) = \sum_{i=1}^{n}\left[y_i . \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} \ln(1 - \pi_i) \tag{28}$$

Da expressão do modelo de Regressão Logística dada pela equação (25), temos que $1 - \pi_i = \left[1 + \exp\left(x_i^T \beta\right)\right]$ e $\ln\left(\pi_i / (1 - \pi_i)\right) = x_i^T \beta$. Então, a log-verossimilhança pode ser reescrita como,

$$\ln\big(L(y_1, y_2,..., y_n, \beta)\big) = \sum_{i=1}^{n} y_i x_i^T \beta - \sum_{i=1}^{n} \ln\left(1 + \exp\left(x_i^T \beta\right)\right) \tag{29}$$

As estimativas de Máxima Verossimilhança, $\hat{\beta}$, são os valores de β que maximizam a log-verossimilhança, dada pela equação (29), e são obtidos através de métodos numéricos. O método numérico mais utilizado é o de Mínimos Quadrados Reponderados (PAULA, 2004).

Seja $\hat{\beta}$ a estimativa dos parâmetros do modelo de Regressão Logística múltipla, obtida utilizando métodos numéricos. Se suposições do modelo são corretas, então podemos mostrar que assintoticamente,

$$E\left(\hat{\beta}\right) = \beta \qquad \text{e} \qquad Var\left(\hat{\beta}\right) = \left(X^T V^{-1} X\right)^{-1} \tag{30}$$

em que, $V = diag\{\pi_1(1-\pi_1), \pi_2(1-\pi_2), ..., \pi_n(1-\pi_n)\}$ [3] e $X = (x_1, x_2, ..., x_n)^T$.

Os testes de significância para cada parâmetro do modelo serão feitos e suas hipóteses são:

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$ $\qquad$ $j=1,2,...,p$

O teste de Wald (MONTGOMERY; PECK; VINING, 2001) pode ser utilizado para avaliar a significância de cada parâmetro e sua estatística é dada por,

$$Z_o = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{31}$$

em que, $se(\hat{\beta}_j)$ é o erro padrão de $\hat{\beta}_j$, dado por $se(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)/n}$.

Sob a hipótese nula, H_0, a estatística do teste de Wald, Z_0, tem distribuição Normal com $\mu = 0$ e $\sigma = 1$ (normal padrão).

O uso de Regressão Logística tem se consagrado em modelos de *Credit Scoring* devido a algumas vantagens oferecidas pela técnica (ROSA, 2000):

- É a mais utilizada entre os profissionais da área (culturalmente difundida);

[3] diag{a,b,c} é uma matriz diagonal com elementos a, b e c.

- Não apresenta problemas sérios de suposições, como, por exemplo, a Análise Discriminante Linear, na qual se pressupõe uma distribuição Normal Multivariada para as variáveis preditoras;
- Facilidade computacional, uma vez que os pacotes estatísticos mais utilizados pelas instituições permitem o seu uso;
- É uma ferramenta poderosa para discriminação e é aplicável aos dados de risco de crédito.

Vasconcellos (2002) também aponta que a Regressão Logística é uma técnica vantajosa devido à velocidade no processo de classificação (quanto tempo o cliente que pede um empréstimo precisa esperar para ter uma resposta afirmativa ou negativa sobre a concessão) e devido à facilidade de revisar o modelo periodicamente.

3.4 Medidas de Avaliação dos Modelos de Classificação

Os modelos de *Credit Scoring* têm como principal objetivo discriminar os clientes que se tornarão inadimplentes com o tempo dos que se manterão bons pagadores. Existem diversas medidas utilizadas para mensurar e comparar os desempenhos de modelos de classificação na realização deste propósito. Na Seção 3.4 serão apresentadas duas medidas de avaliação do poder discriminante dos modelos: a estatística *Kolmogorov-Smirnov* e o coeficiente Gini (ANDERSON, 2007). Além disso, também serão apresentadas outras três taxas de acerto, baseadas da Matriz de Confusão, utilizadas para avaliar a acurácia das previsões dos modelos de classificação: taxas de acerto totais (TAT), taxas de acerto nas previsões dos clientes “bons” (TAB) e dos clientes “maus” (TAM) pagadores (ROSA, 2000). Os melhores modelos de classificação serão os com maior poder discriminante e com maiores taxas de acerto nas previsões dos clientes “bons” e, principalmente, dos clientes “maus” pagadores.

3.4.1 Matriz de Confusão

A Matriz de Confusão de um modelo de *Credit Scoring* é uma maneira fácil de observar se o modelo está prevendo adequadamente os “bons” e “maus” clientes. Para a sua construção deve-se atribuir a cada indivíduo *i* da amostra e validação um *score* s_i. A variável *score* é a probabilidade prevista do cliente se não se tornar inadimplente, dada a observação das variáveis preditoras do modelo, e assume valores entre 0 e 100. Se $s_i > P_c$, então o cliente é classificado como “bom” pagador e, caso contrário, o cliente é classificado como “mau” pagador. O ponto de corte utilizado, P_c, foi de 50. A matriz de confusão (Tabela 2), apresenta as freqüências do cruzamento entre classificações observadas e previstas por um modelo, dado um determinado ponto de corte (ANDERSON, 2007).

Tabela 2 – Matriz de confusão de um modelo de *Credit Scoring*.

Observado	Previsto		Total
	Mau	Bom	
Mau	n_{00}	n_{01}	$n_{0.}$
Bom	n_{10}	n_{11}	$n_{1.}$
Total	$n_{.0}$	$n_{.1}$	$n_{..}$

em que,

n_{oo}: Número de clientes “maus” corretamente classificados como “maus”;

n_{o1}: Número de clientes “maus” incorretamente classificados como “bons”;

n_{1o}: Número de clientes “bons” corretamente classificados como “maus”;

n_{11}: Número de clientes “bons” incorretamente classificados como “bons”;

e, $n_{.0} = n_{00}+n_{10}$; $n_{.1} = n_{01}+n_{11}$; $n_{0.} = n_{00}+n_{01}$; $n_{1.} = n_{10}+n_{11}$; $n_{..}=n_{00}+n_{01}+n_{10}+n_{11}$

Com isso, os modelos de classificação serão avaliados e comparados a partir das três taxas de acerto definidas por,

- Taxa de acerto total: $TAT = (n_{00} + n_{11}) / n_{..}$ (32)
- Taxa de acerto dos “maus”: $TAM = (n_{00}) / n_{0.}$ (33)
- Taxa de acerto dos “bons”: $TAB = (n_{11}) / n_{1.}$ (34)

A taxa de acerto dos “bons” (TAB) também pode ser chamada de sensibilidade ou *true positive rate*, e a taxa de acerto dos “maus” (TAB) também pode ser chamada de especificidade ou *false positive rate.* Outras medidas de avaliação de modelos de classificação binária são os erros tipo I e do tipo II, definidos como (ANDERSON, 2007)

$$\text{Erro tipo I} = (n_{10}) / n_1. \tag{35}$$

$$\text{Erro tipo II} = (n_{01}) / n_0. \tag{36}$$

Uma desvantagem do uso das taxas de acerto obtidas da matriz de confusão para avaliar a assertividade das previsões dos modelos é que estas medidas dependem do ponto de corte escolhido.

3.4.2 Coeficiente Gini

Coeficiente Gini é duas vezes a área entre a curva ROC (*Receiver Operating Characteristic*) e a diagonal da curva (ANDERSON, 2007). O coeficiente Gini é utilizado para avaliar se o *score* previsto discrimina bem os clientes “bons” e “maus” pagadores.

A curva ROC é obtida do gráfico da sensibilidade versus a especificidade das previsões de um modelo de classificação binária (com variável resposta com 2 níveis), com o ponto de corte P_c variando. Quanto maior a sensibilidade e a especificidade melhor o modelo. No entanto, ambas as medidas dependem de P_c, e quanto P_c cresce, a sensibilidade diminui e a especificidade aumenta. Com isso, para a construção da curva ROC, obtém-se as matrizes de confusão para diferentes pontos de corte (P_c) e delas calcula-se a sensibilidade e especificidade. A Figura 9 mostra um exemplo de construção da curva ROC.

O coeficiente Gini é calculado utilizando a seguinte expressão:

$$\textit{Coeficiente Gini} = 1 - \sum_{i=1}^{n} \left(F_M(s_i) - F_M(s_{i-1}) \right) \left(F_B(s_i) - F_B(s_{i-1}) \right) \tag{37}$$

em que, $F_B(s_i)$ é a distribuição acumulada dos clientes “bons” na faixa de *score i*, $F_M(s)$ é a distribuição acumulada do *scores* dos clientes “maus” na faixa de escore *i* e *n* é o número de faixas de score (será aplicado n=1000).

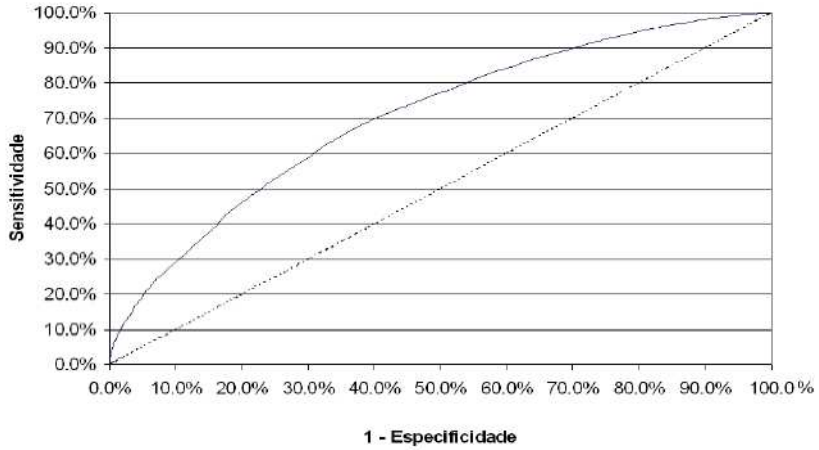

Figura 9 - Exemplo de curva ROC (*Receiver Operating Characteristic*).

O valor do coeficiente Gini representa o poder de discriminação dos clientes “bons” e “maus” por um modelo de classificação binária em todos os intervalos de valores do escore.

3.4.3 Estatística Kolmogorov-Smirnov

A estatística de Kolmogorov-Smirnov (KS) é descrita pela teoria estatística não-paramétrica e é utilizada para testar se as distribuições de dois grupos são iguais (CONOVER, 1999).

Em modelos de *Credit Scoring*, a estatística KS mede a capacidade da variável escore de distinguir “bons” e “maus” clientes, lembrando que a variável escore é o valor da probabilidade prevista do cliente se tornar inadimplente, dada a observação das variáveis preditoras, e assume valores entre 0 e 100.

Para a avaliação da performance de modelos de *Credit Scoring*, a estatística KS é definida como a máxima diferença entre as distribuições acumuladas dos escores dos “bons” e “maus” pagadores (ANDERSON, 2007) e é definida como,

$$KS = \max_{s} \left\{ \left| F_M(s) - F_B(s) \right| \right\} \quad (38)$$

em que, $F_B(s)$ é a distribuição acumulada do escores entre os clientes "bons" e $F_M(s)$ é a distribuição acumulada do escores entre os clientes "maus".

A hipótese da estatística KS supõe que um modelo de classificação com bom desempenho atribui aos clientes "bons" pagadores escores altos e a clientes "maus" pagadores escores baixos. Logo, a distribuição dos escores dos clientes "bons" apresenta maior concentração em valores altos e a distribuição de escores dos clientes "maus" possui maior concentração em valores mais baixos. Além disso, a distribuição acumulada do escore dos "maus" pagadores é superior à distribuição acumulada dos escores dos "bons" pagadores e portanto, o melhor modelo dever prover a maior separação entre clientes adimplentes e inadimplentes ao longo dos valores de escore.

Na Figura 10, é apresentado um exemplo de cálculo da estatística KS. A maior separação entre as distribuições acumuladas de "bons" e "maus" é 30% e portanto, o valor da estatística KS é 30%.

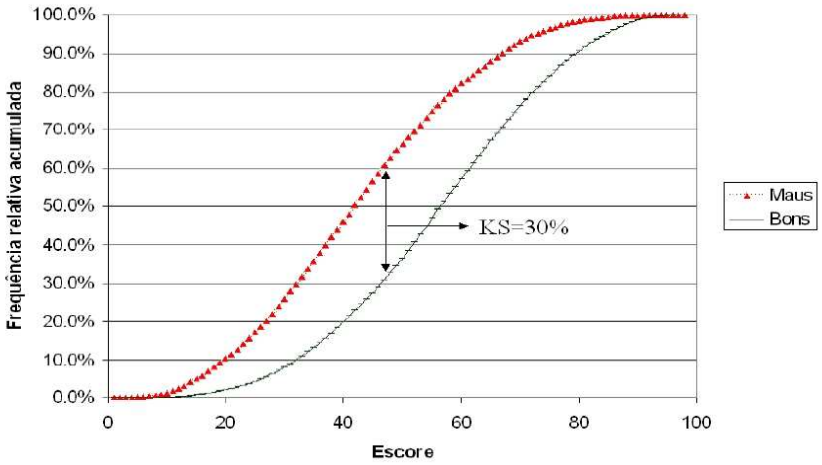

Figura 10 - Exemplo de cálculo da estatística Kolmogorov-Smirnov.

Lecumberri e Duarte (2003) descrevem uma regra prática para a verificação da qualidade de modelos de *Credit Scoring* utilizando à estatística KS (Tabela 3). Esta regra visa auxiliar na interpretação da estatística KS. Por exemplo, no caso de um modelo de *Credit Scoring* cuja distância está abaixo de 20%, há forte indício de um

baixo nível de discriminação no modelo, o que sugere a necessidade de alteração do mesmo.

Tabela 3 - Valores críticos da estatística Kolmogorov-Smirnov aplicada em modelos de *Credit Scoring*.

Estatística KS	**Discrimanação *Credit Scoring***
<15%	Discriminação Muito Baixa
15 a 25%	Discriminação Baixa
25 a 35%	Discriminação Aceitável
35 a 45%	Discriminação Boa
45 a 55%	Discriminação Excelente
55 a 65%	Discriminação Excelente
65 a 75%	Discriminação Excelente
>75%	Discriminação Excelente

4. METODOLOGIA

No Capítulo de Metodologia será descrita a base de dados aplicada, a *German Credit*, e todos os procedimentos adotados para a construção dos modelos de classificação. Na Seção 4.2, será descrita a categorização das variáveis preditoras da base de dados *German Credit*, adotada previamente a aplicação dos modelos. Na Seção 4.3, serão apresentadas as seleções de variáveis que serão empregadas nos Classificadores Bayesianos e da Regressão Logística. Na Seção 4.3, será mostrada a metodologia de construção das amostras de desenvolvimento e de validação obtidas por Cross-Validation com 10 partições (10-fold). Na Seção 4.5, finalmente serão mencionados os softwares empregados na estimação dos modelos de classificação. A Figura 11 apresenta um breve esquema destas metodologias empregadas.

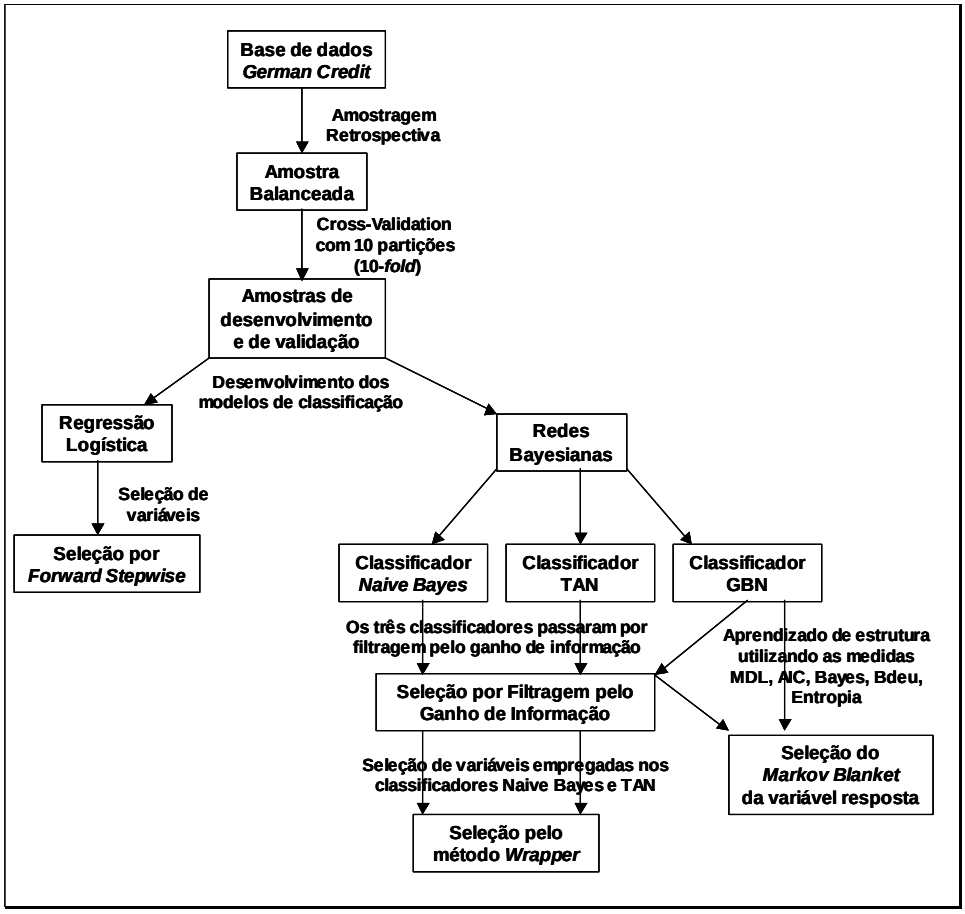

Figura 11 – Esquema das metodologias aplicadas neste trabalho.

4.1 Base de Dados

A base de dados de empréstimos concedidos utilizada será a *German Credit*, disponível no Repositório *UCI Machine Learning* (Blake e Merz, 1998). Nesta base de dados, há informações pessoais e financeiras de clientes proponentes a um empréstimo e a classificação destes clientes de acordo com sua inadimplência ou adimplência (Cliente “bom” ou “mau” pagador) no pagamento do empréstimo.

A classificação dos clientes de acordo com seu risco de inadimplência será a variável resposta, chamada de “Cliente”, assumindo valor 1, se o cliente foi previamente classificado como um “bom” pagador (adimplente), e valor 0, se o cliente foi classificado como um “mau” pagador (inadimplente).

A base de dados contém 1000 clientes, dos quais 700 foram previamente classificados como “bons” pagadores e 300 como “maus” pagadores. Além de, 20 variáveis preditoras categóricas ou contínuas (Tabela 4), representando as informações pessoais e financeiras dos clientes.

A amostra aplicada nos modelos de classificação apresenta mesmo número de clientes “bons” e “maus” pagadores, ou seja, amostra balanceada, e foi obtida por Amostragem Retrospectiva (PAULA, 2004). Este esquema de amostragem consistiu em manter a amostra de clientes “maus”, com 300 clientes, e selecionar aleatoriamente uma amostra de mesmo tamanho (300) de clientes “bons”.

A Amostragem Retrospectiva foi adotada a fim de que a diferença entre os tamanhos das amostras de clientes “bons” e “maus” não criasse nenhum viés nos modelos de classificação e, com isso, evitar que os modelos sejam adequados para discriminar os clientes “bons”, porém ineficientes para discriminar os clientes “maus” (ROSA, 2000).

Tabela 4 - Variáveis da base de dados *German Credit*.

Variável	Tipo de Variável	Núm. de Categorias	Categorias	Nome original (em inglês)
Salário	categórica ordinal	4	X < $0, 0 <= X < 200 , X >=200, Sem remuneração	Status of existing checking account
Duração do Empréstimo	Contínua	-	-	Duration in months
Histórico de Crédito	categórica ordinal	4	Sem empréstimos tomados, Todos os empréstimos pagos pontualmente, Existem empréstimos pagos pontualmente , Histórico de atraso no pagamento, Atraso no pagamento ou com empréstimos tomados em outras Instituições	Credit history
Finalidade	categórica nominal	11	Compra de carro novo, Compra de carro usado, Móveis, Rádio e TV, Utensílios domésticos, Reforma, Educação, Cursos, Negócios, Outros	Purpose
Valor do Empréstimo	Contínua	-	-	Credit amount
Poupança do Cliente	categórica ordinal	5	< $100, $100<= X < $500, $500<= X < $1000, >= $1000, Não possui ou não conhecida	Savings account/bonds
Tempo de Trabalho	categórica ordinal	5	Desempregado, X < 1 ano, 1 <= X < 4 anos, 4 <= X < 7 anos, X >= 7 anos	Present employment since
Taxa de juros em % do valor do empréstimo	Contínua	-	-	Installment rate in % of disposable income
Estado Civil e Sexo	Categórica nominal	5	Masculino divorciado ou separado, Feminimo divorciada, separada ou casada, Masculino solteiro, Masculino casado ou viúvo	Personal status and sex
Outras dívidas ou garantias	categórica nominal	3	Nenhuma, Co-aplicante, Fiador	Other debtors/ Guarantors
Tempo de Residência	Contínua	-	-	Present residence since
Bens	categórica nominal	4	Imóvel, Seguro de Vida, Carro ou outros, Não possui bens	Property
Idade	Contínua	-	-	Age in years
Outros Empréstimos	categórica nominal	3	Bancos, Lojas, Nenhum	Other installment plans
Moradia	categórica nominal	3	Alugada, Própria, Moradia gratuita	Housing
Número de créditos Concedidos em seu banco	Contínua	-	-	Number of existing credits at this bank
Emprego	categórica ordinal	4	Desempregado ou empregado com baixa qualificação ou sem trabalho formal, Empregado com baixa qualificação e com trabalho formal, Empregado qualificado ou funcionário público, Executivo, profissional liberal, empregado altamente qualificado ou oficial	Job title
Número de dependentes	Contínua	-	-	Number of people being liable to provide maintenance for
Telefone Próprio	categórica binária	2	Sim, Não	Telephone
Estrangeiro	categórica binária	2	Sim, Não	Foreign worker
Cliente	categórica binária	2	Bom ou Mau	Good or bad credit rating

A amostra balanceada será utilizada na estimação dos modelos de classificação, mas, para a aplicação dos modelos à população original, é necessário que alguns de seus parâmetros sejam re-calculados.

No modelo de Regressão Logística é necessário que o intercepto seja re-calculado (PAULA, 2004) por:

$$\hat{\beta}_o = \hat{\beta}_0^* - \ln\left(\frac{\gamma_1}{\gamma_2}\right) \tag{39}$$

em que, $\hat{\beta}_0^*$ é o intercepto do modelo logístico ajustado e $\gamma_1 = P\left(Z = 1 \middle| Cliente = 1\right)$ e $\gamma_2 = P\left(Z = 1 \middle| Cliente = 0\right)$, sendo Z uma variável indicadora da seleção amostral em relação à amostra toda.

Nas RB é necessário que a distribuição da variável resposta (e de seus pais, se houverem) seja substituída pela distribuição desta variável na amostra original.

4.2 Categorização das Variáveis

Os modelos de RB avaliados possuem premissa de que todas as suas variáveis sejam discretas (ou categóricas). Como a base de dados *German Credit* apresenta variáveis contínuas (ou numéricas), então será adotado o procedimento de categorização destas variáveis.

Rosa (2000) explica que, se as variáveis do modelo puderem ser mostradas em categorias, tornam-se mais simples a implementação dos modelos e a interpretação dos pesos relativos às categorias das variáveis. Outro inconveniente de se trabalhar com variáveis contínuas é a aparição de valores discrepantes (*outliers*), cuja presença costuma afetar consideravelmente os resultados dos modelos.

A categorização de cada variável contínua consiste na criação de níveis (categorias) de uma variável discreta que correspondam a intervalos de valores da

variável contínua original. Ao final do procedimento, a variável discreta resultante é usada no lugar da variável contínua. As variáveis originalmente categóricas com muitos níveis também tiveram alguns de seus níveis reagrupados para evitar a existência de categorias com número muito pequeno de observações (ou pouco significativo), o que pode prejudicar a estimação dos parâmetros dos modelos.

A categorização de variáveis contínuas adotada consistiu na construção dos níveis das variáveis de acordo com a relação da variável preditora com a variável resposta (Cliente “bom” ou “mau”) (HAND; HENLEY, 1997). Esta análise bivariada da relação da variável preditora com a variável resposta foi feita através do cálculo do Risco Relativo e do WOE de cada nível das variáveis preditoras.

Para as variáveis contínuas, o procedimento consistiu em inicialmente criar uma categorização inicial da variável, a partir dos percentis da distribuição de cada variável contínua. Assim, foram criados 10 níveis com cada faixa de valores correspondendo aos decis (percentil de ordem 10%) da variável contínua original, ou seja, criadas faixas de valores da variável ordenada a cada 10%. A partir desta categorização inicial, foi verificada a freqüência de clientes “bons” e “maus” em cada categoria criada (faixa de valores), a fim de identificar categorias semelhantes com relação a “bons” e “maus” clientes (Tabela 5). Para isso, serão calculadas as seguintes medidas:

- Risco Relativo (AGRESTI, 1999): proporção de “bons” na categoria sobre a proporção de “maus” na categoria;
- “*Weights of Evidence*” (WOE) (HAND; HENLEY, 1997) que é o logaritmo natural do Risco Relativo (Razão de “bons” e “maus”);

Tabela 5 - Exemplo de categorização de uma variável preditora.

Categoria	Número de "bons"	Número de "maus"	%bons	%maus	Risco Relativo	WOE
Categoria 1	b_1	m_1	b_1/b.	m_1/m.	(b_1/b.)/(m_1/m.)	ln[(b_1/b.)/(m_1/m.)]
Categoria 2	b_2	m_2	b_2/b.	m_2/m.	(b2/b.)/(m2/m.)	ln[(b_2/b.)/(m_2/m.)]
Categoria 3	b_3	m_3	b_3/b.	m_3/m.	(b3/b.)/(m3/m.)	ln[(b_3/b.)/(m_3/m.)]
Categoria 4	b_4	m_4	b_4/b.	m_4/m.	(b4/b.)/(m4/m.)	ln[(b_4/b.)/(m_4/m.)]
Categoria 5	b_5	m_5	b_5/b.	m_5/m.	(b5/b.)/(m5/m.)	ln[(b_5/b.)/(m_5/m.)]
Total	b.	m.	1	1	1	0

O Risco Relativo e o WOE são medidas descritivas que auxiliam na identificação de categorias das variáveis com alto ou baixo poder de discriminação dos clientes bons e maus pagadores, e também auxiliam a identificar as categorias que discriminam melhor os clientes “bons” e as que discriminam melhor os clientes “maus“. Estas medidas podem ser analisadas da seguinte maneira:

- WOE = 0 (Risco Relativo = 1): indica que a razão entre “bons” e ”maus” é 1 e, portanto, se a variável assumir o valor correspondente a esta categoria não há nenhum indício do cliente apresentar maior ou menor risco de inadimplência, se comparado à análise desconsiderando esta variável;
- WOE > 0 (Risco Relativo>1): positivo e quanto mais distante de zero, maiores são as chances de o cliente apresentar menor risco de crédito, indicando que a categoria apresenta algum poder para discriminar clientes bons;
- WOE < 0 (Risco Relativo<1): negativo e quanto mais distante de zero, maiores são as chances de o cliente apresentar maior risco de crédito, indicando que a categoria apresenta algum poder para discriminar clientes “maus”;

O Risco Relativo e o WOE também podem ser utilizados para agrupar categorias com valores próximos, ou seja, categorias com risco de inadimplência próximas. No entanto, este agrupamento só pode ser feito se houver interpretação lógica.

A Tabela 6 mostra o resultado da categorização das variáveis contínuas da base de dados *German Credit.* Analisando esta tabela observa-se que as categorias que contribuíram para maiores riscos de inadimplência, sem considerar o efeito das demais, foram: clientes com empréstimos de mais de 3 anos (36 meses), com valor acima de $7500 ou com idade inferior a 25 anos. Por outro lado, as características dos clientes ou do empréstimo que individualmente apresentaram maiores riscos relativos e, portanto, apresentam menores riscos de inadimplência, foram: empréstimos com duração menor que 12 meses, e com valor entre $1000 e $4000, taxas menores que 4% do valor do empréstimo, pessoas entre 30 e 31 anos, entre 35 e 50 anos ou com mais de dois créditos concedidos em seu banco.

As variáveis originalmente categóricas, e que tiveram algumas de suas categorias agrupadas, foram: Histórico de Crédito, Bens, Poupança do Cliente, Outros Empréstimos, Tempo de Trabalho, Moradia, Estado Civil e Sexo, Emprego. A categorização original destas variáveis é mostrada nas A.2 a A.9, do Apêndice.

A Tabela 7 mostra os valores do risco relativo e do *Weights of Evidence* (WOE) das variáveis categóricas da base de dados, que sofreram ou não agrupamento de seus níveis. Não houve agrupamento de níveis da variável "Finalidade", pois não faz sentido para o processo de concessão de crédito agrupar características muito distintas de finalidade do empréstimo. Uma peculiaridade observada na base de dados *German Credit* é a ausência de clientes do sexo feminino e com estado civil "solteira".

Tabela 6 - Categorização das variáveis originalmente contínuas da base de dados *German Credit* e cálculo do Risco Relativo e *Weights of Evidence* (WOE) de cada categoria.

Variável Preditora	Nível	Número de "bons"	Número de "maus"	%bons	%maus	Risco Relativo	WOE
Duração do empréstimo	<12	69	27	23.00%	9.00%	2.556	0.938
	12<= X<16	82	62	27.33%	20.67%	1.323	0.280
	16<= X<36	110	129	36.67%	43.00%	0.853	-0.159
	>=36	39	82	13.00%	27.33%	0.476	-0.743
Valor do Empréstimo	<1000	25	37	8.33%	12.33%	0.676	-0.392
	1000<= X<4000	216	158	72.00%	52.67%	1.367	0.313
	4000<= X<7500	42	61	14.00%	20.33%	0.689	-0.373
	>=7500	17	44	5.67%	14.67%	0.386	-0.951
Taxa de juros em % do valor do empréstimo	<4	170	141	56.67%	47.00%	1.206	0.187
	>=4	130	159	43.33%	53.00%	0.818	-0.201
Tempo de Residência	<3	150	133	50.00%	44.33%	1.128	0.120
	3<= X<4	44	43	14.67%	14.33%	1.023	0.023
	>=4	106	124	35.33%	41.33%	0.855	-0.157
Idade	<25	32	61	10.67%	20.33%	0.525	-0.645
	25<= X<30	63	76	21.00%	25.33%	0.829	-0.188
	30<= X<32	30	22	10.00%	7.33%	1.364	0.310
	32<= X<35	28	33	9.33%	11.00%	0.848	-0.164
	35<= X<51	115	77	38.33%	25.67%	1.494	0.401
	>=51	32	31	10.67%	10.33%	1.032	0.032
Número de Dependentes	<2	248	254	82.67%	84.67%	0.976	-0.024
	>=2	52	46	17.33%	15.33%	1.130	0.123
Número de créditos concedidos em seu banco	< 2	177	200	59.00%	66.67%	0.885	-0.122
	>=2	123	100	41.00%	33.33%	1.230	0.207
Total		300	300	100.00%	100.00%	1.000	0.000

Tabela 7 - Risco Relativo e “Weights of Evidence” (WOE) das variáveis originalmente categóricas da base de dados *German Credit,* após o agrupamento de alguns de seus níveis.

Variável	Categoria	Número de "bons"	Número de "maus"	%bons	%maus	Risco Relativo	WOE
Bens	Imóvel	101	60	33.67%	20.00%	1.683	0.521
	Seguro de Vida, Carro e outros	161	173	53.67%	57.67%	0.931	-0.072
	Não possui bens	38	67	12.67%	22.33%	0.567	-0.567
Salário	X < $0	51	135	17.00%	45.00%	0.378	-0.973
	0 <= X < 200	76	105	25.33%	35.00%	0.724	-0.323
	X >=200	21	14	7.00%	4.67%	1.500	0.405
	Sem remuneração	152	46	50.67%	15.33%	3.304	1.195
Poupança do Cliente	< $500	189	251	63.00%	83.67%	0.753	-0.284
	>= $500	35	17	11.67%	5.67%	2.059	0.722
	Não possui ou não conhecida	76	32	25.33%	10.67%	2.375	0.865
Outros Empréstimos	Nenhum	259	224	86.33%	74.67%	1.156	0.145
	Bancos e Lojas	41	76	13.67%	25.33%	0.539	-0.617
Outras dívidas Ou garantias	Nenhuma	271	272	90.33%	90.67%	0.996	-0.004
	Co-aplicante	11	18	3.67%	6.00%	0.611	-0.492
	Fiador	18	10	6.00%	3.33%	1.800	0.588
Finalidade	Compra de carro novo	57	89	19.00%	29.67%	0.640	-0.446
	Compra de carro usado	36	17	12.00%	5.67%	2.118	0.750
	Móveis	52	58	17.33%	19.33%	0.897	-0.109
	Rádio e TV	96	62	32.00%	20.67%	1.548	0.437
	Utensílios domésticos	2	4	0.67%	1.33%	0.500	-0.693
	Reforma	9	8	3.00%	2.67%	1.125	0.118
	Educação	9	22	3.00%	7.33%	0.409	-0.894
	Cursos	5	1	1.67%	0.33%	5.000	1.609
	Negócios	29	34	9.67%	11.33%	0.853	-0.159
	Outros	5	5	1.67%	1.67%	1.000	0.000
Histórico de Crédito	Sem empréstimos tomados ou todos os empréstimos pagos pontualmente	12	53	4.00%	17.67%	0.226	-1.485
	Existem empréstimos pagos pontualmente	154	169	51.33%	56.33%	0.911	-0.093
	Histórico de atraso no pagamento	28	28	9.33%	9.33%	1.000	0.000
	Atraso no pagamento ou com empréstimos tomados em outras Instituições	106	50	35.33%	16.67%	2.120	0.751
Tempo de Trabalho	Desempregado ou menos de 1 ano	67	93	22.33%	31.00%	0.720	-0.328
	1 <= X < 4 anos	105	104	35.00%	34.67%	1.010	0.010
	X >= 4anos	128	103	42.67%	34.33%	1.243	0.217
Estado Civil e Sexo	Divorciado(a) ou separado(a) ou mulher casada	94	129	31.33%	43.00%	0.729	-0.317
	Masculino solteiro	179	146	59.67%	48.67%	1.226	0.204
	Masculino casado ou viúvo.	27	25	9.00%	8.33%	1.080	0.077
Moradia	Própria	233	186	77.67%	62.00%	1.253	0.225
	Alugada ou Moradia gratuita	67	114	22.33%	38.00%	0.588	-0.532
Emprego	Desempregado ou empregado com baixa qualificação	81	63	27.00%	21.00%	1.286	0.251
	Empregado qualificado ou funcionário público	176	186	58.67%	62.00%	0.946	-0.055
	Executivo, profissional liberal, empregado altamente qualificado ou oficial	43	51	14.33%	17.00%	0.843	-0.171

(continua na próxima página)

Continuação da Tabela 7 - Risco Relativo e “Weights of Evidence” (WOE) das variáveis originalmente categóricas da base de dados *German Credit* após o agrupamento de alguns de seus níveis.

Variável	Categoria	Número de "bons"	Número de "maus"	%bons	%maus	Risco Relativo	WOE
Telefone Próprio	Sim	116	113	38.67%	37.67%	1.027	0.026
	Não	184	187	61.33%	62.33%	0.984	-0.016
Estrangeiro	Sim	20	4	6.67%	1.33%	5.000	1.609
	Não	280	296	93.33%	98.67%	0.946	-0.056
	Total	300	300	100.00%	100.00%	1.000	0.000

Após o tratamento de categorização e de agrupamento de níveis das variáveis preditoras da base de dados *German Credit*, estas variáveis finais (todas categóricas) foram transformadas em variáveis *dummy* (ou indicadoras). As *dummies* são variáveis binárias e assumem valores 1 ou 0. O número de *dummies* é igual ao número de níveis (categorias) da variável preditora (*n*) menos um, ou seja, *n-1*. Cada *dummy* está associada a uma categoria da variável preditora e uma categoria da variável preditora original corresponde à casela de referência. As *dummies* assumem valor 1, se o cliente possui a característica representada por ela, e valor 0, se o cliente não possui tal característica. A casela de referência (categoria de referência) representará a categoria da variável preditora associada a todos os valores zero das *dummies* (e por isso são criadas *n-1 dummies*). Por exemplo, a variável “Duração do Empréstimo” com 4 categorias, "X<12", "12<=X<16", "16<=X<36" e "X>=36", serão criadas 3 *dummies* dadas por,

$$I_{\text{Duração do Empréstimo ("X<12")}} = \begin{cases} 1, \text{ se o empréstimo teve duração menor do que 12 meses} \\ 0, \text{ demais clientes} \end{cases}$$

$$I_{\text{Duração do Empréstimo ("12<=X<16")}} = \begin{cases} 1, \text{ se o empréstimo teve duração entre 12 e 15 meses} \\ 0, \text{ demais clientes} \end{cases}$$

$$I_{\text{Duração do Empréstimo ("16<=X<36")}} = \begin{cases} 1, \text{ se o empréstimo teve duração entre 16 e 35 meses} \\ 0, \text{ demais clientes} \end{cases} \quad (40)$$

A casela de referência será da variável “Duração do Empréstimo” é a dos clientes com empréstimos com duração superior a 36 meses.

Na construção dos modelos de classificação, as *dummies* serão utilizadas no lugar das variáveis preditoras originais e cada *dummy* corresponderá a uma

categoria das variáveis preditoras originas. O uso de *dummies* foi adotado para que nas seleções de variáveis preditoras, descritas na Seção 4.3 a seguir, fossem selecionadas somente as características dos clientes que mais contribuíssem para a discriminação entre os bons e maus pagadores, dado um modelo de classificação. Além disso, também se observou que no desenvolvimento dos Classificadores bayesianos o uso *dummies* contribuiu para uma melhora na assertividade das previsões dos clientes “bons” e “maus” dos modelos de classificação.

4.3 Seleção das Variáveis

Uma vez definido o conjunto de preditoras a ser utilizado no modelo de classificação, resta saber qual a melhor maneira de encontrar um modelo parcimonioso que inclua apenas as variáveis preditoras mais importantes para explicar a probabilidade do cliente ser um “bom” pagador em determinado período. A seleção de variáveis ainda pode melhorar o desempenho dos modelos ajustados, facilitar a visualização e entendimento dos parâmetros estimados, além de prevenir contra problemas de *overfitting* dos modelos estimados (GUYON; ELISEEFF, 2003).

A seleção de variáveis adotada no modelo de Regressão Logística será a *Forward Stepwise* (HOSMER, LEMESHOW, 1989), utilizando como critério a medida AIC (*Akaike information criterion*). O *Forward Stepwise* consiste em um algoritmo de exclusão e inclusão de variáveis preditoras, segundo sua importância de acordo com o critério de Akaike (AIC). A medida AIC leva em consideração tanto a log-verossimilhança (log-likelihood) dos dados, quanto o número de parâmetros do modelo ajustado, sendo que um modelo é melhor do que outro se apresentar menor valor da medida AIC.

Resumidamente, a seleção por *Forward Stepwise* inicia-se pela estimação de um modelo logístico com apenas o intercepto, seguida da estimação dos modelos logísticos com uma variável preditora. A variável incluída é a aquela cujo modelo apresentou menor valor do AIC, em relação ao modelo com somente o intercepto. Partindo do modelo com uma variável incluída, as demais variáveis são introduzidas

individualmente. A próxima variável incluída será aquela cujo modelo ajustado apresentar menor AIC, em relação ao modelo sem inclusão da variável. Enquanto isso, as variáveis que entram no modelo podem ser removidas, se a sua exclusão individual contribuir para uma redução do AIC do modelo com todas as variáveis incluídas. Na seleção *Forward Stepwise*, as variáveis incluídas podem ser removidas, pois na presença de outras variáveis estas podem não ter mais importância para o modelo. A inclusão e remoção de variáveis são repetidas até que nenhuma variável possa mais ser incluída ou excluída do modelo.

A seleção de variáveis dos Classificadores bayesianos será feita em duas etapas. Na primeira, há a ordenação e filtragem de variáveis preditoras pelo seu ganho de informação, em relação à variável reposta. Na segunda etapa, há a seleção de variáveis aplicando o método *Wrapper*, que é um algoritmo de caixa-preta que utiliza o próprio modelo de classificação para selecionar variáveis.

Na primeira etapa, as variáveis preditoras (*dummies*) são ordenadas em função do seu ganho de informação (*Information Gain*), em relação à variável resposta (Cliente “bom” ou “mau”) (Witten e Frank, 2005). Após esta ordenação, as variáveis serão retiradas pelo método *Backward*, ou seja, partindo do modelo saturado (com todas as variáveis) cada variável, que individualmente menos contribui para o ganho de informação da variável resposta, será retirada, sem que isso prejudique a assertividade das previsões dos Classificadores Bayesianos. Os resultados deste procedimento para os Classificadores *Naive Bayes*, TAN e GBN são apresentados na Seção 5.3. Esta seleção descrita tem como finalidade a busca do subconjunto de variáveis que isoladamente mais contribuem para a discriminação entre clientes “bons” e “maus”, sem considerar o efeito das demais.

A literatura científica a seleção pelo ganho de informação como um método de filtragem de variáveis, pois o critério de seleção se baseia no ganho de informação das variáveis preditoras, em relação à resposta, e não no modelo de classificação em si. Com isso, esta seleção pode ser encarada como um pré-processamento de dados (GUYON; ELISSEEFF, 2003). No entanto, nesse trabalho, o modelo de classificação foi utilizado como um critério de parada para a retirada de variáveis, para que fossem

removidas somente as variáveis redundantes e nenhuma variável importante, que pudesse prejudicar na acurácia dos modelos, fosse retirada.

O ganho de informação (*InfoGain*) de uma variável discreta em relação à variável resposta (*Classe*), também discreta, é dado pela diferença,

$$InfoGain(Classe, X) = H(Class) - H\left(Class \middle| X\right) \tag{41}$$

em que, $H(Class)$ é a Entropia da variável resposta e $H\left(Class \middle| X\right)$ é a Entropia condicional da variável resposta dado a variável *X*.

Seja *X* uma variável aleatória discreta com distribuição de probabilidades *P(X)* e com *n* observações dadas por $x_1, x_2, ..., x_n$. A Entropia marginal da variável X é dada por,

$$H(X) = -\sum_{i=1}^{n} P(x_i) \ln\left(P(x_i)\right) \tag{42}$$

E, a Entropia Condicional de X dado Y é dada por,

$$H(X \middle| Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_i) \ln\left(P(x_i \middle| y_i)\right) \tag{43}$$

sendo Y discreta com observações dadas por $y_1, y_2, ..., y_m$.

Após a filtragem, a seleção de variáveis dos Classificadores *Naive Bayes* e TAN realizada aplicando o método *Wrapper* (Witten e Frank, 2005) e nos Classificadores GBN serão selecionadas as variáveis do *Markov Blanket* da variável resposta.

O método *Wrapper* aplica o próprio modelo de classificação para avaliar o conjunto de variáveis preditoras e utiliza o esquema de *Cross-validation* para estimar a acurácia de cada conjunto avaliado. O método leva em consideração as premissas de cada Classificador bayesiano para obter o subconjunto de variáveis preditoras mais importantes para o modelo. Guyon e Elisseeeff (2003), descrevem que o método *Wrapper* é freqüentemente criticado por parecer um método de “força bruta”, exigindo um grande esforço computacional para ser realizado. Mas isso pode não ocorrer necessariamente, pois estratégias eficientes de busca têm sido

desenvolvidas visando à obtenção de métodos de busca mais eficientes e que não prejudiquem o desempenho das previsões do modelo avaliado. Métodos de busca gulosa (*Greedy search*) possuem vantagens computacionais e são robustos contra problemas de *overfitting* dos dados. As metodologias de busca gulosa existentes são a *forward selection* e *backward elimination*, sendo que na *forward selection* as variáveis são progressivamente incorporadas nos subconjuntos de variáveis enquanto a *backward elimination* inicia com um conjunto com todas as variáveis que são eliminadas sem que isso comprometa o desempenho do modelo.

Já nos Classificadores GBN, a seleção do subconjunto de variáveis preditoras do *Markov Blanket* da variável resposta é um procedimento natural de seleção, pois estas variáveis "protegem" a variável resposta da influência de qualquer outra variável de fora do seu *Markov Blanket*. Alguns estudos, mencionados no Capítulo 2, demonstraram que os Classificadores GBN com somente as variáveis do *Markov Blanket* da variável resposta resultam em modelos parcimoniosos e poderosos em diversas aplicações, além de ser uma maneira de contornar o problema de *overfitting* (CHENG; GREINER, 2001).

4.4 Amostra de Desenvolvimento e Validação

Como a base de dados *German Credit* apresentam um número grande de clientes (observações), será utilizado o conceito de *Cross-Validation* (WITTEN; FRANK, 2005) para obtenção das amostras de desenvolvimento e de validação. O procedimento de *Cross-Validation* aplicado consiste em dividir aleatoriamente todos os clientes da base de dados em 10 partições amostrais de tamanhos iguais. Destas partições, serão construídos 10 conjuntos, cada um com 9 partições para desenvolvimento do modelo e 1 para validação do modelo. Com isso, serão aplicadas as técnicas em cada um dos 10 grupos de desenvolvimento do modelo e os resultados serão avaliados em cada um dos respectivos grupos de validação.

As 9 partições de desenvolvimento, ou seja, 90% das observações da base de dados, são utilizadas para estimar os parâmetros dos modelos. As amostras de

validação são utilizadas para avaliar a capacidade de generalização do modelo, ou seja, se o modelo estimado mantiver seu poder de discriminação em amostras provindas da mesma população da amostra de desenvolvimento, então ele é estável. Mas, se seu poder de discriminação variar muito de uma amostra para outra, então ele pode estar com problema de *overfitting*.

Em bases de dados com um número reduzido de observações e muitas variáveis preditoras, como é o caso da *German Credit*, o procedimento de *Cross-Validation* é vantajoso, pois todos os clientes “bons” e “maus” são utilizados tanto para o desenvolvimento dos modelos quanto para a validação dos modelos. Além dos modelos serem desenvolvidos com uma alta proporção de observações da base de dados.

4.5 Softwares Utilizados

O modelo de Regressão Logística com seleção de variáveis por *Forward Stepwise* será estimado utilizando o software estatístico *R,* que é um software livre com as principais ferramentas de estatística desenvolvido por pesquisadores da área de Estatística do mundo todo.

Os Classificadores bayesianos *Naive Bayes*, TAN e GBN serão aplicados utilizando o software Weka (*Waikato Environment for Knowledge Analysis*), versão 3.5.8. O software Weka foi utilizado em todas as etapas de construção dos Classificadores bayesianos, como: a seleção de variáveis pelo método *Wrapper*, filtragem pelo Ganho de Informação em relação à variável resposta, aprendizados de estrutura utilizando as medidas Bayes, MDL, AIC, Bdeu e Entropia e também inferências para a interpretação e predição dos Classificadores.

O software WEKA foi desenvolvido na Universidade de Waikato na Nova Zelândia para tarefas de *Data Mining* e possui uma coleção de algoritmos de Aprendizado de Máquina, área que combina Estatística com Inteligência Artificial.

5. RESULTADOS

No capítulo de Resultados, serão apresentadas as aplicações da Regressão Logística e dos Classificadores Bayesianos em modelos de *Credit Scoring*, utilizando a base de dados *German Credit*. Nas Seções 5.1 e 5.2, serão avaliados e interpretados os modelos de Regressão Logística e os Classificadores Bayesianos *Naive Bayes*, TAN e GBN. Na Seção 5.4, os resultados dos modelos ajustados serão comparados.

Os modelos de classificação foram estimados a partir da amostra balanceada, com 300 clientes “bons” pagadores e 300 clientes “maus” pagadores, obtida por Amostragem Retrospectiva, como foi descrito na Seção 4.1. A amostra balanceada contém 47 variáveis *dummy*, equivalentes às variáveis originais, como também foi apresentado na Seção 4.2. A variável resposta “Cliente” é binária (*dummy*), assumindo valor 0 se o cliente for um “mau” pagador (casela de referência) e valor 1 se o cliente for um “bom” pagador.

Para interpretar os modelos de classificação, deve-se considerar que cada *dummy* (ou variável binária) corresponderá a um nível de uma variável preditora (categórica) e o último nível da variável preditora será a casela de referência. As *dummies* assumem valor 1, se o cliente possui a característica correspondente a esta variável, e 0, caso contrário. A casela de referência corresponde ao nível da variável preditora original associado a todos os valores zero das *dummies*.

As amostras de desenvolvimento e de validação foram obtidas por Cross-Validation com 10 partições. O desempenho dos modelos de classificação serão avaliados e comparados em relação à estatística Kolmogorov-Smirnov, coeficiente Gini e as taxas de acerto obtidas da Matriz de Confusão, descritas na Seção 3.4. O ponto de corte adotado para classificar os clientes da amostra de validação foi de 50. Logo, na amostra de validação, se o modelo ajustado previu *score* acima de 50, então o cliente foi classificado como um “bom” pagador, e caso contrário, o cliente foi classificado como “mau” pagador.

5.1 Regressão Logística

Neste trabalho, a análise de Regressão Logística foi realizada utilizando o software estatístico R. Inicialmente, foi estimado o modelo saturado (com todas as variáveis) e subseqüentemente foi estimado o modelo com variáveis selecionadas por *Forward Stepwise*. As medidas de desempenho destes modelos, obtidas das amostras de validação construídas por *Cross-Validation* com 10 partições (10-*fold*), são mostradas na Tabela 8.

As medidas de desempenho dos modelos logísticos ajustados indicam que a seleção de variáveis foi eficiente, pois tornou o modelo mais parcimonioso (com um número menor de variáveis) e as taxas de acerto, a estatística KS e o coeficiente Gini apresentaram maiores valores, após a seleção por *Forward Stepwise*. A seleção de variáveis por *Forward Stepwise* possibilita a redução do problema de dependência entre variáveis preditoras (multicolinearidade) pois, durante a seleção, são retiradas as variáveis que apresentem alta associação com outras.

Tabela 8 - Estatística KS, coeficiente Gini, taxas de acerto total (TAT), dos clientes "bons" (TAB), dos clientes "maus" (TAM) e número de variáveis dos modelos de Regressão Logística ajustados.

Modelo	TAT	TAB	TAM	KS	Gini	Nº de Variáveis
Regressão Logística (Modelo Saturado)	72.7%	70.7%	74.7%	45.33%	55.94%	48
Regressão Logística com Seleção por *Forward Stepwise*	74.83%	74.00%	75.67%	50.33%	61.46%	28

A amostra utilizada na estimação dos modelos logísticos foi a balanceada, obtida por Amostragem Retrospectiva na base de dados *German Credit*. Com isso, para a aplicação do modelo na população original é necessário que o intercepto do modelo logístico seja re-calculado, pela expressão 39 da Seção 4.1, que resulta em,

$$\hat{\beta}_o = 1.6039 - \ln\left(\frac{300/700}{1}\right) \cong 2.451 \qquad (44)$$

com, $\gamma_1 = 300/700$, $\gamma_2 = 1$ e $\hat{\beta}_0^* = 1.6039$ (intercepto do modelo logístico ajustado).

O intercepto do modelo logístico é utilizado para o cálculo da probabilidade do cliente ser um “bom” pagador, quando todas as variáveis *dummy* do modelo assumem valor zero. Dado o modelo logístico ajustado, esta probabilidade é dada por,

$$P(Cliente = 1) = \frac{\exp(2.451)}{1 + \exp(2.451)} = 0.9206 \quad (45)$$

As estimativas dos demais coeficientes do modelo com seleção de variáveis por *Forward Stepwise* (modelo logístico final), juntamente com algumas medidas descritivas de cada variável, com seus respectivos níveis (*dummies*), são apresentadas na Tabela 9. Estas medidas descritivas auxiliam na interpretação dos coeficientes.

As caselas de referência são os níveis associados aos valores zero das *dummies* e estão representadas na Tabela 9 por um traço (-). A variável resposta “Cliente” assume valor 0 se o cliente for um “mau” pagador e assume valor 1 se o cliente for um “bom” pagador.

Os coeficientes de um modelo de Regressão Logística múltipla podem ser interpretados como o aumento (ou redução) na probabilidade do cliente não se tornar inadimplente, em determinado período, dado o acréscimo de uma unidade na variável preditora (*dummy*), sem considerar a alteração das demais variáveis do modelo (MONTGOMERY et al., 2001).

Assim, coeficientes negativos indicam que a observação da característica do cliente (*dummy* de uma variável preditora) contribui para uma redução na probabilidade do cliente não se tornar inadimplente, em determinado período (aumento no risco de inadimplência), em relação aos demais clientes; e coeficientes positivos evidenciam que a observação da *dummy* produz um aumento na probabilidade do cliente não se tornar inadimplente, em determinado período (redução no risco de inadimplência), em relação aos demais clientes.

Tabela 9 - Estimativas dos coeficientes (Coef.) do modelo de Regressão Logística Final, juntamente as medidas descritivas: Risco Relativo, Número de Clientes "bons" (#Bons) e "maus" (#Maus), Total de clientes (Total) por categoria, Percentual de clientes da categoria em relação ao total de clientes (%Total) e Percentual de Maus por categoria (*dummy*) (*Bad Rate*).

Variável Preditora	Níveis (Variável *Dummy)*	Coef.	Risco Relativo	#Bons	#Maus	Total	%Total	*Bad Rate*
Intercepto		2.451	-	-	-	-	-	-
Bens	Imóvel	-	1.683	101	60	161	26.8%	37.3%
	Seguro de Vida, Carro e outros	-	0.931	161	173	334	55.7%	51.8%
	Não possui bens	-	0.567	38	67	105	17.5%	63.8%
Salário	X < $0	-1.988	0.378	51	135	186	31.0%	72.6%
	0 <= X < 200	-1.262	0.724	76	105	181	30.2%	58.0%
	X >=200	-0.943	1.500	21	14	35	5.8%	40.0%
	Sem remuneração	-	3.304	152	46	198	33.0%	23.2%
Poupança do Cliente	< $500	-0.871	0.753	189	251	440	73.3%	57.0%
	>= $500	-	2.059	35	17	52	8.7%	32.7%
	Não possui ou não conhecida	-	2.375	76	32	108	18.0%	29.6%
Outros Empréstimos	Nenhum	0.886	1.156	259	224	483	80.5%	46.4%
	Bancos e Lojas	-	0.539	41	76	117	19.5%	65.0%
Outras dívidas ou garantias	Nenhuma	-1.073	0.996	271	272	543	90.5%	50.1%
	Co-aplicante	-1.470	0.611	11	18	29	4.8%	62.1%
	Fiador	-	1.800	18	10	28	4.7%	35.7%
Finalidade	Compra de carro novo	-2.033	0.640	57	89	146	24.3%	61.0%
	Compra de carro usado	-	2.118	36	17	53	8.8%	32.1%
	Móveis	-1.387	0.897	52	58	110	18.3%	52.7%
	Rádio e TV	-1.279	1.548	96	62	158	26.3%	39.2%
	Utensílios domésticos	-1.685	0.500	2	4	6	1.0%	66.7%
	Reforma	-1.610	1.125	9	8	17	2.8%	47.1%
	Educação	-2.345	0.409	9	22	31	5.2%	71.0%
	Cursos	-	5.000	5	1	6	1.0%	16.7%
	Negócios	-1.488	0.853	29	34	63	10.5%	54.0%
	Outros	-	1.000	5	5	10	1.7%	50.0%
Histórico de Crédito	Sem empréstimos tomados ou todos os empréstimos pagos pontualmente	-1.278	0.226	12	53	65	10.8%	81.5%
	Existem empréstimos pagos pontualmente	-0.487	0.911	154	169	323	53.8%	52.3%
	Histórico de atraso no pagamento	-	1.000	28	28	56	9.3%	50.0%
	Atraso no pagamento ou com empréstimos tomados em outras Instituições	-	2.120	106	50	156	26.0%	32.1%
Tempo de Trabalho	Desempregado ou menos de 1 ano	-	0.720	67	93	160	26.7%	58.1%
	1 <= X < 4 anos	-	1.010	105	104	209	34.8%	49.8%
	X >= 4anos	-	1.243	128	103	231	38.5%	44.6%
Estado Civil e Sexo	Divorciado(a) ou separado(a) ou mulher casada	-	0.729	94	129	223	37.2%	57.8%
	Masculino solteiro	0.448	1.226	179	146	325	54.2%	44.9%
	Masculino casado ou viúvo.	-	1.080	27	25	52	8.7%	48.1%

(continua na próxima página)

Continuação da Tabela 9 - Estimativas dos coeficientes (Coef.) do modelo de Regressão Logística Final, juntamente as medidas descritivas: Risco Relativo, Número de Clientes "bons" (#Bons) e "maus" (#Maus) , Total de clientes (Total) por categoria e Percentual de clientes da categoria em relação ao total de clientes (%Total) e Percentual de Maus por categoria (*dummy*) (*Bad Rate*).

Variável Preditora	Níveis (Variável *Dummy)*	Coef.	Risco Relativo	#Bons	#Maus	Total	%Total	*Bad Rate*
Moradia	Própria	0.489	1.253	233	186	419	69.8%	44.4%
	Alugada ou Moradia gratuita	-	0.588	67	114	181	30.2%	63.0%
Emprego	Desempregado ou empregado com baixa qualificação	-	1.286	81	63	144	24.0%	43.8%
	Empregado qualificado ou funcionário público	-	0.946	176	186	362	60.3%	51.4%
	Executivo, profissional liberal, empregado altamente qualificado ou oficial	-	0.843	43	51	94	15.7%	54.3%
Telefone Próprio	Sim	-	1.027	116	113	229	38.2%	49.3%
	Não	-	0.984	184	187	371	61.8%	50.4%
Estrangeiro	Sim	1.858	0.946	280	296	576	96.0%	51.4%
	Não	-	5.000	20	4	24	4.0%	16.7%
Duração do empréstimo	<12	0.960	2.556	69	27	96	16.0%	28.1%
	12<= X<16	0.678	1.323	82	62	144	24.0%	43.1%
	16<= X<36	-	0.853	110	129	239	39.8%	54.0%
	>=36	-	0.476	39	82	121	20.2%	67.8%
Valor do Empréstimo	<1000	1.003	0.676	25	37	62	10.3%	59.7%
	1000<= X<4000	1.686	1.367	216	158	374	62.3%	42.2%
	4000<= X<7500	0.926	0.689	42	61	103	17.2%	59.2%
	>=7500	-	0.386	17	44	61	10.2%	72.1%
Taxa de juros em % do valor do empréstimo	>=4	-0.465	0.818	130	159	289	48.2%	55.0%
	<4	-	1.206	170	141	311	51.8%	45.3%
Tempo de Residência	<3	-	1.128	150	133	283	47.2%	47.0%
	3<= X<4	-	1.023	44	43	87	14.5%	49.4%
	>=4	-	0.855	106	124	230	38.3%	53.9%
Idade	<25	-	0.525	32	61	93	15.5%	65.6%
	25<= X<30	-	0.829	63	76	139	23.2%	54.7%
	30<= X<32	0.630	1.364	30	22	52	8.7%	42.3%
	32<= X<35	-	0.848	28	33	61	10.2%	54.1%
	35<= X<51	0.556	1.494	115	77	192	32.0%	40.1%
	>=51	-	1.032	32	31	63	10.5%	49.2%
Número de Dependentes	>=2	-	1.130	52	46	98	16.3%	46.9%
	<2	-	0.976	248	254	502	83.7%	50.6%
Número de créditos concedidos em seu banco	>=2	-	1.230	123	100	223	37.2%	44.8%
	< 2	-	0.885	177	200	377	62.8%	53.1%

A análise dos coeficientes do modelo logístico final mostra que as características dos clientes (*dummies*), que individualmente contribuem para redução no risco de inadimplência, em relação aos demais clientes, foram: "Outros Empréstimos (Nenhum)", "Salário (Sem Remuneração)", "Outras dívidas ou garantias (Fiador)", "Estado Civil e Sexo (Masculino solteiro)", "Moradia (Própria)", "Estrangeiro (Sim)",

“Duração do empréstimo (<=16 meses)”, “Valor do Empréstimo (<=$7500)” “Idade (30<= X<32 e 35<= X<51)”.

Por outro lado, as características que individualmente apresentam maiores riscos de inadimplência, em relação aos demais clientes, são: “Histórico de Crédito (sem empréstimos tomados ou com todos os empréstimos pagos pontualmente ou com empréstimos pagos pontualmente)”, “Valor do Empréstimo (>=7500)”, “Outros Empréstimos (bancos e lojas)”, “Duração do Empréstimo (>=16 meses)” e taxa de juros de mais de 4% do valor do empréstimo.

Algumas *dummies* com riscos relativos próximos a 1, consideradas com risco de inadimplência neutro, foram retiradas do modelo durante a seleção de variáveis. Estas *dummies* são: “Histórico de Crédito (Histórico de atraso no pagamento)”, “Tempo de Trabalho (1<=X<4 anos)”, “Emprego (Empregado qualificado ou funcionário público)”, “Telefone Próprio (Sim)”, “Tempo de Residência (3<= X<4)” e “Número de Dependentes (<2)”.

5.2 Classificadores Bayesianos

Nesta Seção são descritos os resultados das aplicações dos Classificadores Bayesianos *Naive Bayes*, TAN e GBN em modelos de *Credit Scoring*, utilizando a base de dados *German Credit* e *Cross-Validation* com 10 partições para a obtenção das amostras de desenvolvimento e de validação.

A seleção de variáveis nos Classificadores *Naive Bayes* e TAN, descrita na Seção 4.3, passou por duas etapas: uma de filtragem pelo ganho de informação e outra de seleção pelo método *Wrapper*. As seleções de variáveis, inferências, os aprendizados de parâmetros e de estrutura foram feitas utilizando o software WEKA.

As *dummies* da base de dados *German Credit* ordenadas por sua contribuição para o ganho de informação da variável resposta são mostradas na Tabela 10; e desta tabela pode-se observar que a *dummy* “Histórico de Crédito (Histórico de

atraso no pagamento)" é a que menos contribui para a previsão dos clientes bons e maus pagadores, seguida das *dummies* "Tempo de Trabalho (1<=X<4 anos)" e "Outras dívidas ou garantias (Nenhuma)", e assim por diante.

Da base de dados, foram retiradas 18 *dummies*, desde a variável "Histórico de Crédito (Histórico de atraso no pagamento)" até a variável "Outras dívidas ou garantias (Co-aplicante)". Os resultados dos Classificadores Bayesianos estimados após esta filtragem pelo ganho de informação. Após a filtragem, os Classificadores *Naive Bayes* e TAN passaram por uma seleção de variáveis pelo método *Wrapper.* Já os Classificadores GBN tiveram suas variáveis selecionadas pelo *Markov Blanket* da variável resposta. Os resultados destes modelos serão apresentados nas Seções 5.3.1 a 5.3.4.

Tabela 10 - Variáveis preditoras ordenadas pela sua contribuição individual para o ganho de informação em relação à variável resposta ("Cliente" bom ou mau pagador).

VariávelPreditora (Nível)	Ganho de Informação	VariávelPreditora (Nível)	Ganho de Informação
Salário (X < $0)	0.067934	Emprego (Desempregado ou empregado com baixa qualificação)	0.003567
Poupança do Cliente (< $500)	0.040192	Finalidade (Cursos)	0.003532
Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)	0.037357	Valor do Empréstimo (<1000)	0.003132
Valor do Empréstimo (1000<= X<4000)	0.028955	Duração do empréstimo (16<= X<36)	0.003021
Duração do empréstimo (<12)	0.027069	Tempo de Residência (<3)	0.002325
Moradia (Própria)	0.021194	Outras dívidas ou garantias (Co-aplicante)	0.002155
Bens (Imóvel)	0.017302	Idade (25<= X<30)	0.001905
Outros Empréstimos (Nenhum)	0.015834	Histórico de Crédito (Existem empréstimos pagos pontualmente)	0.001815
Estrangeiro (Sim)	0.014534	Salário (X >=200)	0.001799
Idade (35<= X<51)	0.013363	Idade (30<= X<32)	0.001626
Idade (<25)	0.013051	Bens (Seguro de Vida, Carro e outros)	0.001170
Finalidade (Rádio e TV)	0.012013	Emprego (Empregado qualificado ou funcionário público)	0.000837
Finalidade (Compra de carro novo)	0.011215	Finalidade (Utensílios domésticos)	0.000825
Estado Civil e Sexo (Divorciado(a) ou separado(a) ou mulher casada)	0.010544	Idade (32<= X<35)	0.000549
Finalidade (Compra de carro usado)	0.009168	Finalidade (Negócios)	0.000534
Estado Civil e Sexo (Masculino solteiro)	0.008808	Número de Dependentes (>=2)	0.000528
Poupança do Cliente (>= $500)	0.008359	Finalidade (Móveis)	0.000482
Salário (0 <= X < 200)	0.008025	Telefone Próprio (Sim)	0.000076
Finalidade (Educação)	0.007118	Finalidade (Reforma)	0.000073
Tempo de Trabalho (Desempregado ou menos de 1 ano)	0.006950	Outras dívidas ou garantias (Nenhuma)	0.000023
Taxa de juros em % do valor do empréstimo (<4)	0.006760	Tempo de Residência (3<= X<4)	0.000016
Valor do Empréstimo (4000<= X<7500)	0.005111	Tempo de Trabalho (1 <= X < 4 anos)	0.000009
Número de créditos concedidos em seu banco (>=2)	0.004545	Histórico de Crédito (Histórico de atraso no pagamento)	0.000000
Duração do empréstimo (12<= X<16)	0.004405	-	-

5.2.1 Classificador *Naive Bayes*

O desempenho dos Classificadores *Naive Bayes* em todas as etapas de seleção de variáveis, com relação às taxas de acerto da matriz de confusão, a estatística Kolmogorov-Smirnov e ao coeficiente Gini é mostrado na Tabela 11.

Nos Classificadores *Naive Bayes*, a filtragem de variáveis pelo ganho de informação foi eficiente, pois a retirada das 18 variáveis contribuiu para um modelo mais parcimonioso (com menor número de variáveis) e mais poderoso. A retirada das variáveis, que individualmente menos contribuíam para o ganho de informação da variável resposta, auxiliou no aumento das taxas de acerto nas previsões dos clientes "maus" pagadores e no aumento das medidas de discriminação entre clientes "bons" e "maus" pagadores (KS e Gini).

Tabela 11 - Estatística Kolmogorov-Smirnov, coeficiente Gini, taxas de acerto total (TAT), dos clientes "bons" (TAB) e dos clientes "maus" (TAM) e número de variáveis dos Classificadores *Naive Bayes* ajustados utilizando *Cross-Validation*.

Modelo	TAT	TAB	TAM	KS	Gini	Nº de Variáveis
Modelo Saturado	72.00%	68.30%	75.70%	44.33%	56.83%	48
Filtragem pelo Ganho de Informação	72.50%	68.70%	76.30%	46.67%	59.39%	31
Seleção *Wrapper* Forward	74.00%	73.00%	75.00%	48.33%	54.28%	12
Seleção *Wrapper* Backward	**74.50%**	**72.30%**	**76.70%**	**49.33%**	**57.47%**	**28**

Após a filtragem, foi realizada a seleção de variáveis pelo método *Wrapper* e utilizando a busca *Forward Selection* e *Backward Elimination*, para efeito de comparação. Da Tabela 11, pode-se observar que as seleções pelo método *Wrapper* apresentaram melhores desempenhos, do que os modelos saturado e com filtragem de variáveis, em relação à assertividade das previsões dos clientes "bons" e a estatística KS.

Apesar do modelo *Naive Bayes* com busca por *Forward Selection* ser o mais parcimonioso, será escolhido o Classificador *Naive Bayes* com seleção pelo método *Wrapper* com busca por *Backward Elimination* como o melhor modelo (Classificador *Naive Bayes* final); e está indicado em negrito na Tabela 11. Pois, este é o modelo

que discrimina um pouco melhor os clientes bons e maus pagadores, analisando a estatística KS e o coeficiente Gini, e também apresentou maior assertividade nas previsões dos clientes “maus” pagadores.

Para a aplicação do Classificador *Naive Bayes* final na população original, foi necessária a substituição da distribuição da variável resposta “Cliente” pela sua distribuição na amostra original, que é $P(Cliente = 1) = 0.7$ e $P(Cliente = 0) = 0.3$. Esta substituição foi feita, pois a estimação do modelo foi feita a partir da amostra balanceada, obtida por amostragem retrospectiva.

Apenas para ilustrar, é apresentada na Figura 12 a estrutura da RB do Classificador *Naive Bayes* final.

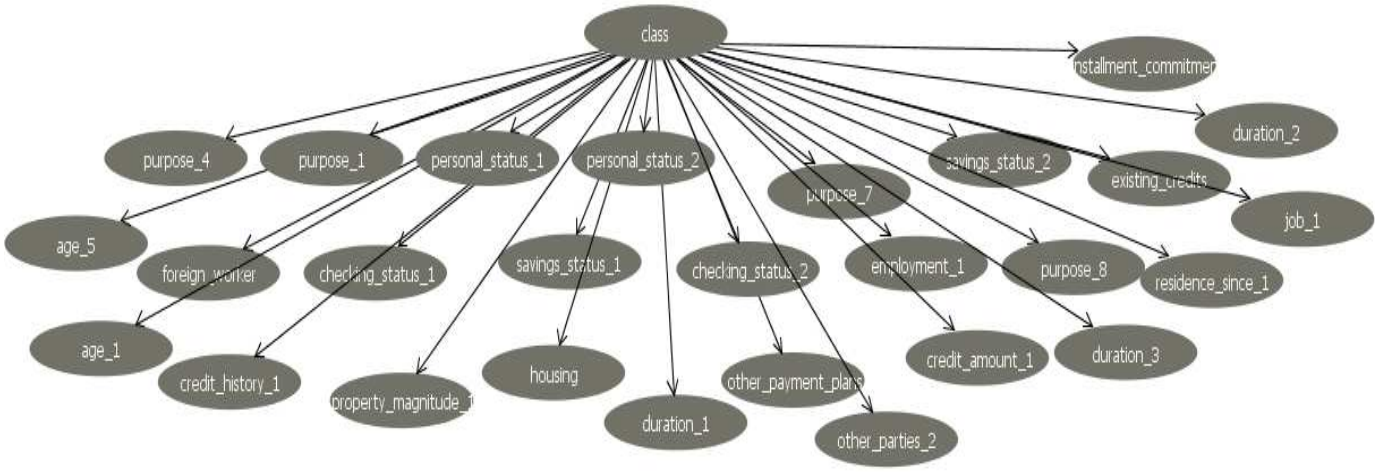

Figura 12 - Estrutura da Rede Bayesiana do Classificador *Naive Bayes* com seleção de variáveis pelo método *Wrapper* com busca *Backward Elimination*.

A Tabela 12 contém as probabilidades condicionais de um cliente pretendente ao crédito não se tornar inadimplente, em determinado período, dada à observação de uma variável preditora (*dummy*), sem que haja alteração nos valores das demais variáveis. As probabilidades condicionais desta tabela informam o aumento na probabilidade do cliente ser “bom” pagador, dada a observação de uma característica do cliente, sem que outras variáveis do modelo recebam evidências; e foram obtidas por Inferências utilizando o algoritmo *Junction Tree*.

Tabela 12 - Probabilidades dos clientes se tornarem “bons” e “maus” pagadores, dado a observação dos níveis das variáveis preditoras, sem que haja alteração nas categorias das demais variáveis, do Classificador *Naive Bayes* final.

VariávelPreditora	Nível(variável *dummy*)	P(Cliente=1| Categoria=1)	P(Cliente=0| Categoria=1)	P(Cliente=1| Categoria=0)	P(Cliente=0| Categoria=0)
Bens	Imóvel	0.796	0.204	0.659	0.341
Salário	X < $0	0.472	0.528	0.778	0.222
	0 <= X < 200	0.629	0.371	0.728	0.272
Poupança do Cliente	< $500	0.638	0.362	0.839	0.161
	>= $500	0.824	0.176	0.686	0.314
Outros Empréstimos	Nenhum	0.729	0.271	0.560	0.440
Outras dívidas ou garantias	Co-aplicante	0.596	0.404	0.705	0.295
Finalidade	Compra de carro novo	0.601	0.399	0.729	0.271
	Rádio e TV	0.782	0.218	0.667	0.333
	Educação	0.504	0.496	0.709	0.291
	Cursos	0.875	0.125	0.697	0.303
Histórico de Crédito	Sem empréstimos tomados ou todos os empréstimos pagos pontualmente	0.360	0.640	0.731	0.269
Tempo de Trabalho	Desempregado ou menos de 1 ano	0.628	0.372	0.724	0.276
Estado Civil e Sexo	Divorciado(a) ou separado(a) ou mulher casada	0.630	0.370	0.737	0.263
	Masculino solteiro	0.741	0.259	0.647	0.353
Moradia	Própria	0.745	0.255	0.580	0.420
Emprego	Desempregado ou empregado com baixa qualificação	0.749	0.251	0.683	0.317
Estrangeiro	Não	0.907	0.093	0.688	0.312
Duração do empréstimo	<12	0.854	0.146	0.664	0.336
	12<= X<16	0.755	0.245	0.681	0.319
	16<= X<36	0.666	0.334	0.722	0.278
Valor do Empréstimo	<1000	0.615	0.385	0.709	0.291
Taxa de juros em % do valor do empréstimo	>=4	0.656	0.344	0.738	0.262
Tempo de Residência	<3	0.724	0.276	0.677	0.323
Idade	<25	0.554	0.446	0.723	0.277
	35<= X<51	0.776	0.224	0.660	0.340
Número de créditos concedidos em seu banco	>=2	0.741	0.259	0.674	0.326

Analisando Tabela 12 verifica-se que as *dummies* que contribuem para o aumento no risco de inadimplência, quando analisadas isoladamente em relação aos demais clientes, são: “Salário (X < $0 ou 0 <= X < 200)”, “Poupança do Cliente (<$500)”, “Outros Empréstimos (Bancos e Lojas)”, “Outras dívidas ou garantias (Co-aplicante)”, “Finalidade (Compra de Carro Novo ou Educação)”, “Histórico de Crédito

(Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)", "Tempo de Trabalho (Desempregado ou menos de 1 ano)", "Estado Civil e Sexo (Divorciado(a) ou separado(a) ou mulher casada)", "Duração do Empréstimo (16<= X<36)", "Moradia (Alugada ou Moradia gratuita)", "Valor do Empréstimo (<1000)", "Taxa de juros em % do valor do empréstimo (>=4)", "Idade (<25)" e "Número de créditos concedidos em seu banco(<2)".

5.2.2 Classificador *TAN*

Os Classificadores bayesianos *Tree Augmented Naive Bayes* (TAN) foram construídos aplicando a filtragem pelo ganho de informação, seguida da seleção pelo método *Wrapper* com busca por *Backward Elimination*. A Tabela 13 apresenta as medidas do desempenho do Classificador TAN nas etapas de seleção de variáveis.

Os aprendizados de estrutura realizados empregaram as cinco medidas, descritas na Seção 3.1.2: MDL, Bayes, AIC, Bdeu e Entropia. No Classificador TAN, busca da melhor estrutura é restrita ao espaço de estruturas do tipo árvore. Com isso, cada variável preditora (*dummy*) deve ter no máximo um pai, além da variável resposta "Cliente", ou seja, o pai de uma variável será outra *dummy* com maior dependência condicional, dada a variável resposta.

O software Weka foi utilizado nas seleções de variáveis, aprendizados e inferências. A busca por *Backward Elimination* foi adotada no método *Wrapper*, pois era a única disponível no software Weka.

A análise da Tabela 13 indica que a filtragem de variáveis pelo ganho de informação melhorou o desempenho dos Classificadores TAN, em relação a todas as taxas de acerto e em relação às estatísticas KS e coeficiente Gini. No entanto, a seleção de variáveis pelo método *Wrapper* piorou o desempenho dos Classificadores TAN com aprendizado de estrutura utilizando as medidas MDL e AIC, em relação às taxas de acerto dos clientes "bons" pagadores.

O Classificador TAN com aprendizado de estrutura utilizando a medida *Bayes* foi selecionado como melhor Classificador (Classificador TAN final), pois foi o modelo que apresentou maiores taxas de acerto nas previsões dos "maus" e altos valores da estatística KS e do coeficiente Gini. O Classificador TAN final está indicado em negrito na Tabela 13.

Tabela 13 - Estatística Kolmogorov-Smirnov, coeficiente Gini, taxas de acerto total (TAT), dos clientes "bons" (TAB) e dos clientes "maus" (TAM) e número de variáveis dos Classificadores TAN com aprendizado de estrutura utilizando diferentes medidas.

Medida	Modelo	TAT	TAB	TAM	KS	Gini	Nº de Variáveis
MDL	Modelo Saturado	72.83%	70.00%	75.70%	47.00%	58.74%	48
	Filtragem pelo Ganho de Informação	74.67%	71.30%	78.00%	50.67%	62.05%	31
	Seleção *Wrapper*	73.17%	68.70%	77.70%	50.67%	61.61%	27
Bayes	Modelo Saturado	73.67%	71.70%	75.70%	47.67%	58.66%	48
	Filtragem pelo Ganho de Informação	75.17%	72.30%	78.00%	50.67%	61.29%	31
	Seleção *Wrapper*	**75.33%**	**71.00%**	**79.70%**	**51.33%**	**60.81%**	**27**
AIC	Modelo Saturado	72.83%	70.00%	75.70%	47.00%	58.73%	48
	Filtragem pelo Ganho de Informação	74.67%	71.30%	78.00%	50.67%	62.05%	31
	Seleção *Wrapper*	73.17%	68.70%	77.70%	50.67%	61.59%	27
Bdeu	Modelo Saturado	73.17%	72.00%	74.30%	46.67%	58.17%	48
	Filtragem pelo Ganho de Informação	74.50%	72.00%	77.00%	50.00%	60.85%	31
	Seleção *Wrapper*	74.00%	72.30%	75.70%	50.00%	61.77%	26
Entropia	Modelo Saturado	72.83%	70.00%	75.70%	47.00%	58.73%	48
	Filtragem pelo Ganho de Informação	74.67%	71.30%	78.00%	50.67%	62.05%	31
	Seleção *Wrapper*	73.17%	69.70%	77.70%	50.67%	61.59%	27

A estrutura da RB do Classificador TAN final (Figura 13) mostra claramente a premissa do Classificador de que cada variável preditora pode ter no máximo um pai na RB, ou seja, a estrutura de um grafo de árvore. As variáveis (*dummies*) do Classificador TAN final, e seus respectivos pais, são apresentadas na Tabela 14.

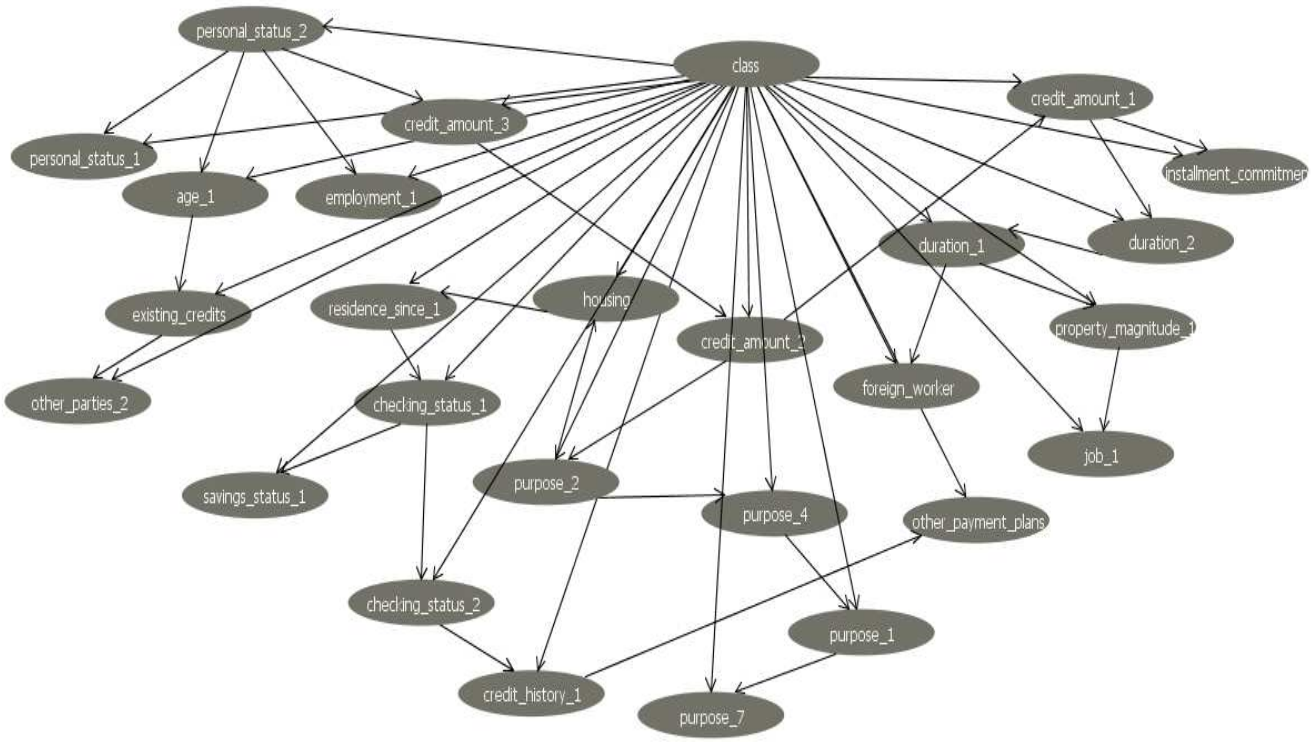

Figura 13 - Estrutura da RB do Classificador TAN com aprendizado de estrutura utilizando a medida *Bayes* e seleção de variáveis pelo método *Wrapper* com busca por *Backward Elimination*.

Tabela 14 - Variáveis preditoras e respectivos pais do Classificador TAN final.

Variável Preditora (Nível)	**Pais (Nível)**	**Label da Variável Preditora**	**Label de seus Pais**
Bens (Imóvel)	Duração do empréstimo (X <12), Cliente	property_magnitude_1	duration_1, class
Salário (X < $0)	Tempo de Residência (X <3), Cliente	checking_status_1	residence_since_1, class
Salário (0 <= X < 200)	Salário (X < $0) , Cliente	checking_status_2	checking_status_1, class
Poupança do Cliente (X < $500)	Salário (X < $0) , Cliente	savings_status_1	checking_status_1, class
Outros Empréstimos (Nenhum)	Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente) , Cliente	other_payment_plans	credit_history_1, class
Outras dívidas ou garantias (Co-aplicante)	Número de créditos concedidos em seu banco (X >=2) , Cliente	other_parties_2	existing_credits, class
Finalidade (Compra de carro novo)	Finalidade (Rádio e TV) , Cliente	purpose_1	purpose_4, class
Finalidade (Compra de carro usado)	Valor do Empréstimo (1000<= X <4000) , Cliente	purpose_2	credit_amount_2, class
Finalidade (Rádio e TV)	Finalidade (Compra de carro usado) , Cliente	purpose_4	purpose_2, class

(continua na próxima página)

Continuação da Tabela 14 - Variáveis preditoras e respectivos pais do Classificador TAN final.

Variável Preditora (Nível)	Pais (Nível)	Label da Variável Preditora	Label de seus Pais
Finalidade (Educação)	Finalidade (Compra de carro novo) , Cliente	purpose_7	purpose_1, class
Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)	Salário (0 <= X < 200) , Cliente	credit_history_1	checking_status_2, class
Tempo de Trabalho (Desempregado ou menos de 1 ano)	Estado Civil e Sexo (Masculino solteiro) , Cliente	employment_1	personal_status_2, class
Estado Civil e Sexo (Divorciado(a) ou separado(a) ou mulher casada)	Estado Civil e Sexo (Masculino solteiro) , Cliente	personal_status_1	personal_status_2, class
Estado Civil e Sexo (Masculino solteiro)	Cliente-	personal_status_2	Class
Moradia (Própria)	Finalidade (Compra de carro usado) , Cliente	housing	purpose_2, class
Emprego (Desempregado ou empregado com baixa qualificação)	Bens (Imóvel) , Cliente	job_1	property_magnitude_1, class
Estrangeiro (Não)	Duração do empréstimo (X <12) , Cliente	foreign_worker	duration_1, class
Duração do empréstimo (X <12)	Duração do empréstimo (12<= X <16) , Cliente	duration_1	duration_2, class
Duração do empréstimo (12<= X <16)	Valor do Empréstimo (X <1000) , Cliente	duration_2	credit_amount_1, class
Valor do Empréstimo (X <1000)	Valor do Empréstimo (1000<= X <4000) , Cliente	credit_amount_1	credit_amount_2, class
Valor do Empréstimo (1000<= X <4000)	Valor do Empréstimo (4000<= X <7500) , Cliente	credit_amount_2	credit_amount_3, class
Valor do Empréstimo (4000<= X <7500)	Estado Civil e Sexo (Masculino solteiro) , Cliente	credit_amount_3	personal_status_2, class
Taxa de juros em % do valor do empréstimo (X <4)	Valor do Empréstimo (X <1000) , Cliente	installment_commitment	credit_amount_1, class
Tempo de Residência (X <3)	Moradia (Própria) , Cliente	residence_since_1	housing, class
Idade (X <25)	Estado Civil e Sexo (Masculino solteiro) , Cliente	age_1	personal_status_2, class
Número de créditos concedidos em seu banco (X >=2)	Idade (X <25) , Cliente	existing_credits	age_1, class

As dependências entre variáveis preditoras obtidas do aprendizado de estrutura serão analisadas a partir do cálculo das probabilidades dos clientes se tornarem “bons” (ou “maus”) pagadores, em determinado período, dada a observação das variáveis (filhos) e de seus pais, sem que as demais variáveis recebam evidências na RB. Estas probabilidades são mostradas na Tabela 17.

Para efeito de comparação, também foram calculadas as probabilidades dos clientes se tornarem inadimplentes (ou não), em determinado período, dada a

observação das variáveis preditoras isoladamente, sem considerar o efeito das demais, mostradas na Tabela 15.

Tabela 15 - Probabilidades dos clientes se tornarem inadimplentes (ou não), dada a observação de cada variável preditora isoladamente, no Classificador TAN com aprendizado utilizando a medida Bayes e com seleção com o método *Wrapper* com busca por Backward Elimination.

Variável (Nível**)**	**P(Cliente=1| Categoria=1)**	**P(Cliente=0| Categoria=1)**	**P(Cliente=1| Categoria=0)**	**P(Cliente=0| Categoria=0)**
Bens (Imóvel)	0.639	0.361	0.713	0.287
Salário (X < $0)	0.548	0.452	0.763	0.237
Salário (0 <= X < 200)	0.023	0.977	0.820	0.180
Poupança do Cliente (X < $500)	0.736	0.264	0.546	0.454
Outros Empréstimos (Nenhum)	0.639	0.361	0.814	0.186
Outras dívidas ou garantias (Co-aplicante)	0.745	0.255	0.697	0.303
Finalidade (Compra de carro novo)	0.050	0.950	0.781	0.219
Finalidade (Compra de carro usado)	0.123	0.877	0.735	0.265
Finalidade (Rádio e TV)	0.209	0.791	0.750	0.250
Finalidade (Educação)	0.136	0.864	0.717	0.283
Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)	0.833	0.167	0.673	0.327
Tempo de Trabalho (Desempregado ou menos de 1 ano)	0.524	0.476	0.754	0.246
Estado Civil e Sexo (Divorciado(a) ou separado(a) ou mulher casada)	0.009	0.991	0.923	0.077
Estado Civil e Sexo (Masculino solteiro)	0.741	0.259	0.647	0.353
Moradia (Própria)	0.407	0.593	0.851	0.149
Emprego (Desempregado ou empregado com baixa qualificação)	0.831	0.169	0.640	0.360
Estrangeiro (Não)	0.890	0.110	0.691	0.309
Duração do empréstimo (X <12)	0.094	0.906	0.732	0.268
Duração do empréstimo (12<= X <16)	0.895	0.105	0.502	0.498
Valor do Empréstimo (X <1000)	0.025	0.975	0.766	0.234
Valor do Empréstimo (1000<= X <4000)	0.029	0.971	0.879	0.121
Valor do Empréstimo (4000<= X <7500)	0.811	0.189	0.675	0.325
Taxa de juros em % do valor do empréstimo (X >=4)	0.805	0.195	0.431	0.569
Tempo de Residência (X <3)	0.833	0.167	0.577	0.423
Idade (X <25)	0.463	0.537	0.735	0.265
Número de créditos concedidos em seu banco (X >=2)	0.566	0.434	0.748	0.252

Tabela 16 - Probabilidades dos clientes se tornarem inadimplentes (ou não), dada a observação de cada variável preditora isoladamente, no Classificador TAN Classificador TAN com aprendizado utilizando a medida Bayes e com seleção com o método *Wrapper* com busca por Backward Elimination.

Filhos (Nível**)**	**Pais (**Nível**)**	**P(Cliente=1| Filho=1,Pai=1)**	**P(Cliente=0| Filho=1,Pai=1)**	**P(Cliente=1| Filho=1,Pai=0)**	**P(Cliente=0| Filho=1,Pai=0)**
Bens (Imóvel)	Duração do empréstimo (X <12)	0.180	0.820	0.680	0.320
Salário (X < $0)	Tempo de Residência (X <3)	0.764	0.236	0.479	0.521
Outros Empréstimos (Nenhum)	Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)	0.768	0.232	0.614	0.386
Outras dívidas ou garantias (Co-aplicante)	Número de créditos concedidos em seu banco (X >=2)	0.211	0.789	0.578	0.422
Finalidade (Compra de carro usado)	Valor do Empréstimo (1000<= X <4000)	0.015	0.985	0.372	0.628
Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)	Salário (0 <= X < 200)	0.117	0.883	0.848	0.152
Tempo de Trabalho (Desempregado ou menos de 1 ano)	Estado Civil e Sexo (Masculino solteiro)	0.583	0.417	0.469	0.531
Moradia (Própria)	Finalidade (Compra de carro usado)	0.088	0.912	0.455	0.545
Emprego (Desempregado ou empregado com baixa qualificação)	Bens (Imóvel)	0.827	0.173	0.833	0.167
Estrangeiro (Não)	Duração do empréstimo (X <12)	0.382	0.618	0.920	0.080
Duração do empréstimo (12<= X <16)	Valor do Empréstimo (X <1000)	0.050	0.950	0.935	0.065
Valor do Empréstimo (4000<= X <7500)	Estado Civil e Sexo (Masculino solteiro)	0.916	0.084	0.708	0.292
Taxa de juros em % do valor do empréstimo (X >=4)	Valor do Empréstimo (X <1000)	0.030	0.970	0.848	0.152
Tempo de Residência (X <3)	Moradia (Própria)	0.574	0.426	0.935	0.065
Idade (X <25)	Estado Civil e Sexo (Masculino solteiro)	0.509	0.491	0.432	0.568
Número de créditos concedidos em seu banco (X >=2)	Idade (X <25)	0.246	0.754	0.624	0.376

Analisando as estimativas das probabilidades condicionais das Tabela 15 e Tabela 17, nota-se que a observação da variável “Histórico de Crédito (Sem empréstimos tomados ou com todos os empréstimos pagos pontualmente)” isoladamente contribuiu para uma redução no risco de inadimplência, se comparada aos demais clientes. No entanto, quando é analisado o efeito de interação entre esta variável com seu pai na RB, “Salário (0<=X<200)”, observa-se outra interpretação. A

interação mostra que, entre os clientes sem empréstimos tomados ou com todos os empréstimos pagos pontualmente (em dia), aqueles com salário inferior a $200 possuem alto risco de inadimplência, se comparados demais clientes com outras faixas salariais e com mesmo histórico de crédito. Estes clientes com outras faixas salariais e com mesmo histórico de crédito, por sua vez, possuem baixo risco de inadimplência, em determinado período. Com isso, fica clara a importância de se avaliar o efeito combinado entre variáveis e de seus pais na RB.

As dependências entre variáveis indicadas no aprendizado de estrutura se confirmam na amostra, pois seus riscos relativos (Tabela A. 10 do Apêndice) apresentaram valores superiores a 1, indicando alta associação entre as *dummies* preditoras e seus pais.

A análise das interações entre variáveis na RB permite a avaliação dos efeitos da combinação de cada *dummy* preditora com seu pai na RB. Esta análise é importante, pois pode levar a informações a respeito do perfil de risco de inadimplência que não podem ser observadas analisando as variáveis individualmente, como pode ser observado anteriormente.

Analisando a Tabela 17, observa-se que as seguintes interações contribuem para maiores riscos de inadimplência, enquanto individualmente a observação da *dummy* preditora contribui para menores riscos de inadimplência, em relação aos outros clientes: variável "Outras dívidas ou garantias(Co-aplicante)" com pai "Número de créditos concedidos em seu banco(X>=2)", variável "Histórico de Crédito(Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)" com pai "Salário(0<=X<200)", variável "Estrangeiro(Sim)" com pai "Duração do empréstimo(X<12)", variável "Duração do empréstimo(12<=X<16)" com pai "Valor do Empréstimo(X<1000)", variável "Taxa de juros em % do valor do empréstimo(X>=4)" com pai "Valor do Empréstimo(X<1000)", variável "Tempo de Residência(X<3)" com pai "Moradia(Própria)".

5.2.3 Classificador *GBN*

No aprendizado aplicado para a obtenção da estrutura do Classificador GBN a partir da amostra balanceada da base de dados *German Credit*, foi empregado o algoritmo *Hill Climbing* para a busca da estrutura que melhor descreva a base de dados. As medidas utilizadas para avaliar as possíveis estruturas durante sua busca foram: MDL, Bayes, AIC, Bdeu e Entropia

O algoritmo *Hill Climbing* busca a melhor estrutura da RB, adicionando e removendo arestas, além de ajustar arestas reversas (verifica o sentido das arestas). A melhor estrutura é aquela que maximiza (ou minimiza) as medidas acima. A busca não é restrita a ordem das variáveis, como ocorre no algoritmo K2.

Os Classificadores GBN, assim os *Naive Bayes* e TAN, também tiveram suas variáveis filtradas pelo seu ganho de informação à variável resposta “Cliente”. No entanto, a Tabela A. 1 mostra que esta filtragem prejudicou o desempenho dos aprendizados de estrutura com as medidas: AIC, Bdeu e Entropia. Por isso, os Classificadores GBN tiveram suas variáveis selecionadas somente pelo Markov da variável resposta, sem passarem pela etapa de filtragem.

O fraco desempenho da filtragem pelo ganho de informação evidencia que, as variáveis removidas individualmente podem contribuir menos para o ganho de informação da variável resposta, mas combinadas com outras variáveis, elas podem ser importantes para a classificação dos clientes “bons” e “maus” pagadores.

Analisando a Tabela 17, nota-se que a seleção de variáveis pelo *Markov Blanket* da variável resposta melhorou o desempenho dos Classificadores GBN nos aprendizados de estrutura empregando as medidas MDL, Bayes e Bdeu em relação à estatística KS, coeficiente Gini e taxas de acerto.

Comparando os resultados destas três medidas observam-se resultados distintos entre elas, sendo que a medida Bdeu apresentou maiores valores para as taxas de acerto (TAT, TAB e TAM), estatística KS, coeficiente Gini e também apresentou Classificador GBN mais parcimonioso (com 15 variáveis).

Tabela 17 - Estatística Kolmogorov-Smirnov, coeficiente Gini, taxas de acerto total (TAT), dos clientes "bons" (TAB), dos clientes "maus" (TAM) e número de variáveis dos Classificadores GBN com aprendizado de estrutura utilizando diferentes medidas.

Medida	Modelo	TAT	TAB	TAM	KS	Gini	Nº de Variáveis
MDL	Modelo Saturado	70.67%	69.30%	72.00%	42.00%	53.69%	48
	Seleção *Markov Blanket*	73.00%	72.70%	73.30%	46.00%	59.40%	16
Bayes	Modelo Saturado	68.50%	65.30%	71.70%	40.00%	51.47%	48
	Seleção *Markov Blanket*	71.67%	67.70%	75.70%	43.67%	57.03%	30
AIC	Modelo Saturado	72.50%	70.30%	74.70%	45.33%	55.70%	48
	Seleção *Markov Blanket*	70.33%	68.70%	72.00%	43.00%	54.54%	37
Bdeu	Modelo Saturado	71.50%	71.30%	71.70%	44.67%	55.84%	48
	Seleção *Markov Blanket*	**74.67%**	**76.30%**	**76.30%**	**49.67%**	**58.84%**	**15**
Entropia	Seleção *Markov Blanket*	65.83%	67.70%	64.00%	33.00%	43.23%	48

O Classificador GBN com seleção de variáveis pelo *Markov Blanket* e aprendizado de estrutura utilizando a medida Bdeu será chamado de Classificador GBN final (em negrito na Tabela 17). Sua estrutura é apresentada na Figura 14 e a Tabela 18 mostra suas variáveis e respectivos pais na RB. Observando a estrutura do Classificador GBN, nota-se claramente a premissa do Classificador GBN de que a variável resposta não necessariamente é pai de todas as variáveis preditoras.

O desempenho dos Classificadores GBN com seleções de variáveis e aprendizados de estrutura com as medidas AIC e Entropia apresentaram piores resultados, em relação a todas as medidas de desempenho dos modelos.

Esta baixa performance verificada nos aprendizados de estrutura com as medidas Entropia e AIC indicam que a escolha do *Markov Blanket* da variável resposta não foi adequada. Pois, a retirada das variáveis de fora do *Markov Blanket* da variável resposta prejudicou o desempenho dos modelos, indicando que a resposta não depende somente das variáveis do seu *Markov Blanket*.

Este fraco desempenho ocorre devido ao número de variáveis do modelo e ao número de observações da base de dados *German Credit*. Resultado semelhante a estes também foi obtido por Friedman e Goldszmidt (1996), que observaram um fraco desempenho nos Classificadores GBN em bases de dados com mais de 15 atributos. Os autores concluíram que esta baixa assertividade e baixo poder

discriminante observados em alguns aprendizados dos Classificadores GBN podem ocorrer devido ao grande número de dependências avaliadas em conjuntos de dados com mais do que 15 atributos.

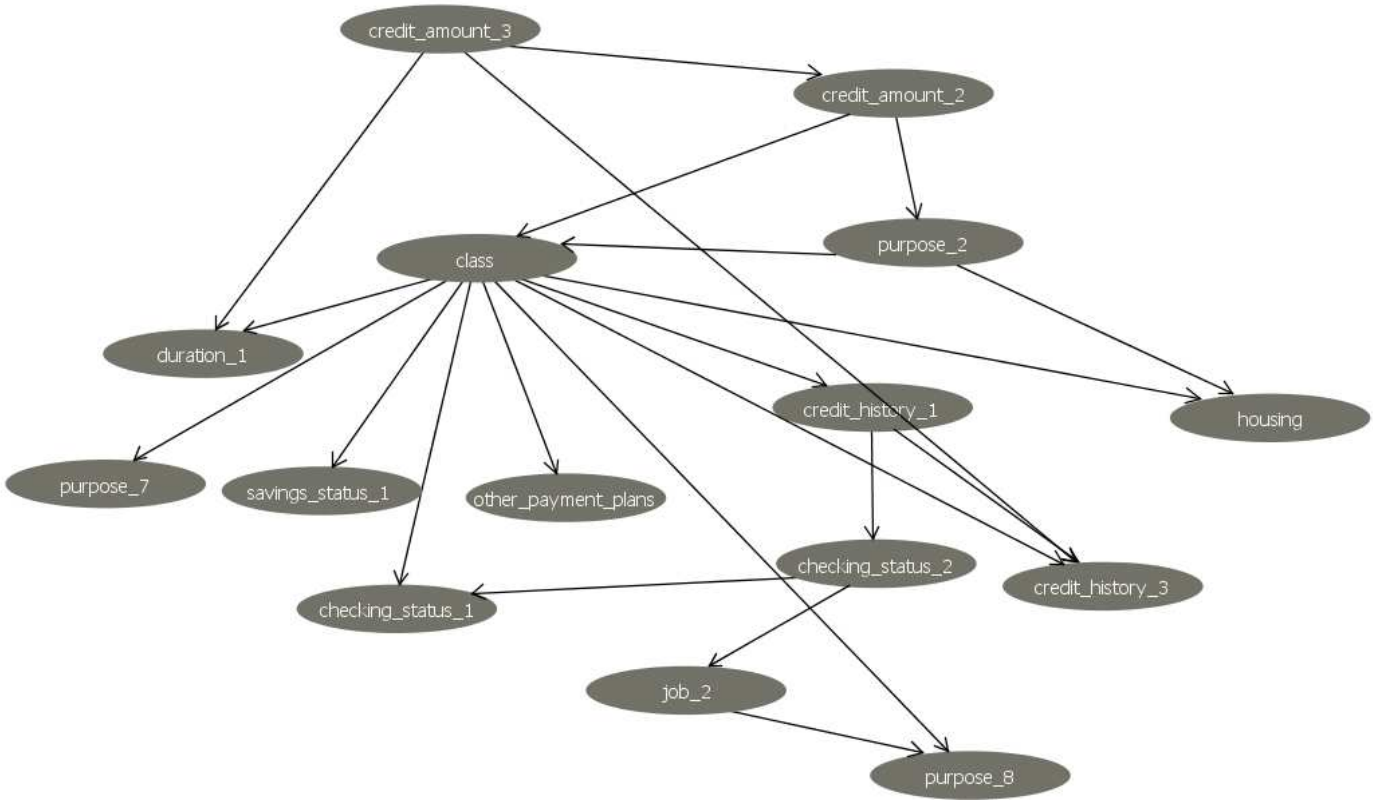

Figura 14 - Estrutura da RB do Classificador GBN com aprendizado de estrutura utilizando a medida BDeu e com seleção de variáveis pelo *Markov Blanket* da variável resposta.

A interpretação do Classificador GBN final será feita a partir da análise das probabilidades condicionais dos clientes se tornarem inadimplentes (ou não), em determinado período, dada a observação de cada *dummy* preditora individualmente, mostradas na Tabela 19.

A análise das dependências entre variáveis preditoras será realizada pela análise dos efeitos de interação entre *dummies*. Para isso, foram calculadas as probabilidades dos clientes se tornarem “bons” (ou “maus”) pagadores, dada as observações das *dummies* preditoras e dada a observação (ou não) de seus pais na RB, nas Tabela 20 e Tabela 21.

Tabela 18 - Variáveis preditoras (*dummies*) e respectivos pais do Classificador GBN com aprendizado de estrutura utilizando a medida BDeu e com seleção de variáveis pelo *Markov Blanket* da variável resposta.

Filho (Nível)	Pais (Nível)	Label Filhos	Label Pais
Salário (X < $0)	Cliente, Salário (0 <= X < 200)	checking_status_1	class, checking_status_2
Salário (0 <= X < 200)	Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)	checking_status_2	credit_history_1
Poupança do Cliente (< $500)	Cliente	savings_status_1	class
Outros Empréstimos (Nenhum)	Cliente	other_payment_plans	class
Finalidade (Compra de carro usado)	Valor do Empréstimo (1000<= X<4000)	purpose_2	credit_amount_2
Finalidade (Educação)	Cliente	purpose_7	class
Finalidade (Cursos)	Cliente, Emprego (Empregado qualificado ou funcionário público)	purpose_8	class, job_2
Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)	Cliente	credit_history_1	class
Histórico de Crédito (Histórico de atraso no pagamento)	Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente), Valor do Empréstimo(4000<= X<7500), Cliente	credit_history_3	credit_history_1, credit_amount_3, class
Moradia (Própria)	Cliente, Finalidade (Compra de carro usado)	housing	class, purpose_2
Emprego (Empregado qualificado ou funcionário público)	Salário (0 <= X < 200)	job_2	checking_status_2
Duração do empréstimo (<12)	Cliente, Valor do Empréstimo (4000<= X<7500)	duration_1	class, credit_amount_3
Valor do Empréstimo (1000<= X<4000)	Valor do Empréstimo (4000<= X<7500)	credit_amount_2	credit_amount_3
Valor do Empréstimo (4000<= X<7500)	-	credit_amount_3	-
Cliente	Valor do Empréstimo (1000<= X<4000), Finalidade (Compra de carro usado)	class	credit_amount_2, purpose_2

Tabela 19 - Probabilidades dos clientes se tornarem inadimplentes (ou não), dada a observação de cada variável preditora isoladamente, do Classificador GBN com aprendizado de estrutura utilizando a medida BDeu e com seleção de variáveis pelo *Markov Blanket* da variável resposta.

Variável Preditora (Nível)	P(Cliente=1| Categoria=1)	P(Cliente=0| Categoria=1)	P(Cliente=1| Categoria=0)	P(Cliente=0| Categoria=0)
Salário (X < $0)	0.4459	0.5541	0.7884	0.2116
Salário (0 <= X < 200)	0.6845	0.3155	0.7067	0.2933
Poupança do Cliente (< $500)	0.6375	0.3625	0.8402	0.1598
Outros Empréstimos (Nenhum)	0.7296	0.2704	0.5588	0.4412
Finalidade (Compra de carro usado)	0.1649	0.8351	0.6846	0.3154
Finalidade (Educação)	0.4964	0.5036	0.7096	0.2904
Finalidade (Cursos)	0.8719	0.1281	0.6974	0.3026
Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)	0.3529	0.6471	0.7313	0.2687
Histórico de Crédito (Histórico de atraso no pagamento)	0.6964	0.3036	0.7005	0.2995
Moradia (Própria)	0.7446	0.2554	0.5808	0.4192
Emprego (Empregado qualificado ou funcionário público)	0.7012	0.2988	0.6984	0.3016
Duração do empréstimo (<12)	0.8546	0.1454	0.6637	0.3363
Valor do Empréstimo (1000<= X<4000)	0.7524	0.2476	0.6078	0.3922
Valor do Empréstimo (4000<= X<7500)	0.6078	0.3922	0.7177	0.2823

Tabela 20 - Probabilidades dos clientes se tornarem inadimplentes (ou não), dada a observação de cada variável preditora isoladamente e dada a observação (ou não) de seus pais, do Classificador GBN com aprendizado de estrutura utilizando a medida Bdeu e com seleção de variáveis pelo *Markov Blanket* da variável resposta.

Variável Preditora (Nível)	Pai (Nível)	P(Cliente=1| Filho=1,Pai=1)	P(Cliente=0| Filho=1,Pai=1)	P(Cliente=1| Filho=1,Pai=0)	P(Cliente=0| Filho=1,Pai=0)
Salário (0 <= X < 200)	Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)	0.3529	0.6471	0.7313	0.2687
Finalidade (Compra de carro usado)	Valor do Empréstimo (1000<= X<4000)	0.0232	0.9768	0.2666	0.7334
Finalidade (Cursos)	Emprego (Empregado qualificado ou funcionário público)	0.4525	0.5475	0.9591	0.0409
Moradia (Própria)	Finalidade (Compra de carro usado)	0.0648	0.9352	0.7302	0.2698
Emprego (Empregado qualificado ou funcionário público)	Salário (0 <= X < 200)	0.6845	0.3155	0.7067	0.2933
Duração do empréstimo (<12)	Valor do Empréstimo (4000<= X<7500)	0.3089	0.6911	0.8653	0.1347

Tabela 21 - Probabilidades dos clientes proprietários de imóveis se tornarem inadimplentes (ou não), dada à observação (ou não) e seus pais, do Classificador GBN com aprendizado de estrutura utilizando a medida BDeu e com seleção de variáveis pelo *Markov Blanket* da variável resposta.

Filho (Nível)	Pai 1 (Nível)	Pai 2 (Nível)		
Histórico de Crédito (Histórico de atraso no pagamento)	Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)	Valor do Empréstimo (4000<= X<7500)	P(Cliente=1| Filho,Pai1,Pai2)	P(Cliente=0| Filho,Pai1,Pai2)
1	0	1	0.7276	0.2724
1	0	0	0.6741	0.3259
0	1	1	0.1575	0.8425
0	1	0	0.6335	0.3665
0	0	1	0.6176	0.3824
0	0	0	0.754	0.246

A análise das Tabelas 19 a 21 indica que a variável “Moradia (Própria)” contribui isoladamente para a redução do risco de inadimplência, em relação aos demais estados de “Moradia (Alugada)” ou com “Moradia (Gratuita)”. No entanto, a interpretação da interação desta *dummy* combinada ao seu pai “Finalidade(Compra de carro usado)”, indica que, entre os clientes que possuem casa própria, aqueles que adquirem o empréstimo para a compra de carro usado apresentam maior risco de inadimplência, em relação aos demais clientes com moradia própria.

Será analisada a interação entre a *dummy* “Histórico de Crédito (Histórico de atraso no pagamento)” e de seus pais na RB, as *dummies*: “Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)” e “Valor do Empréstimo(1000<= X<4000)”. Desta análise, pode-se observar que, entre os clientes com atraso no pagamento, aqueles que também possuem empréstimos entre $4000 e $7500 apresentam menores riscos de inadimplência, do que outros clientes com outros valores de empréstimos e com mesmo histórico de crédito.

Além disso, entre os clientes sem empréstimos tomados ou todos os empréstimos pagos pontualmente, aqueles que também possuem empréstimos entre $4000 e $7500 possuem um maior risco de inadimplência, em relação aos clientes com mesmo histórico de crédito e que contratam outros valores de empréstimo.

5.3 Comparação dos Modelos de Classificação

Nesta Seção, será descrito o estudo comparativo das aplicações dos Classificadores Bayesianos e da Regressão Logística em modelos de *Credit Scoring.*

Para a comparação dos modelos, foi construída a Tabela 22, com as medidas de avaliação dos Classificadores *Naive Bayes*, TAN e GBN finais e da Regressão Logística com seleção de variáveis por *Forward Stepwise,* calculadas a partir das amostras de validação obtidas por *Cross-Validation* com 10 partições.

Tabela 22 - Estatística Kolmogorov-Smirnov, coeficiente Gini, taxas de acerto total (TAT), dos clientes "bons" (TAB), dos clientes "maus" (TAM) e número de variáveis dos modelos de classificação finais.

Modelos de Classificação	Medida	Modelo	TAT	TAB	TAM	KS	Gini	Nº de Variáveis
Regressão Logística	-	Seleção por Forward Stepwise	74.83%	74.00%	75.67%	50.33%	61.46%	28
Naive Bayes	-	Seleção *Wrapper* Backward	74.50%	72.30%	76.70%	49.33%	57.47%	28
TAN	Bayes	Seleção *Wrapper* e com filtragem pelo Ganho de Informação	75.33%	71.00%	79.70%	51.33%	60.81%	27
GBN	Bdeu	Seleção *Markov Blanket*	74.67%	76.30%	76.30%	49.67%	58.84%	15

A Tabela 22 mostra que as taxas de acerto total (TAT) dos modelos finais apresentaram valores próximos. Mas, analisando a taxa de acerto dos "maus", o Classificador TAN final apresentou melhor desempenho, pois sua assertividade nas previsões dos clientes "maus" pagadores foi mais de 3% superior às taxas observdas nos modelos de classificação.

Comparando as medidas Bayes, AIC, MDL, Bdeu e Entropia de avaliação da estrutura durante seu aprendizado observa-se que, nos Classificadores TAN, a medida Bayes apresentou resultados pouco melhores e os aprendizados com as demais medidas não apresentaram diferenças significativas. Nos Classificadores GBN também foi observado o mesmo resultado, exceto no aprendizado de estrutura

com a medida Entropia, que apresentou pior desempenho, em relação a todas as taxas de acerto das previsões, a estatística KS e ao coeficiente Gini.

A filtragem de variáveis pelo ganho de informação contribuiu um aumento nas taxas de acerto das previsões e no poder discriminante dos Classificadores *Naive Bayes* e TAN. No entanto, esse resultado não foi observado nos Classificadores GBN, pois algumas variáveis podem não contribuir individualmente para o ganho de informação da variável reposta, mas combinadas com outras variáveis, elas se tornam importantes para a classificação dos clientes “bons” e “maus” pagadores.

Os Classificadores TAN e GBN permitem analisar as dependências que possam existir entre variáveis preditoras, o que não pode ser feito no Classificador *Naive Bayes* e na Regressão Logística. As análises dos efeitos de interação entre *dummies* preditoras e seus pais na RB indicaram a presença de combinações de variáveis que contribuem para aumentos nos riscos de inadimplência dos clientes, enquanto a observação individual da *dummy* contribui para uma redução nos riscos de inadimplência. Com isso, a análise dos efeitos de interação auxilia na identificação de perfis de risco dos clientes.

6. CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho foram analisadas comparativamente aplicações de Redes Bayesianas (RB) e Regressão Logística em modelos de *Credit Scoring*. Foram avaliadas as RB conhecidas como Classificadores Bayesianos, com estruturas do tipo: *Naive Bayes*, Tree Augmented *Naive Bayes* (TAN) e General Bayesian Network (GBN).

Nos resultados observou-se que os Classificadores *Naive Bayes*, TAN e GBN finais apresentaram desempenho similar ao da Regressão Logística, em relação às taxas de acerto totais, à estatística Kolmogorov-Smirnov e ao coeficiente Gini.

Além disso, o Classificador TAN com aprendizado de estrutura empregando a medida Bayes apresentou a maior taxa de acerto das previsões dos clientes “maus”, sendo esta 4% superior às observadas na Regressão Logística. Portanto, o uso deste modelo auxilia Instituições Credoras a reduzir erros em concessões incorretas a clientes que possam se tornar inadimplentes, em determinado período.

Por isso, o Classificador TAN foi escolhido como o melhor modelo, pois apresentou o melhor desempenho nas previsões dos clientes “maus” pagadores e permitiu uma análise dos efeitos de interação entre variáveis.

O uso das taxas de acerto obtidas da matriz de confusão auxiliou na identificação das categorias de variável resposta (cliente “bom” ou “mau” pagador) cujo modelo é mais assertivo. Podendo ser utilizadas como medidas de desempenho complementares à estatística Kolmogorov-Smirnov e ao coeficiente Gini.

Uma contribuição deste trabalho está na análise comparativa do uso das medidas Bayes, AIC, MDL, Bdeu e Entropia no aprendizado de estrutura em RB. Nos Classificadores TAN, o aprendizado de estrutura com a medida Bayes apresentou resultados pouco melhores e os aprendizados com as demais medidas não apresentaram diferenças significativas. Nos Classificadores GBN também foi observado o mesmo resultado, exceto no aprendizado de estrutura com a medida

Entropia, que apresentou pior desempenho, em relação a todas as taxas de acerto das previsões, a estatística KS e ao coeficiente Gini.

Outra contribuição deste estudo está em avaliar diferentes metodologias de seleção de variáveis nos Classificadores Bayesianos. Com isso, observou-se que a filtragem de variáveis pelo seu ganho de informação para a variável resposta contribuiu para uma melhora no desempenho dos Classificadores *Naive Bayes* e TAN, pois auxiliou no aumento das taxas de acerto das previsões e no poder discriminante destes modelos. No entanto, este resultado não foi observado nos Classificadores GBN. Portanto, nos Classificadores GBN não é adequada a retirada individual das variáveis que menos contribuíam para o ganho de informação da variável reposta, pois estas variáveis na presença de outras podem tornar-se importantes para o modelo de classificação.

A seleção de variáveis pelo método *Wrapper* com busca por *Backward Elimination,* empregada nos Classificadores *Naive Bayes* e TAN, e a seleção das variáveis do *Markov Blanket* da variável resposta “Cliente”, empregada nos Classificadores GBN, auxiliaram no aumento da assertividade das previsões e no poder discriminante dos Classificadores, em relação ao modelo saturado (com todas as variáveis).

Apesar disso, os Classificadores GBN com seleções de variáveis pelo *Markov Blanket* da variável resposta e com aprendizados de estrutura utilizando as medidas AIC e Entropia apresentam fracos desempenhos em todas as medidas de avaliação. Este fato está associado ao grande número de dependências avaliadas durante o seu aprendizado de estrutura, que se agrava quando a base de dados possui um grande número de variáveis e um número restrito de observações.

Este problema não é observado nos aprendizados realizados no Classificador TAN, pois a busca da sua melhor estrutura é restrita ao espaço de estruturas do tipo árvore, na qual são avaliadas dependências entre pares de variáveis preditoras. Portanto, no aprendizado dos Classificadores TAN é avaliado um número menor de dependências, sendo mais adequado na presença de uma base de dados reduzida.

Os Classificadores TAN e GBN permitem a análise de dependências entre variáveis preditoras, assumindo premissas diferentes a respeito de sua estrutura. Neste trabalho, esta análise se resumiu à análise dos efeitos de interação entre variáveis *dummy*.

A análise dos efeitos de interação indicou que algumas combinações de *dummies* (variáveis preditoras) contribuíram para aumentos nos riscos de inadimplência, enquanto a análise individual das *dummies* indicou que sua observação reduz os riscos de inadimplência. O contrário também foi observado, ou seja, algumas combinações *dummies* apresentaram altas probabilidades dos clientes se tornarem bons pagadores, enquanto a observação individual da *dummy* contribuiu para o aumento no risco de inadimplência.

Portanto, a análise dos efeitos de interação entre variáveis na RB permitem comparar combinações de *dummies* com seu respectivo pai na RB e levou a informações a respeito do perfil de risco de inadimplência, que não puderam ser observadas na análise individual das variáveis (*dummies*).

Uma limitação observada neste trabalho está na presença de poucas observações na base de dados *German Credit.* Em Instituições Financeiras são encontradas comumente bases de dados com grande número de observações, com isso é proposta a aplicação dos Classificadores Bayesianos neste tipo de base.

Em trabalhos futuros também é sugerida uma análise mais detalhada da variabilidade das partições da amostra de validação obtidas por Cross-Validation com *10-fold*, a fim de se ter uma análise da estabilidade dos Classificadores Bayesianos.

Para uma análise mais detalhada das dependências obtidas nos aprendizados de estrutura, é sugerida a aplicação de aprendizados a partir de uma estrutura conhecida. A base de dados utilizada pode ser gerada a partir da distribuição conjunta desta RB, podendo ser geradas amostras com diferentes números de observações e com diferentes números de variáveis preditoras. Este estudo permitirá uma análise da estabilidade dos Classificadores Bayesianos com alterações em parâmetros amostrais.

Outra proposta de aplicação dos Classificadores Bayesianos é a obtenção de um modelo híbrido a partir dos Classificadores e da Regressão Logística, como também foi feito no artigo de Arminger, Enache e Bonne (1997). O trabalho de Arminger, Enache e Bonne (1997) avaliou um procedimento combinado de três modelos utilizando seus valores previstos e observados; e os modelos aplicados foram: Regressão Logística, Árvore de Classificação e um tipo de Rede Neural chamada *Feedforward Network*.

REFERÊNCIAS

AGRESTI, W.J. *Practical nonparametric statistics*. 3ª ed. New York: John Wiley and Sons, 1999. 584p.

ANDERSON, R. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. 1ª ed. New York: Oxford University Press, 2007. 731p.

ARMINGER, G.; ENACHE, D.; BONNE, *T. Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feedforward Networks.* Computational Statistics, v.12, n.2, p.293-310, 1997.

BAESENS, B. et al. *Learning bayesian network Classifiers for Credit Scoring Using Markov Chain Monte Carlo Search.* In: Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02), v.3, p.49-52, 2002.

BAESENS, B. et al. *Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers.* European Journal of Operational Research, v. 127, n.2, p. 508-523, 2004.

BANCO CENTRAL DO BRASIL. *Relatório de Inflação*. v.9, n.3. Brasília: 2007. Disponível em: <http://www.bcb.gov.br/htms/relinf/port/2007/09/ri200709P.pdf> Acesso em: 02/12/2007.

BANCO CENTRAL DO BRASIL. *Relatório de Inflação*. v.9, n.1. Brasília: 2007. Disponível em: <http://www.bcb.gov.br/htms/relinf/port/2007/03/ri200703P.pdf> Acesso em: 02/12/2007.

BANCO CENTRAL DO BRASIL. *Resolução 2682*. Brasília: Banco Central do Brasil,1999.

BLAKE, C.; MERZ, C. *UCI repository of machine learning databases.* 1998. Disponível em: <http://www.ics.uci.edu/~mlearn/MLRepository.html>. Acesso em: 01 dez. 2007

BOLFARINE, H ; SANDOVAL, M. C. *Introdução à Inferência Estatística.* Rio de janeiro: Sociedade Brasileira de Matemática, 2001. 125 p.

BUNTINE, W. *A Guide to the Literature on Learning Probabilistic Networks from Data*. IEEE Transactions on Knowledge Data Engineering. v.8, n.2, p.195-210, 1996.

CHANG, K.C. et al. *Bayesian Networks applied to Credit Scoring*. IMA Journal of Mathematics Applied in Business an Industry, v.11, n.1., p.1-18, 2000.

CHENG,J.;GREINER,R. *Comparing bayesian network classifiers*. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI'99). Morgan Kaufmann, p.101-107, 1999.

CHENG,J.;GREINER,R. *Learning Bayesian Belief Network Classifiers: Algorithms and System.* In: Proceedings of 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, v.2056, p.141-151, 2001.

CHICKERING, D.M.; GEIGER, D.; HECKERMAN, D.E. *Learning Bayesian Networks is NP-Hard.* Microsoft Research Technical Report, MSR-TR-94-17, 1994.

CHOW, C.K.; LIU, C.N. *Approximating discrete probability distributions with dependence trees.* IEEE Transactions on Information Theory, v.14, n.3, p. 462-467, 1968.

CONOVER, W.J. *Practical nonparametric statistics*. 3ª ed. New York: John Wiley and Sons, 1999. 584p.

EISENBEIS, R.A*. Problems in Applying Discriminant Analysis in Credit Scoring Models.* Journal of Banking and Finance, v.2, p.205-219, 1978.

FOLHA ON LINE. *Entenda a crise com o mercado imobiliário nos EUA.* Folha de São Paulo, 16/08/2007. Disponível em: <http://www1.folha.uol.com.br/folha/dinheiro/ult91u320606.shtml> Acesso em: 02/12/2007.

FOLHA ON LINE. *Entenda a crise financeira que atinge a economia dos EUA*. Folha de São Paulo, 15/09/2008. Disponível em: <http://www1.folha.uol.com.br/folha/dinheiro/ult91u320606.shtml> Acesso em: 16/11/2008.

FRIEDMAN,N.;GEIGER,D.;GOLDSZMIDT,M. *Bayesian Network Classifiers.* Machine Learning, v.29, p131-163, 1997.

FRIEDMAN,N.; GOLDSZMIDT,M. *Building Classifiers Using Bayesian Networks*. In: Proceedings of Thirteenth National Conference on Artificial Intelligence (AAAI96), v.2, p.1277-1284, 1996.

GIUDICI, P. *Integration of Qualitative and Quantitative Operational Risk Data: A Bayesian Approach.* Operational Risk Modelling and Analysis: Theory and Practice, p.131-138, 2004.

GOUVÊA, M.A.; GONÇALVES, E.B. *Análise de Risco de Crédito com o uso de Modelos de Redes Neurais e Algoritmos Genéticos.* In: IX SEMEAD Seminários em Administração, 2006.

GUYON, I.; ELISSEEFF, A. *An Introduction to variable and feature selection.* Journal of Machine Learning Research, v.3, p.1157-1182, 2003.

HAND, D. J.; HENLEY; W. E. *Statistical classification methods in consumer Credit Scoring: a review.* Journal of the Royal Statistical Society: Series A (Statistics in Society), v.160, n.3, p.523-541, 1997.

HAND, D. J. *Modelling Consumer Credit Risk.* IMA Journal of Management Mathematics, v.12, n.2, p.139-155, 2001.

HECKERMAN, D.; GEIGER, D.;CHICKERING, D. *Learning Bayesian networks: The combination of knowledge and statistical data*. Machine Learning, v.20, n.3, p.197-243, 1995.

HECKERMAN, D. *Bayesian Networks for Data Mining.* Data Mining and Knowledge Discovery, v.1, p.79-119, 1997.

HECKERMAN, D. *A tutorial on learning with bayesian networks.* Data Mining and Knowledge Discovery, v.1, p.79-119, 1997.

HOSMER, D.W.; LEMESHOW, S. *Applied Logistic Regression.* 1ª ed. New York: John Wiley, 1989. 392p.

JENSEN, F.V. *Bayesian Networks and Decision Graphs.* New York: Springer, 2001. 268p.

LANGLEY, P.; SAGE, S. *Induction of Selective Bayesian Classifiers*. In: Proceedings of the 10th Annual Conference on Uncertainty in Artificial Intelligence (UAI-94). San Francisco: Morgan Kaufmann, 1994. p. 399-406.

LECUMBERRI, L.F.L; DUARTE, A.M. *Uma metodologia para o gerenciamento de modelos de escoragem em operações de crédito de varejo no Brasil*. Revista de Economia Aplicada, v.7,n.4, p. 795-818, 2003.

LOURENÇO, F. C. *Vantagens do uso de métodos quantitativos no ciclo do crédito*. Revista Business da Equifax, n.251, 2005. Disponível em: <http://www.equifax.com.br/rev_bus/05_abr/pag_pvi.asp>. Acesso em: 01dez.2007.

MADDEN, M. G. *The performance of Bayesian network classifiers constructed using different techniques.* In: Proceedings of the 14th European Conference on Machine Learning, Workshop on Probabilistic Graphical Models for Classification. p. 59–70, 2003.

MARCHESINI, A *Em 2008, volume de crédito deve somar 38% do PIB; juros cairão.* Infomoney, 23/11/2007. Disponível em: <http://web.infomoney.com.br/templates/news/view.asp?codigo=864761&path=/suasfinancas/> Acesso em: 02/12/2007.

MATSUURA, J.P. *Discretização para Aprendizagem Bayesiana: Aplicação no Auxílio à Validação de Dados em Proteção ao Voô.* 2003. 81p. Dissertação (Mestrado) - Instituto Tecnológico de Aeronáutica, São José dos Campos, 2003.

MCCULLAGH, P; NELDER, J. A. *Generalized Linear Models.* 2ª ed. London: Chapman and Hall, 1989. 511p.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to Linear Regression Analysis.* 3ª ed. New York: John Wiley, 2001. 672p.

NEAPOLITAN, R.E. *Learning Bayesian Networks.* New Jersey: Prentice Hall, 2004. 674 p.

PAIVA, P. *Operações de empréstimos já substituíram aplicações no mercado como maior fonte de ganho das instituições.* Estados de Minas, 19/11/2007. Disponível em: <http://www.uasf.sebrae.com.br/uasfgestao/uasfnoticias/nov%202007/not3332/view > Acesso em: 02/12/2007.

PAULA, G.A. (2004). *Modelos de Regressão com Apoio Computacional.* São Paulo: Instituto de Matemática e Estatística Universidade São Paulo. 2004. 245p. Disponível em: <http://www.ime.usp.br/~giapaula/livro.pdf>. Acesso em: 01dez.2007.

PEREIRA, G. H. A. *Modelos de Risco de Crédito de Clientes: Uma Aplicação a Dados Reais*. 2004. 96p. Dissertação (Mestrado) – Instituto de Matemática e Estatística, Universidade São Paulo, São Paulo, 2004.

POKU, K.A. *Operational Risk management - Implementing a Bayesian Network for Foreign Exchange and Money Market Settlement.* 2005. 134p. Ph.D. Thesis - Faculty of Economics and Business Administration, University of Göttingen, Alemanha, 2005.

ROSA, P. T. M. *Modelos de Credit Scoring: Regressão Logística, CHAID e REAL.* 2000. 68p. Dissertação (Mestrado) – Instituto de Matemática e Estatística, Universidade São Paulo, São Paulo, 2000.

SAHEKI, A. H. Construção de uma Rede Bayesiana aplicada ao diagnóstico de *doenças cardíacas*. 2005. 70p. Dissertação (Mestrado) – Escola Politécnica, Universidade São Paulo, São Paulo, 2005.

SECURATO, J. R. *Crédito: Análise e Avaliação do Risco – Pessoas Físicas e Jurídicas.* 1ª ed. São Paulo: Saint Paul, 2002. 354 p.

SIQUEIRA, J. *Expansão do crédito em 2007 supera estimativas do Itaú.* Reuters, 06/11/2007. Disponível em: <http://oglobo.globo.com/economia/mat/2007/11/06/327051496.asp> Acesso em: 02/12/2007.

VASCONCELLOS, M. S. *Proposta de Método para análise de concessões de Crédito a Pessoas Físicas.* 2002. 119p. Dissertação (Mestrado) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2002.

ZHANG, N.; POOLE, D. *Exploiting Causal Independence in Bayesian Network Inference.* Journal of Artificial Intelligence Research, v. 5, p. 301-328, 1996.

WEST, D. *Neural Network Credit Scoring Models.* Computers and Operations Research, v. 27, n.11, pp. 1131-1152, 2000.

WITTEN, I. H.; FRANK, E. Data Mining: Practical Machine Learning Tools and Techniques. 2ª ed. San Francisco: Morgan Kaufmann, 2005. 525p.

APÊNDICE DE TABELAS

Tabela A. 1 - Nomes das variáveis da base de dados *German Credit*.

Variável	Nome original (em inglês)
Salário	*Status of existing checking account*
Duração do Empréstimo	*Duration in months*
Histórico de Crédito	*Credit history*
Finalidade	*Purpose*
Valor do Empréstimo	*Credit amount*
Poupança do Cliente	*Savings account/bonds*
Tempo de Trabalho	*Present employment since*
Taxa de juros em % do valor do empréstimo	*Installment rate in percentage of disposable income*
Estado Civil e Sexo	*Personal status and sex*
Outras dívidas ou garantias	*Other debtors/guarantors*
Tempo de Residência	*Present residence since*
Bens	*Property*
Idade	*Age in years*
Outros Empréstimos	*Other installment plans*
Moradia	*Housing*
Número de créditos concedidos em seu banco	*Number of existing credits at this bank*
Emprego	*Job title*
Número de Dependentes	*Number of people being liable to provide maintenance for*
Telefone Próprio	*Telephone*
Estrangeiro	*Foreign worker*
Classificação do cliente como "bom" ou "mau" pagador	*Good or bad credit rating*

Tabela A. 2 - Valores do Risco Relativo, *Weights of Evidence* (WOE) e outras medidas descritivas da variável “Histórico de Crédito”, antes do agrupamento de algumas de suas categorias.

Categoria	Número de "bons"	Número de "maus"	%bons	%maus	Risco Relativo	WOE
Sem empréstimos tomados	5	25	1.67%	8.33%	0.200	-1.609
Todos os empréstimos pagos pontualmente	7	28	2.33%	9.33%	0.250	-1.386
Existem empréstimos pagos pontualmente	154	169	51.33%	56.33%	0.911	-0.093
Histórico de atraso no pagamento	28	28	9.33%	9.33%	1.000	0.000
Atraso no pagamento ou com empréstimos tomados em outras Instituições	106	50	35.33%	16.67%	2.120	0.751
Total	300	300	100.00%	100.00%	1.000	0.000

Tabela A. 3 - Valores do Risco Relativo, *Weights of Evidence* (WOE) e outras medidas descritivas da variável “Bens”, antes do agrupamento de algumas de suas categorias.

Categoria	Número de "bons"	Número de "maus"	%bons	%maus	Risco Relativo	WOE
Imóvel	101	60	33.67%	20.00%	1.683	0.521
Seguro de Vida	66	71	22.00%	23.67%	0.930	-0.073
Carro ou outros	95	102	31.67%	34.00%	0.931	-0.071
Não possui bens	38	67	12.67%	22.33%	0.567	-0.567
Total	300	300	100.00%	100.00%	1.000	0.000

Tabela A. 4 - Valores do Risco Relativo, *Weights of Evidence* (WOE) e outras medidas descritivas da variável “Poupança do Cliente”, antes do agrupamento de algumas de suas categorias.

Categoria	Número de "bons"	Número de "maus"	%bons	%maus	Risco Relativo	WOE
< $100	163	217	54.33%	72.33%	0.751	-0.286
$100<= X < $500	26	34	8.67%	11.33%	0.765	-0.268
$500<= X < $1000	19	11	6.33%	3.67%	1.727	0.547
>= $1000	16	6	5.33%	2.00%	2.667	0.981
Não possui ou não conhecida	76	32	25.33%	10.67%	2.375	0.865
Total	300	300	100.00%	100.00%	1.000	0.000

Tabela A. 5 - Valores do Risco Relativo, *Weights of Evidence* (WOE) e outras medidas descritivas da variável “Outros Empréstimos”, antes do agrupamento de algumas de suas categorias.

Categoria	Número de "bons"	Número de "maus"	%bons	%maus	Risco Relativo	WOE
Bancos	32	57	10.67%	19.00%	0.561	-0.577
Lojas	9	19	3.00%	6.33%	0.474	-0.747
Nenhum	259	224	86.33%	74.67%	1.156	0.145
Total	300	300	100.00%	100.00%	1.000	0.000

Tabela A. 6 - Valores do Risco Relativo, *Weights of Evidence* (WOE) e outras medidas descritivas da variável “Tempo de Trabalho”, antes do agrupamento de algumas de suas categorias.

Categoria	Número de "bons"	Número de "maus"	%bons	%maus	Risco Relativo	WOE
Desempregado	19	23	6.33%	7.67%	0.826	-0.191
X < 1 ano	48	70	16.00%	23.33%	0.686	-0.377
1 <= X < 4 anos	105	104	35.00%	34.67%	1.010	0.010
4 <= X < 7 anos	46	39	15.33%	13.00%	1.179	0.165
X >= 7 anos	82	64	27.33%	21.33%	1.281	0.248
Total	300	300	100.00%	100.00%	1.000	0.000

Tabela A. 7 - Valores do Risco Relativo, *Weights of Evidence* (WOE) e outras medidas descritivas da variável “Moradia”, antes do agrupamento de algumas de suas categorias.

Categoria	Número de "bons"	Número de "maus"	%bons	%maus	Risco Relativo	WOE
Alugada	41	70	13.67%	23.33%	0.586	-0.535
Própria	233	186	77.67%	62.00%	1.253	0.225
Moradia gratuita	26	44	8.67%	14.67%	0.591	-0.526
Total	300	300	100.00%	100.00%	1.000	0.000

Tabela A. 8 - Valores do Risco Relativo, *Weights of Evidence* (WOE) e outras medidas descritivas da variável “Estado Civil e Sexo”, antes do agrupamento de algumas de suas categorias.

Categoria	Número de "bons"	Número de "maus"	%bons	%maus	Risco Relativo	WOE
Masculino divorciado ou separado	15	20	5.00%	6.67%	0.750	-0.288
Feminino divorciada, separada ou casada	79	109	26.33%	36.33%	0.725	-0.322
Masculino solteiro	179	146	59.67%	48.67%	1.226	0.204
Masculino casado ou viúvo	27	25	9.00%	8.33%	1.080	0.077
Total	300	300	100.00%	100.00%	1.000	0.000

Tabela A. 9 - Valores do Risco Relativo, *Weights of Evidence* (WOE) e outras medidas descritivas da variável “Emprego”, antes do agrupamento de algumas de suas categorias.

Categoria	Número de "bons"	Número de "maus"	%bons	%maus	Risco Relativo	WOE
Desempregado ou empregado com baixa qualificação ou sem trabalho formal	8	7	2.67%	2.33%	1.143	0.134
Empregado com baixa qualificação e com trabalho formal	73	56	24.33%	18.67%	1.304	0.265
Empregado qualificado ou funcionário público	176	186	58.67%	62.00%	0.946	-0.055
Executivo, profissional liberal, empregado altamente qualificado ou oficial	43	51	14.33%	17.00%	0.843	-0.171
Total	300	300	100.00%	100.00%	1.000	0.000

Tabela A. 10 - Risco Relativo das variáveis preditoras (*dummies)* e de seus pais no Classificador TAN com aprendizado de estrutura utilizando a medida Bayes e com seleção com o método *Wrapper* com busca por *Backward Elimination*.

Variável Preditora (Nível)	Pais (Nível)	Risco Relativo
Bens (Imóvel)	Duração do empréstimo (X <12)	2.654
Salário (X < $0)	Tempo de Residência (X <3)	0.516
Poupança do Cliente (X < $500)	Salário (X < $0)	2.641
Outros Empréstimos (Nenhum)	Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)	0.307
Outras dívidas ou garantias (Co-aplicante)	Número de créditos concedidos em seu banco (X >=2)	0.751
Finalidade (Compra de carro usado)	Valor do Empréstimo (1000<= X <4000)	0.361
Histórico de Crédito (Sem empréstimos tomados ou todos os empréstimos pagos pontualmente)	Salário (0 <= X < 200)	2.030
Tempo de Trabalho (Desempregado ou menos de 1 ano)	Estado Civil e Sexo (Masculino solteiro)	0.355
Estado Civil e Sexo (Masculino solteiro)		1.560
Moradia (Própria)	Finalidade (Compra de carro usado)	0.292
Emprego (Desempregado ou empregado com baixa qualificação)	Bens (Imóvel)	3.492
Estrangeiro (Não)	Duração do empréstimo (X <12)	8.434
Duração do empréstimo (12<= X <16)	Valor do Empréstimo (X <1000)	5.961
Valor do Empréstimo (4000<= X <7500)	Estado Civil e Sexo (Masculino solteiro)	2.013
Taxa de juros em % do valor do empréstimo (X <4)	Valor do Empréstimo (X <1000)	2.923
Tempo de Residência (X <3)	Moradia (Própria)	3.974
Idade (X <25)	Estado Civil e Sexo (Masculino solteiro)	0.305
Número de créditos concedidos em seu banco (X >=2)	Idade (X <25)	0.354

Tabela A. 11 - Estatística Kolmogorov-Smirnov, coeficiente Gini, taxa de acerto total (TAT), taxa de acerto dos clientes "bons" (TAB), taxa de acerto dos clientes "maus" (TAM) e número de variáveis dos Classificadores GBN com aprendizado de estrutura utilizando diferentes medidas.

Medida	Modelo	TAT	TAB	TAM	KS	Gini	Nº de Variáveis
MDL	Modelo Saturado	70.67%	69.30%	72.00%	42.00%	53.69%	48
	Seleção *Markov Blanket* com Filtragem pelo Ganho de Informação	73.17%	71.30%	75.00%	47.00%	59.00%	13
	Seleção *Markov Blanket*	73.00%	72.70%	73.30%	46.00%	59.40%	16
Bayes	Modelo Saturado	68.50%	65.30%	71.70%	40.00%	51.47%	48
	Seleção *Markov Blanket* com Filtragem pelo Ganho de Informação	70.83%	69.70%	72.00%	42.33%	51.79%	21
	Seleção *Markov Blanket*	71.67%	67.70%	75.70%	43.67%	57.03%	30
AIC	Modelo Saturado	72.50%	70.30%	74.70%	45.33%	55.70%	48
	Seleção *Markov Blanket* com Filtragem pelo Ganho de Informação	69.50%	69.00%	70.00%	39.67%	51.27%	30
	Seleção *Markov Blanket*	70.33%	68.70%	72.00%	43.00%	54.54%	37
Bdeu	Modelo Saturado	71.50%	71.30%	71.70%	44.67%	55.84%	48
	Seleção *Markov Blanket* com Filtragem pelo Ganho de Informação	60.17%	59.00%	61.30%	24.67%	27.26%	17
	Seleção *Markov Blanket*	74.67%	76.30%	76.30%	49.67%	58.84%	15
Entropia	Seleção *Markov Blanket*	65.83%	67.70%	64.00%	33.00%	43.23%	48
	Seleção *Markov Blanket* com Filtragem pelo Ganho de Informação	62.00%	65.00%	59.00%	29.33%	38.07%	31