APLICAÇÃO DE REDES BAYESIANAS NA ANÁLISE DE RISCO DE CRÉDITO BANCÁRIO

Amanda Trojan Fenerich
Pontifícia Universidade Católica do Paraná – PPGEPS PUCPR
R. Imaculada Conceição, 1155 .Bloco 2 - 2º Andar. Prado Velho – Curitiba, PR
amanda.fenerich@gmail.com

Maria Teresinha Arns Steiner
: Pontifícia Universidade Católica do Paraná, Escola Politécnica – PPGEPS PUCPR
R. Imaculada Conceição, 1155 .Bloco 2 - 2º Andar. Prado Velho – Curitiba, PR
maria.steiner@pucpr.br

Pedro José Steiner
Universidade Positivo – PMDA
Praça General Osório, 125 - Centro, Curitiba – PR
pedrosteiner@ufpr.br

Edevar André Tochetto
Universidade Positivo – PMDA
Praça General Osório, 125 - Centro, Curitiba – PR
edevar_t@hotmail.com

Diego Paolo Tsutsumi
Universidade Tecnológica Federal do Paraná – PPGCA
. Av. Sete de Setembro 3165, Bloco B - 1° andar. Rebouças – Curitiba, PR
diegotsu@gmail.com

Resumo: As instituições financeiras têm importantes decisões a serem tomadas diariamente, dentre as quais está a concessão de crédito para pessoas físicas e/ou jurídicas. Esta decisão deve ser tomada com base em uma rigorosa avaliação, uma vez que envolve riscos que podem trazer sérias consequências para a instituição. Isso exige o desenvolvimento de modelos cada vez mais sofisticados para as decisões de concessão de crédito, de forma a torná-las mais objetivas, rápidas e confiáveis. O objetivo é a aplicação das Redes *Bayesianas* (RB) para auxiliar as instituições financeiras na análise de concessão de crédito bancário. Foi realizado um levantamento de dados em uma grande instituição financeira nacional, sendo consideradas 8.358 instâncias e oito atributos para a construção do modelo. Os resultados da aplicação das RB no problema de avaliação de risco de crédito para pessoa jurídica foram satisfatórios, com o valor da acurácia em torno de 85% nas configurações consideradas.
Palavras-chave: Aprendizagem de Máquina; Modelos de Classificação de Riscos; Pessoa Jurídica.

Abstract: Financial institutions have important decisions to make every day, among them is the granting of credit to individuals and /or legal entities. This decision must be made on the basis of a rigorous evaluation, since it involves risks that can have serious consequences for the institution. This requires the development of increasingly sophisticated models for lending decisions, in order to make them more objective, faster and more reliable. The goal is to apply Bayesian Networks (BN) in order to assist financial institutions in the analysis of bank credit granting. A data survey was carried out in a large national financial institution with 8,358 instances and eight attributes being considered for the construction of the model. The results of the application of the BNs on the problem of credit risk assessment for legal entities were satisfactory, with an accuracy value of around 85% in the configurations considered here.
Key words: Machine Learning; Credit Scoring; Companies.

https://proceedings.science/p/85557?lang=pt-br

1. Introdução

A tomada de decisão é um processo de identificação de um problema, ou de uma oportunidade, e a seleção de uma linha de ação para resolvê-lo. E este processo pode ser aplicado nos três níveis hierárquicos de uma empresa, estratégico, gerencial e operacional. Quando os gerentes se veem diante de uma situação na qual uma decisão deve ser tomada, duas opções básicas se apresentam: usar apenas a intuição gerencial e/ou realizar um processo de modelagem da situação, aplicando técnicas apropriadas para encontrar a melhor solução para o problema ou oportunidade (Lachtermacher, 2016). Ainda segundo este autor, até recentemente, não existiam nem dados, nem informações sobre os problemas, nem poder computacional para resolvê-los. Com o advento dos microcomputadores e com o aprimoramento da tecnologia, um número cada vez maior de empresas passou a optar pela elaboração de modelos para auxiliar esse processo.

Conforme Hillier e Lieberman (2013), com o aumento da complexidade dos problemas, a necessidade de encontrar a melhor maneira para solucioná-los criaram as condições necessárias para o surgimento de técnicas de Pesquisa Operacional (PO), que é uma área multidisciplinar voltada para a aplicação de métodos analíticos avançados para auxiliar na tomada de melhores decisões (*Institute for Operations Research and the Management Sciences*; INFORMS).

Devido à natureza computacional e estatística da maioria das técnicas, a PO também possui fortes laços com a ciência da computação e ciência de dados. Alguns dos principais campos teóricos nos quais as técnicas da PO se apoiam são: mineração de dados; aprendizagem de máquina; análise de decisão; modelagem matemática; engenharia; teoria dos jogos; otimização; probabilidade e estatística; logística; redes sociais; gestão de projetos; entre outras.

Neste artigo, especial atenção será dada ao campo referente à aprendizagem de máquina (*Machine Learning*; ML), com o objetivo de aplicar uma técnica da aprendizagem supervisionada, denominada Redes *Bayesianas* (RB), para auxiliar as instituições financeiras na análise de concessão de crédito bancário. O termo "crédito", quando relacionado ao ambiente bancário, é um tipo de transação, ou contrato, pelo qual a instituição financeira dispõe certa quantia de dinheiro a alguém (pessoa física ou pessoa jurídica), mediante o compromisso de devolvê-lo posteriormente, em uma ou mais parcelas. Como se trata de uma operação de risco, uma vez que na concessão de crédito existe a possibilidade da "não devolução" do dinheiro emprestado, para os bancos esta é uma decisão complexa que pode envolver sérias consequências futuras e, portanto, deve ser examinada com muito rigor. Nas operações de crédito a pessoas jurídicas que, em conjunto, podem chegar a volumes significativos de recursos financeiros, a avaliação do risco de crédito se torna ainda mais importante, pois nos casos extremos da pior das hipóteses, estas operações podem levar à liquidação da instituição financeira (Rodrigues Capelletto e Corrar, 2008).

Os métodos tradicionais de decisões sobre concessão de crédito a clientes são fundamentados em julgamentos humanos a partir de experiências anteriores dos gestores sendo, portanto, bastante subjetivos. As pressões econômicas decorrentes da elevada demanda por crédito, a grande competição comercial do setor e o surgimento de novas tecnologias computacionais têm levado ao desenvolvimento de modelos cada vez mais sofisticados, para as decisões de concessão de crédito, de forma a torná-las mais objetivas, rápidas e confiáveis (Albuquerque; Medina e Silva, 2017).

A análise discriminante, a regressão logística, as árvores de decisão, utilizadas por Medina e Selva (2013), bem como a programação linear, podem ser destacadas como técnicas para construção de modelos de risco de crédito (Samejima; Doya e Kawato, 2003). O uso de métodos com abordagem de decisão multicritério também têm sido utilizados para avalição do risco de crédito (Doumpos et al., 2002). O constante aprimoramento das técnicas para auxiliar nesta tomada de decisão tem sido feito, principalmente por pesquisadores da área de Inteligência Artificial (IA). O uso de ferramentas de análise de banco de dados, Mineração de Dados (*Data Mining; DM*) e Aprendizado de Máquina, têm sido algumas das novas tecnologias utilizadas para aumentar a qualidade e a eficiência das decisões em instituições bancárias (Steiner et al., 2007).

Este artigo está estruturado da seguinte forma: a seção 2 apresenta o referencial teórico sobre a aprendizagem de máquina e RB; em seguida, na seção 3, estão detalhados alguns trabalhos correlatos sobre modelos para avaliação do risco de crédito; na seção 4 é apresentada a metodologia; e nas seções 5 e 6, estão os resultados e as discussões, respectivamente.

2. Aprendizagem de Máquina

A aprendizagem de máquina é um ramo da IA, que engloba várias técnicas, algoritmos e teorias, tais como álgebra linear, probabilidade, grafos, entre outros. Estas visam um objetivo comum, o de possibilitar que uma máquina possa, através de observações de dados reais do modelo, reconhecer padrões, tomar decisões, classificar, prever acontecimentos, extrair e representar conhecimento. Por exemplo, um servidor de *e-mails* pode aprender a classificar se um *e-mail* é ou não um *spam*, sendo que o conceito de *spam* à primeira vista é subjetivo, porém uma máquina pode aprender este conceito apenas com observações de humanos.

Segundo Alpaydin (2010), a aprendizagem de máquina é programar computadores para otimizar os parâmetros de um modelo usando uma base de treinamento ou experiências passadas, sendo que tal modelo é dito preditivo quando faz previsões sobre o futuro ou descritivo quando extrai conhecimento dos dados, ou ambos. Para Russell e Norvig (2004), a ideia da aprendizagem de máquina não é ser um modelo estático, mas sim que a máquina aprenda continuamente, para que ela esteja sempre diminuindo os erros da tarefa designada (predição, classificação, etc.). A Figura 1 mostra uma classificação, na qual são elencados os paradigmas, teorias, métodos e modelos inerentes ao aprendizado de máquina.

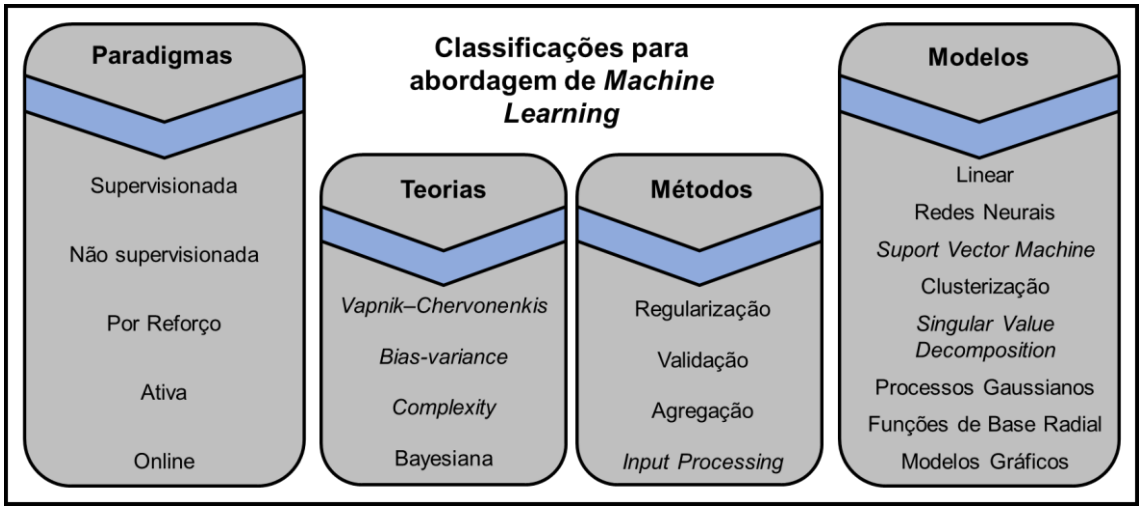


Figura 1 - Classificações para abordagem de *Machine Learning*.
Fonte: Abu-Mostafa, *California Institute of Technology*, 2016.

Existem três paradigmas principais, reconhecidos pela comunidade científica, de processos de aprendizado de máquina: 1) Aprendizagem Supervisionado; 2) Aprendizagem Não-Supervisionado; 3) Aprendizagem por reforço. Os conceitos de outros tipos de paradigmas, como ativa e online, vêm sendo ao pouco introduzidos, mas ainda não são totalmente reconhecidos.

O problema da aprendizagem supervisionada envolve aprendizagem de uma função a partir de exemplos de suas entradas e saídas, ou seja, existe um supervisor ensinando qual saída é correta para cada entrada. Portanto, neste paradigma existe sempre uma base de dados de treinamento, que contém as saídas corretas para cada entrada. Com esta base, a função é aprendida e poderá ser usada em outra base (que possui somente entradas) para prever saídas. Na aprendizagem não-supervisionada, o objetivo é reconhecer padrões dentro de uma base de dados, e não segue a lógica entrada-saída como no paradigma anterior, portanto, não há um supervisor. Na base de dados, por exemplo, pode existir alguma repetição, uma ordem ou uma aglomeração

de informação, e a extração desse tipo de informação é o núcleo da aprendizagem não-supervisionada

Russell e Norvig (2004) mencionam que o problema da aprendizagem por reforço é o mais geral das três categorias. Neste caso, ao invés de ser informado sobre o que fazer por um supervisor, a máquina deve aprender como o ambiente funciona a partir de consequências das suas ações no ambiente. Por exemplo, a ação de acelerar muito o carro em uma estrada gera a consequência de levar uma multa; então a máquina deverá aprender com essa experiência e agir de forma diferente em situações futuras. Enfatizando o caso da aprendizagem por reforço, Apaldyn (2010) relata que em algumas aplicações, o objetivo do problema é encontrar uma sequência de ações dentro de uma política e, nessas situações, a máquina deve ser capaz de identificar o quão bom é a política.

Dentre todas as teorias, os métodos e os modelos que tangem a aprendizagem de máquina, especial atenção será dada à Teoria Bayesiana e aos Modelos Gráficos, pois estes serão a base para a implementação dos objetivos desta pesquisa. O teorema matemático de *Bayes* serve de base para muitas aplicações em aprendizagem de máquina, tanto para métodos de treinamento quanto para modelos representativos, podendo ser usado também na estimativa de parâmetros. Além do mais, esta teoria serve de base para os modelos gráficos probabilísticos, que representam uma distribuição de probabilidade conjunta de várias variáveis a partir da construção de grafos.

Segundo Koller e Friedman (2009), os modelos gráficos probabilísticos usam uma representação baseada em grafos para codificar uma distribuição complexa sobre um espaço de alta dimensão. Neste tipo de representação gráfica, geralmente, os nós correspondem às variáveis em nosso domínio e as arestas correspondem a interações probabilísticas diretas entre elas. Ainda segundo a autora, a linguagem gráfica explora a estrutura que aparece presente em muitas distribuições que queremos codificar na prática: a propriedade de que as variáveis tendem a interagir diretamente uma com as outras.

Dentre as vantagens de utilizar um modelo gráfico, destacam-se a possibilidade de tratar de forma natural e compacta as distribuições, mesmo em casos em que a representação explicita da distribuição conjunta é “astronomicamente” grande; a mesma estrutura permite que a distribuição seja usada para inferência; e facilita a construção dos modelos que aprendem a partir de dados que fornecem uma boa aproximação à experiência passada (Koller e Friedman, 2009).

2.1 Redes Bayesianas

Uma rede bayesiana é um modelo gráfico que consiste de duas partes, *G* e *P*, onde *G* é um gráfico acíclico dirigido (*Directed Acyclic Graph*; DAG), cujos nós representam variáveis aleatórias e os arcos entre nós representam dependência condicional das variáveis aleatórias; e *P* é um conjunto de distribuições das probabilidades condicionais (*Conditional Probability Distribution*; CPD) para cada nó condicionada em seus pais (Dean, 2014).

As distribuições de probabilidade condicional para cada nó podem ser obtidas de maneira prévia, a partir do conhecimento de especialistas, ou então aprendidas, a partir de dados. Koller e Friedman (2009) ressaltam que o gráfico *G* pode ser visto de duas maneiras diferentes, porém equivalentes: “como uma estrutura de dados que fornece o esqueleto para representar uma distribuição conjunta de forma compacta numa forma fatorizada; como uma representação compacta para um conjunto de suposições de independência condicional sobre uma distribuição”.

A Figura 2 mostra um exemplo de uma rede Bayesiana para o problema de compra de roupa, que leva em consideração os seguintes atributos: a estação do ano; a localização; o peso e a cor do tecido. A resolução deste problema fornece, por exemplo, as probabilidades de comprar ou um casaco, ou uma bermuda, ou uma camisa, levando em consideração a probabilidade dos atributos anteriormente citados. Sendo assim, se a estação do ano for primavera, e a localização for New York, a probabilidade de comprar uma bermuda será maior do que a probabilidade de comprar um casaco.

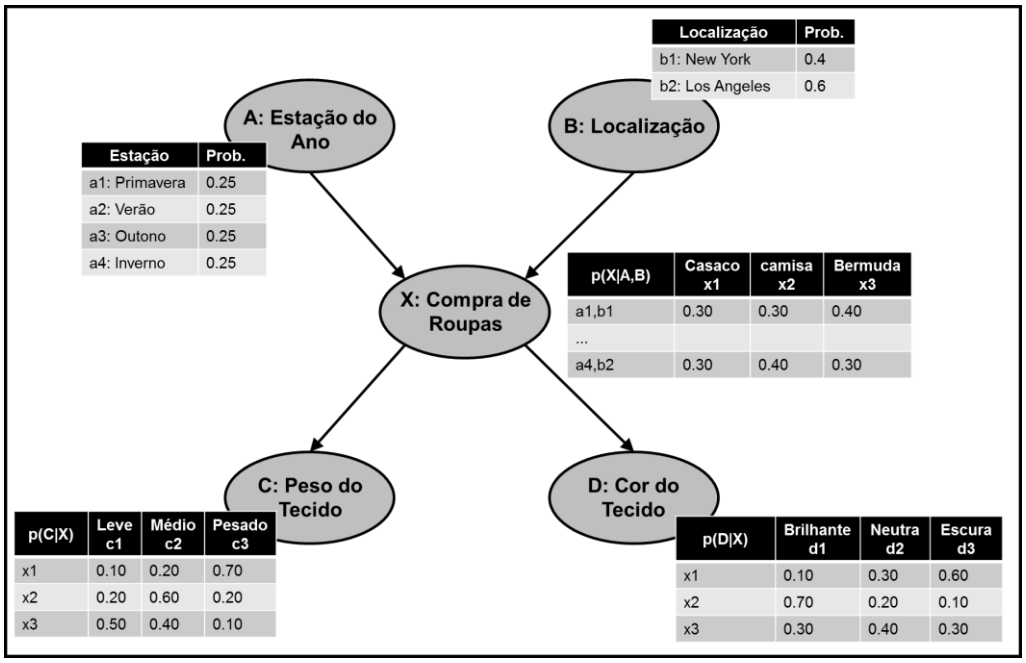


Figura 0 – Exemplo de uma RB aplicada ao problema de compra de roupa.

Apesar do nome, as RBs não implicam necessariamente que elas dependam de estatísticas bayesianas. Pelo contrário, eles são assim chamados porque usam a regra de Bayes para inferência probabilística. É possível usar estatísticas bayesianas para treinar uma RB, mas há também muitas outras técnicas que estão mais intimamente relacionadas com os métodos estatísticos tradicionais (Dean, 2014). Russell *et al.* (1995) sugerem o método do gradiente descendente para a aprendizagem em RBs, análogo ao caso de aprendizagem de redes neurais, que pode ser aplicado para encontrar o conjunto ideal de pesos (valores de probabilidade).

3 Modelos para Avaliação do Risco de Crédito

Toda a operação de crédito concedido, por envolver expectativa futura de pagamentos, está exposta a riscos. Frascaroli, Paes e Ramos (2010) destacam quatro tipos de riscos existentes na literatura, sendo eles: o risco financeiro; risco operacional; risco legal e risco sistêmico, como mostrado na Figura 3.

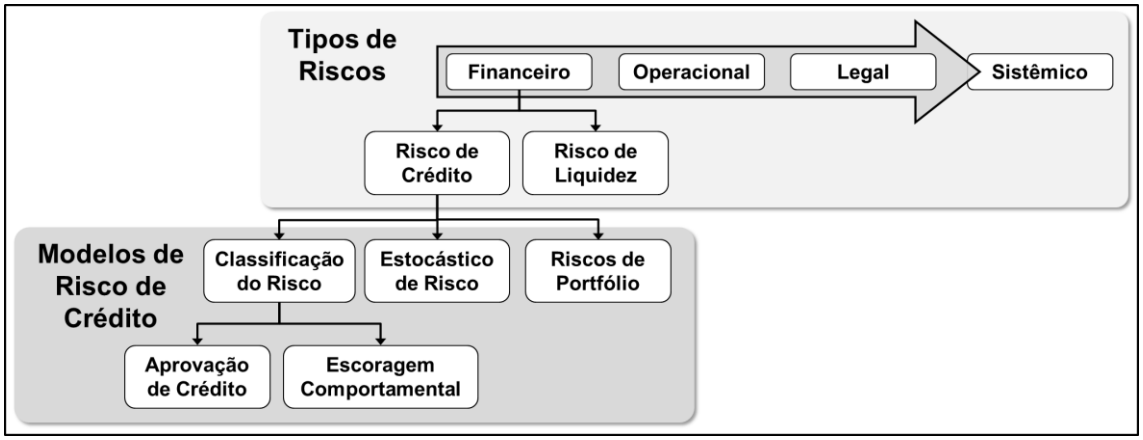


Figura 3 – Tipos de riscos e classificação dos modelos de riscos de crédito.

Os riscos financeiros são os riscos de crédito e de liquidez. O risco operacional é aquele decorrente de falhas humanas, de equipamentos e de sistemas operacionais, aumentando os riscos de crédito e de liquidez. Já o risco legal é aquele decorrente de um embasamento legal e regulamentar inadequado aos sistemas de liquidação cuja atividade destina-se ao controle dos processos. Estes poderão, em conjunto ou separadamente, provocar o risco sistêmico, que pode ser definido como a possibilidade de falhas de crédito e/ou liquidez capazes de provocar a inadimplência generalizada dos participantes de um sistema financeiro (Frascaroli, Paes e Ramos, 2010). O risco de crédito é definido como a possibilidade de ocorrência de perdas financeiras associadas ao não cumprimento das obrigações nos termos pactuados, à desvalorização de contrato de crédito, à redução de ganhos ou remunerações, às vantagens concedidas na renegociação e aos custos de recuperação (Albuquerque, Medina e Silva; 2017).

Os modelos que avaliam e/ou mensuram os riscos de crédito, podem ser classificados em três grupos, conforme Silva Brito e Assaf Neto (2008): modelos de classificação de risco, modelos estocásticos de risco de crédito e modelos de risco de portfólio. Os modelos de classificação de risco são aqueles que procuram quantificar o risco de um tomador ou de uma operação e atribuem nota ao devedor ou ao empréstimo efetuado. Os modelos estocásticos de risco de crédito são utilizados para determinar o comportamento estocástico do risco de crédito ou das variáveis inseridas no seu cálculo. Os modelos de risco de portfólio têm a função de estimar a distribuição de probabilidades das perdas ou do valor de uma carteira de crédito e, assim, produzir estimativas do risco presente. Os modelos de classificação de riscos, ou seja, de *credit scoring*, se dividem em duas categorias: modelos de aprovação de crédito e modelos de escoragem comportamental, também conhecidos como *behavioural scoring*. A principal diferença entre as duas categorias é que neste último, a instituição financeira analisa o comportamento do cliente em operações anteriores, enquanto que nos modelos de aprovação de crédito, a instituição financeira não conhece os aspectos comportamentais e as atividades do cliente na instituição (histórico de compras e pagamentos). Uma decisão importante a ser considerada para construção dos modelos de *credit scoring* é o tipo de pessoa, física ou jurídica, ao qual o modelo é empregado. Isto porque os atributos a serem considerados serão distintos, como mostram alguns exemplos de modelos de *credit scoring* no Quadro 1, aplicados em trabalhos correlatos.

	Autor(es)	Atributos Considerados	Técnica(s) Utilizada(s)	Acurácia Máxima
Pessoa Física	Albuquerque, Medina e Silva (2017)	Idade; Renda; Grau de Instrução; Tempo de Relacionamento; Prazo Contratado; Taxa SELIC; Coordenadas Geográficas.	Regressão Logística; Regressão Logística Geograficamente Ponderada	65%
	Gonçalves, Gouvêa e Mantovani (2013)	Gênero; Estado civil; Tempo no emprego; Tempo de residência fixa; Idade; Endereço e Telefone (comercial e residencial); Salário do cônjuge; Profissão; Número de parcelas; Valor da parcela sobre o salário; Valor do empréstimo; Tipo de crédito; Relação entre empréstimo; e Salário e aquisição do empréstimo.	Regressão Logística	70%
	Araújo e Carmona (2007)	Estado Civil; Gênero; Residência; Idade; Número de Dependentes; Receita, Despesa e Resultado Líquido (familiar e do Avalista); Valor total e número de Parcelas (último empréstimo); Percentual de Endividamento; Número de Créditos Anteriores; Histórico com a Instituição (titular e avalista); Agente de Crédito Responsável pelo Empréstimo	Regressão Logística	81%

	Amorim Neto e Carmona (2002)	Grau de escolaridade; Residência; Natureza da ocupação; Idade; Restrição de crédito; Financiamento de veículos; Cartão de crédito do banco.	Regressão Logística; e Análise Discriminante	81%
Pessoa Jurídica	Gevert et al. (2011)	Restrições (empresa e sócios); Tempo como cliente; Setor de atividade, Tempo de atividade; Número de funcionários; Imóvel; Principais clientes; Faturamento anual; conta em outro banco; Valor de bens (empresa e sócios); Seguro empresarial; Aplicações financeiras; Porcentagem de vendas a prazo; Experiência de crédito; Histórico da conta corrente; Sociedade de cônjuges.	Regressão Logística; Redes Neurais Artificiais; e *Support Vector Machine*.	99,05%
	Amaral Junior e Távora Junior (2010)	Desemprego; Índice de produção; Taxa SELIC; Taxa de compulsório; e *spread* bancário.	Redes Neurais Artificiais; e Regressão Linear	Coeficiente de determinação R^2
	Steiner et al. (2007)	Restrições (empresa e sócios); Tempo como cliente; Setor de Atividade; Tempo de Atividade; Número de Funcionários; Sede da Empresa; Bairro; Principais clientes; Faturamento Bruto Anual; Cliente em outro banco; Bens (empresa e sócios); Seguro Empresarial; Aplicações Financeiras; Vendas a Prazo; Experiência de Crédito; Histórico da Conta Corrente; Sociedade entre Cônjuges; Risco.	Regras de decisão e Redes Neurais Artificiais	89,57%

Quadro 1 - Características dos modelos construídos para pessoas física e jurídica.

Como pode-se notar, os modelos podem levar em consideração diferentes tipos e número de atributos, sendo que os atributos dos modelos de pessoa física consideram basicamente as informações de quem está solicitando o crédito como, por exemplo, idade, sexo e estado civil, enquanto que os modelos de pessoa jurídica consideram as informações da empresa e, eventualmente, de seus sócios como, por exemplo, se possuem ou não alguma restrição ao crédito. Além destes, outros atributos podem ser incluídos no modelo para previsão de inadimplência como, por exemplo, aqueles relacionados a Psicologia Econômica, a fim de determinar características psicológicas e comportamentais associadas ao grupo de indivíduos em situação de inadimplência (Rogers, Rogers e Securato, 2015).

O número de atributos considerados para construção dos modelos pode variar bastante. Dos trabalhos contidos na Tabela1, este número varia de cinco (Amaral Junior e Távora Junior, 2010) até em mais de 20 atributos (Steiner et al., 2007; Gevert et al, 2011; Araújo e Carmona, 2007). Vale ressaltar que, a complexidade de cada modelo aumenta em função do número de atributos considerados, mas em contrapartida, maior também será a capacidade do mesmo em melhor representar a realidade (no caso dos dados serem confiáveis). Por isso, tem-se um *trade-off* entre a complexidade e a representatividade do modelo, visando uma escolha adequada dos atributos.

Outra característica importante para a construção do modelo é o número de exemplos (ou instâncias) que, no caso do *credit scoring*, são os registros históricos de clientes, podendo este número variar de acordo com a disponibilidade dos mesmos. Em Araújo e Carmona, 2007, os pesquisadores trabalharam com 200 instâncias, já em Gonçalves, Gouvêa e Mantovani, 2013; e Albuquerque, Medina e Silva, 2017, com mais de 20.000 operações de crédito.

A classificação dos modelos de *credit scoring* geralmente se dá pela definição de clientes como “Adimplentes” ou “Inadimplentes”, sendo que o limiar que considera se um cliente é

inadimplente ou não pode variar de um estudo para o outro. Por exemplo, o cliente é considerado inadimplente se possui saldos com mais de 60 dias em atraso (Gonçalves, Gouvêa e Mantovani, 2013; Araújo e Carmona, 2007); ou com mais de 90 dias em atraso (Amaral Junior e Távora Junior, 2010; Albuquerque, Medina e Silva, 2017); ou então com atraso no pagamento de suas operações superior a 180 dias e inferior a 360 dias (Amorim Neto e Carmona, 2002).

As técnicas utilizadas para treinar e testar o modelo também podem variar, como mostrado na Tabela 1, e, consequentemente, o desempenho do modelo, medido geralmente pela acurácia, na tarefa de predição de clientes "adimplentes" e "inadimplentes" poderá ser diferente.

Todas estas diferenças mencionadas são devido à complexidade e tamanho da maioria dos problemas modernos, em que a melhor prática para se obter resultados satisfatórios, segundo Dean (2014), é utilizar diferentes técnicas de modelagem ou algoritmos, usando diferentes configurações ou parâmetros, de forma comparativa. Desta forma, a necessidade do desenvolvimento e da utilização de modelos e ferramentas mais sofisticadas, para a resolução de problemas cada vez mais complexos, é cada vez maior.

4. Metodologia

Para o problema analisado, realizou-se uma pesquisa para o levantamento de dados, cujas etapas resumem-se na coleta dos dados, na definição dos atributos, na aplicação da técnica de RB para classificação e, por fim, na avaliação do modelo preditor.

4.1 Descrição do Problema

As instituições financeiras têm importantes decisões a serem tomadas diariamente, dentre elas está a concessão de crédito para pessoas físicas e/ou jurídicas. Esta decisão deve ser feita com base em uma rigorosa avaliação, uma vez que envolve riscos que podem trazer sérias consequências para a empresa.

A avaliação de risco de crédito se refere à possibilidade de um devedor (empresa, organização ou pessoa) não poder cumprir suas obrigações de dívida com seus credores. Tal tipo problema é amplamente estudado na literatura, com o desenvolvimento e a utilização de técnicas de classificação, que visam atribuir as empresas em grupos, de acordo com o seu nível de risco de crédito (Doumpos e Zopounidis, 2002 e Zopounidis e Doumpos, 2002). Normalmente, são utilizados dois grupos: empresas para as quais o crédito deve ser concedido; e empresas para as quais o crédito deve ser rejeitado.

A avaliação do risco de crédito é, portanto, um procedimento que as instituições financeiras possuem para minimizar o potencial de risco e, assim, garantir que a decisão de conceder ou não crédito à uma pessoa não trará grandes consequências, no caso de um cenário negativo da pessoa não cumprir com as obrigações de pagamento.

Levando em consideração que as avaliações tradicionais de concessão de crédito se baseiam em análises subjetivas do potencial de pagamento, podendo ou não solicitar alguma forma de garantia de pagamento, as técnicas de Pesquisa Operacional, como já mencionado anteriormente, contribuem de maneira substancial no suporte à decisão da gerência.

4.2 Coleta de dados e definição dos atributos

O problema investigado, denominado como "Avaliação de Risco de Crédito de Pessoa Jurídica", considera uma instituição financeira, que conta com mais de 8 mil clientes (pessoa jurídica ou empresas) cadastrados. A referida instituição coloca à disposição de seus clientes linhas de crédito, sendo que para este tipo de transação a sua clientela é constituída, basicamente, empresas que se classificam como micro, pequena, média e empresas.

O banco de dados atualizado da grande instituição financeira nacional se refere à concessão de crédito para pessoas jurídicas (empresas) localizadas no Estado do Paraná (PR). Foram utilizados registros históricos do período de janeiro de 1996 a junho de 2017, com mais de

39 mil operações de créditos, realizadas por mais de 8 mil empresas. Atualmente, no banco de dados estão 6.494 clientes adimplentes, ou seja, que não possuem débito vencido, e 1.864 inadimplentes. Se somados, os débitos vencidos de todos os inadimplentes chegam ao total de R$ 1.679.831.282,76.

Para a construção do modelo para o problema, foram considerados todas as 8.358 instâncias (empresas) e oito atributos: Faturamento Bruto Anual; Quantidade de Funcionários; Tempo de Empresa; Tempo de Cliente do Banco; Risco; Restrição; Limite estabelecido; Endividamento Total. O Quadro 2 mostra a descrição de cada um dos atributos, bem como seu tipo e as estatísticas descritivas.

Atributo	Tipo	Descrição	Descrição Estatística
Faturamento	Intervalar	Representa o faturamento bruto da Pessoa Jurídica por ano	Intervalo: 0 a 769.665.000 Média: 3.923.404,78 Desvio Padrão: 14.229.457,21
Funcionários	Intervalar	Representa a quantidade de funcionários que a empresa possui	Intervalo: 0 a 5.113 Média: 13,515 Desvio Padrão: 77,56
Tempo de Empresa	Intervalar	Representa o tempo em anos que a empresa foi fundada até os dias atuais	Intervalo: 0 a 134 Média: 14,36 Desvio Padrão: 9,76
Tempo de Cliente	Intervalar	Representa o tempo em anos desde que a empresa é cliente da instituição financeira	Intervalo: 0 a 72 Média: 9,214 Desvio Padrão: 6,20
Restrição	Ordinal	Indica se a Pessoa Jurídica possui ou não algum tipo de restrição perante ao banco.	Frequência em cada classe: Restrição Impeditiva: 2.790 Restrição Fraca: 1.775 Sem Restrições: 3.793
Limite Estabelecido	Intervalar	Indica o limite de crédito que a instituição financeira disponibiliza para a pessoa Jurídica	Intervalo: 0 a 35.000.000 Média: 501.015,381 Desvio Padrão: 714.741,149
Endividamento	Intervalar	Demonstra quanto a Pessoa Jurídica está endividada	Intervalo: 201 a 355.563.956 Média: 468.577,441 Desvio Padrão: 4.686.694,62
Inadimplência	Nominal	Saída de Classificação da Rede, indicando se a Pessoa Jurídica está ou não adimplente com relação a seus débitos.	Frequência em cada classe: Adimplente: 6.494 Inadimplente: 1.864

Quadro 2 - Dados (atributos) do problema "Avaliação de Risco de Crédito de Pessoa Jurídica"

5. Obtenção e Discussão dos Resultados

A modelagem do problema com as RBs, considerando os atributos mencionados, resulta em um grafo com suas respectivas distribuições de probabilidades, obtidas com o treinamento a partir do banco de dados. O modelo de treinamento da rede foi construído utilizando o *software* WEKA® (*Waikato Environment for Knowledge Analysis*, disponível em: http://www.cs.waikato.ac.nz/ml/weka) e o teste foi aplicado com a opção de validação *cross-validation* 10-*folds*.

Para a avaliação do modelo foi considerada a métrica de acurácia, baseada na matriz de confusão, e medida pela Taxa de Acerto Total (TAT) como em (1).

$$Acurácia = TAT = (n_{aa} + n_{ii})/n_T \quad (1)$$

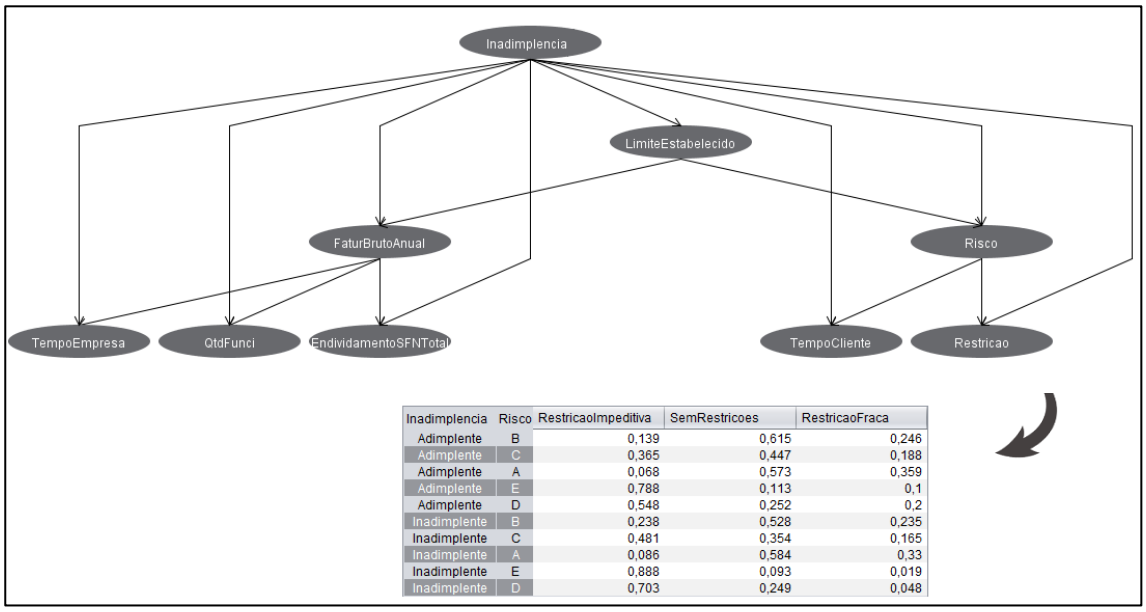
onde n_{aa} representa o número de instâncias adimplentes e corretamente classificados como "Adimplente"; n_{ii}, o número de instâncias inadimplentes e corretamente classificados como "Inadimplente"; n_T, o total de instâncias consideradas.

Os resultados assim obtidos utilizando as RBs estão apresentados na matriz de confusão do Quadro 3.

Classificado como →	Adimplente	Inadimplente
Adimplente	5.884	610
Inadimplente	635	1.229

Quadro 3 - Matriz de Confusão para o problema

Nessas configurações, a rede obteve um resultado de acurácia de 0,8510, ou seja, aproximadamente 85,10% das instâncias foram classificadas corretamente. A Figura 4 apresenta o grafo resultante, construído pelo *software* WEKA. As tabelas de distribuições de probabilidade foram ocultadas para melhor visualização da rede, ficando exposta somente a do atributo "Restrição" que, como pode ser observado, é dependente do atributo "Risco".



Inadimplencia	Risco	RestricaoImpeditiva	SemRestricoes	RestricaoFraca
Adimplente	B	0,139	0,615	0,246
Adimplente	C	0,365	0,447	0,188
Adimplente	A	0,068	0,573	0,359
Adimplente	E	0,788	0,113	0,1
Adimplente	D	0,548	0,252	0,2
Inadimplente	B	0,238	0,528	0,235
Inadimplente	C	0,481	0,354	0,165
Inadimplente	A	0,086	0,584	0,33
Inadimplente	E	0,888	0,093	0,019
Inadimplente	D	0,703	0,249	0,048

Figura 4 – Modelagem da RB para o problema analisado (Crédito).

Um segundo teste realizado com os dados, foi considerando uma diferenciação pelo tipo de Segmento de Pessoa Jurídica, ou seja, separando as instâncias que compreendem Microempresa (3.242 instâncias); Pequena Empresa (3.810 instâncias); Média Empresa (192 instâncias) e Empresa (1.036 instâncias), segundo uma classificação realizada pela própria instituição bancária. Os resultados obtidos, considerando esta configuração para o treinamento e teste das redes, estão dispostos no Quadro 4.

Microempresa	Classificado como →	Adimplente	Inadimplente
	Adimplente	2.395	232
	Inadimplente	257	358
Pequena Empresa	Classificado como →	Adimplente	Inadimplente
	Adimplente	2.499	294
	Inadimplente	258	759
Média Empresa	Classificado como →	Adimplente	Inadimplente
	Adimplente	166	6
	Inadimplente	12	8
Empresa	Classificado como →	Adimplente	Inadimplente

	Adimplente	749	75
	Inadimplente	69	143

Quadro 4 – Matrizes de confusão, discriminando as empresas por segmento (microempresa; pequena empresa; média empresa e empresa)

Neste segundo teste (com uma configuração considerando os tipos de Segmentos), a acurácia de cada rede ficou em 84,92% para Microempresa; 85,51% para Pequena Empresa; 90,62% para Média Empresa e 86,10% para Empresa. A separação dos dados conforme o Segmento de empresas, para o treinamento e teste das RB, resultou em valores de acurácia similares à configuração com todos os dados juntos, de aproximadamente 85% e, somente para o caso de Média Empresa, que o valor de acurácia teve um aumento, atingindo 90,62%, porém isso se deve ao fato deste segmento ter um desbalanceamento na classe de adimplente/inadimplente.

6. Considerações Finais

O presente artigo que objetiva discriminar clientes pessoa jurídica em adimplentes ou inadimplentes, apresentou resultados com desempenho razoável quanto a acurácia da técnica de aprendizagem (RBs), em torno de 85%.

Ressalta-se também que melhores resultados poderiam ser, provavelmente, obtidos mudando a configuração da rede, ou seja, construindo por conta própria e com o auxílio de especialista, os nós e arcos da rede bayesiana. Desta forma, ter-se-ia condições de determinar quais atributos possuem maior dependência um com o outro e, com esta configuração de rede de dependências, treinar os dados (instâncias) para obter os valores de probabilidades condicionais.

Além disso, outras estratégias poderiam ser adotadas para alavancar o poder de acurácia da rede, como: a) introduzir outros atributos que caracterizem melhor as empresas à classe adimplente/inadimplente; b) utilizar outro algoritmo de aprendizagem supervisionada, na expectativa de obter resultados com melhores desempenhos; c) utilizar técnicas de balanceamento de classes, entre outras.

Referências

ABU-MUSTAFA, Y. (2016). Machine Learning. California Institute of Technology, Caltech. Online Course. Disponível em: <http://www.work.caltech.edu/>. Acesso em 23 nov. 2017.

ALBUQUERQUE, P. H. M., MEDINA, F. A. S.; DA SILVA, A. R. (2017). Regressão Logística Geograficamente Ponderada Aplicada a Modelos de Credit Scoring. Revista Contabilidade & Finanças, 28(73), 93-112.

ALPAYDIN, E. (2010) Introduction to Machine Learning. MIT press.

AMARAL JÚNIOR, J. B.; TÁVORA JÚNIOR, J. L. (2010). Uma análise do uso de redes neurais para a avaliação do risco de crédito de empresas. Revista do BNDES, Rio de Janeiro, (34), 133-180.

AMORIN NETO, A. A.; CARMONA, C. U. D. M. (2002). Modelagem do risco de crédito: um estudo do segmento de pessoas físicas em um banco de varejo. Revista Eletrônica de Administração, 10(4).

ARAÚJO, E. A.; CARMONA, C. U. D. (2007). Desenvolvimento de modelos Credit Scoring com abordagem de regressão logística para a gestão da inadimplência de uma instituição de microcrédito. Contabilidade Vista & Revista, 18(3).

DEAN, J. (2014). Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. John Wiley & Sons.

DOUMPOS, M.; KOSMIDOU, K.; BAOURAKIS, G.; ZOPOUNIDIS, C. (2002). Credit risk assessment using a multicriteria hierarchical discrimination approach: A comparative analysis. European Journal of Operational Research, 138(2), 392-412.

DOUMPOS, M.; ZOPOUNIDIS, C. (2002) Multicriteria decision aid classification methods. Springer Science & Business Media.

FRASCAROLI, B. F.; PAES, N. L.; RAMOS, F. D. S. (2010). A indústria brasileira e o racionamento de crédito: uma análise do comportamento dos bancos sob informações assimétricas. Revista Economia, 11(2), 403-433.

GEVERT, V. G.; DA SILVA, A. C. L.; GEVERT, F.; ALES, V. T. (2011). Modelos de regressão logística, redes neurais e support vector machine (svm s) na análise de crédito a pessoas jurídicas. RECEN-Revista Ciências Exatas e Naturais, 12(2), 269-293.

GONÇALVES, E. B.; GOUVÊA, M. A.; MANTOVANI, D. M. N. (2013). Análise de risco de crédito com o uso de regressão logística. Revista Contemporânea de Contabilidade, 10(20), 139-160.

HILLIER, F. S.; LIEBERMAN, G. J. (2013). Introdução à pesquisa operacional. Porto Alegre: McGraw Hill Brasil.

INFORMS. What is Operations Research?. Disponível em: < https://www.informs.org/About-INFORMS/What-is-Operations-Research>. Acesso em: 23 nov. 2017.

KOLLER, D.; FRIEDMAN, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.

LACHTERMACHER, G. (2016). Pesquisa operacional na tomada de decisões. 5. ed. Rio de Janeiro: LTC.

MEDINA, R. P.; SELVA, M. L. (2013). Análisis del credit scoring. RAE-Revista de Administração de Empresas, 53(3).

RODRIGUES CAPELLETTO, L.; CORRAR, L. J. (2008). Índices de risco sistêmico para o setor bancário. Revista Contabilidade & Finanças-USP, 19(47).

ROGERS, P.; ROGERS, D.; SECURATO, J. R. (2015). Sobre variáveis psicológicas em modelos de application scoring. RAE-Revista de Administração de Empresas, 55(1), 38-49.

RUSSELL, B.; BINDER, J.; KOLLER, D.; KANAZAWA, K. (1995) Local learning in probabilistic networks with hidden variables. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1146–1152. San Francisco: Morgan Kaufmann.

RUSSELL, S.; NORVIG, P. (2004) Inteligência artificial. Elsevier.

SAMEJIMA, K.; DOYA, K.; KAWATO, M. (2003). Inter-module credit assignment in modular reinforcement learning. Neural Networks, 16(7), 985-994.

SILVA BRITO, G. A.; ASSAF NETO, A. (2008). Modelo de classificação de risco de crédito de empresas. Revista Contabilidade & Finanças-USP, 19(46).

STEINER, M. T. A.; NIEVOLA, J. C.; SOMA, N. Y.; SHIMIZU, T.; STEINER NETO, P. J. (2007). Extração de regras de classificação a partir de redes neurais para auxílio à tomada de decisão na concessão de crédito bancário. Pesquisa Operacional, 27(3), 407-426.

ZOPOUNIDIS, C.; DOUMPOS, M. (2002). Multi-group discrimination using multi-criteria analysis: Illustrations from the field of finance. European Journal of Operational Research, v. 139, n. 2, p. 371-389.