PRODUÇÃO ONLINE
REVISTA CIENTÍFICA ELETRÔNICA DE ENGENHARIA DE PRODUÇÃO
ISSN 1676-1901

REDES NEURAIS ARTIFICIAIS PARA CLASSIFICAÇÃO DE RISCO DE CRÉDITO

ARTIFICIAL NEURAL NETWORKS FOR CREDIT RISK CLASSIFICATION

Annanda Martins Silveira* E-mail: annandams@gmail.com
Mariana Kleina* E-mail: marianakleina11@gmail.com
*Universidade Federal do Paraná (UFPR), Curitiba, PR

Resumo: A atividade de concessão de empréstimos visa o lucro obtido por meio de juros sobre o montante do empréstimo. Esta atividade é predominantemente realizada por grandes instituições financeiras, como os bancos. Porém, existem plataformas *online* em que é possível obter ou oferecer empréstimos entre apenas as pessoas interessadas, o *peer-to-peer lending*. Entretanto, esta atividade oferece riscos ao agente concessor e portanto, deve ser bem analisada. Para diminuir os riscos de concessão de crédito, existem diversos métodos de avaliação e previsão de comportamento de variáveis. Este trabalho tem o objetivo de propor dois métodos de previsão de risco de crédito e classificar os potenciais mutuários em três classes: adimplentes (que pagam o empréstimo), temporariamente inadimplentes (que pagam o empréstimo com atraso) e inadimplentes (que não pagam o empréstimo). Os métodos utilizados são redes neurais artificiais com algoritmo de aprendizado supervisionado *backpropagation* e regressão linear múltipla. O primeiro obteve um desempenho de 60,8% de assertividade em suas classificações, enquanto o segundo, 49,7%. Apesar de as redes neurais artificiais apresentarem melhores resultados, ambos métodos apresentam resultados insatisfatórios para a aplicação proposta, visto que um grande número de potenciais mutuários de crédito teve uma previsão incorreta do seu comportamento.

Palavras-chave: Redes neurais artificiais. Regressão linear múltipla. Risco de crédito. *Peer-to-peer lending.*

Abstract: The credit granting activity aims the profit obtained by the interests applied on the loan amount. This activity is predominantly carried out by major financial institutions, such as banks. However, there are online platforms in which it is possible to get or offer loans between the interested people, the peer-to-peer lending platforms. However, this activity entails risks to the lender and, therefore, must be well analysed. To reduce credit risk, there are several methods for evaluating and forecasting the behavior variables. This paper aims to propose two credit risk forecasting methods and classify the potential borrowers in three classes: payers (who repay the loan), temporary defaulters (who repay the loan lately) and defaulters (who do not repay the loan). The methods applied are artificial neural networks with backpropagation supervised learning algorithm and multiple linear regression. The former achieved a performance of 60.8% assertiveness in its classifications, while the latter achieved 49.7%. Although the artificial neural networks performed better, both methods present unsatisfactory results for the proposed application, since a large number of potential borrowers had an incorrect forecasting of their behavior.

Keywords: Artificial neural networks. Multiple linear regression. Credit risk. Peer-to-peer lending.

1 INTRODUÇÃO

Em 1970, a revolução industrial substituiu a força física dos seres humanos pela força desempenhada por máquinas. Hoje, a inteligência artificial (IA) tem o propósito de substituir atividades de inteligência desempenhada pelos homens pela inteligência das máquinas e programas computacionais (MUNAKATA, 2008).

Dentre as atividades de inteligência, a IA tem aplicações em diversas áreas, como em problemas de classificação e predição, raciocínio sobre informações incertas ou incompletas e dedução baseada em conhecimento. Para tanto, os programas e computadores operam de maneira a reproduzir como o cérebro processa informações (MUNAKATA, 2008).

Neste contexto, a IA busca a execução de tarefas que necessitem de inteligência, porém, a partir de programas e algoritmos que simulem o cérebro. Dentre as aplicações da IA citadas, a dedução baseada em conhecimento tem grande aplicabilidade em sistemas de tomada de decisão, pois é necessário o raciocínio ágil sobre o comportamento, muitas vezes incerto, de diversas variáveis. Um exemplo disso é a decisão em investimentos e concessões de crédito.

Todavia, existem outros métodos que podem ser aplicados como ferramenta nas decisões de concessões de crédito. Segundo Draper e Smith (2014), é possível descrever as relações entre as variáveis de um sistema em uma equação matemática, que pode ser extremamente útil para prever o comportamento de uma variável resposta.

Um método de obter tal equação matemática é o de regressão linear múltipla (RLM), que recebe esse nome por abrager duas ou mais variáveis independentes. Desta forma, a RLM, assim como a RNA, pode ser utilizada na previsão de comportamento de potencias clientes de crédito.

Dentro do cenário de concessão de crédito, os bancos são os tradicionais intermediários do sistema financeiro, ou seja, a ponte entre os investidores e tomadores de empréstimos. Com o colapso do sistema financeiro que teve início em 2008, a confiança pública depositada nos bancos foi abalada. Surgiu, então, a ideia

de “desintermediação” como uma alternativa para se obter crédito sem a participação dos bancos (MATEESCU, 2015).

Dessa forma, o *peer-to-peer lending* surgiu como um sistema sem intermediação financeira de maneira relativamente simples, em que os interessados em empréstimos ou investimentos poderiam realizar as transações online e entre si. Além disso, este novo método veio com a ideia de que os empréstimos seriam mais rápidos em razão da menor burocracia, e, além de outras características, seria mais transparente, uma vez que os históricos de empréstimos seriam disponibilizados ao público (MATEESCU, 2015).

No contexto de utilização de IA e modelos de regressão para tomadas de decisão no meio financeiro, este trabalho tem como objetivo a classificação dos clientes de uma grande empresa de *peer-to-peer lending*, a fim de auxiliar na decisão de concessão de crédito. Para isso, foram utilizadas as redes neurais artificiais do tipo *perceptron* de multi camadas e o método de RLM.

Os dados dos clientes foram disponibilizados pelo site da empresa e fornecem os atributos de entrada de ambos os métodos. Assim, os modelos classificaram os clientes de acordo com as classes propostas por Assef (2018): adimplentes, inadimplentes ou temporariamente inadimplentes.

2 REVISÃO DE LITERATURA

2.1 Risco de crédito

De acordo com Centa (2005), a concessão de crédito se dá quando uma pessoa física ou jurídica cede uma parte do seu patrimônio a um terceiro, com a expectativa de recebê-lo de volta em um prazo estipulado. Dessa forma, a concessão de crédito envolve uma decisão a ser tomada pelo dono do patrimônio.

Dentro do cenário econômico financeiro, existem agentes com excesso de recursos monetários – os agentes superavitários – e agentes com falta de recursos monetários – os agentes deficitários. Com o intuito de canalizar os recursos entre estes agentes, os intermediários financeiros recebem dinheiro dos agentes

superavitários, remunerando-os, e emprestam recursos para os agentes deficitários, cobrando-os juros. Os agentes intermediários visam o lucro nesta atividade de intermediação, porém, ela envolve riscos (LEMES Jr., RIGO e CHEROBIM, 2010).

Segundo Hoji (2010), o risco é a incerteza de ocorrer aquilo que foi planejado. No caso de intermediação financeira, o risco se dá pela probabilidade de os tomadores de recursos não pagarem o valor acordado no ato do empréstimo. Dessa forma, os aplicadores cobram taxas de juros maiores para mutuários com maiores riscos, ou seja, com maiores incertezas de retorno (LEMES Jr., RIGO e CHEROBIM, 2010).

Portanto, os analistas de concessão de crédito devem ser criteriosos na avaliação dos riscos e fundamentar a decisão de conceder ou não o crédito em informações técnicas, a fim de diminuir ao máximo o número de empréstimos inadimplentes (SEHN, CARLINI Jr., 2007).

A atividade de avaliação de riscos na concessão de crédito é conhecida como política de gestão de risco de crédito. Esta política tem como objetivo a administração do capital, fazendo com que sejam aproveitadas as oportunidades de lucro da organização e, ao mesmo tempo, não sejam extrapolados os limites de riscos estabelecidos pela organização concessora de crédito (AMARAL Jr.; TÁVORA Jr., 2010).

De acordo com Centa (2005), para que uma análise de crédito seja consistente, é importante que a quantificação dos riscos seja bem avaliada e as conclusões e recomendações devem ser feitas de maneira prática e viável. Segundo o autor, a análise de crédito é, portanto, o ponto inicial no momento de decisão de concessão de crédito.

Para Zhang, Tadikamalla e Shang (2016), a análise de crédito pode ser executada a partir de cinco métodos. São eles: método estatístico, método de decisão de teoria, redes neurais artificiais (RNA), *support vector machines* (SVM) e análise por envoltória de dados. Neste trabalho, foram utilizadas as redes neurais artificiais para a análise de concessão de crédito. Além disso, é proposto um método de Regressão Linear Múltipla para comparação dos resultados.

2.2 Redes neurais artificiais

As redes neurais artificiais tiveram seus princípios de funcionamento motivados pela percepção de que o cérebro humano processa informações de uma maneira totalmente diferente dos computadores convencionais. O cérebro é altamente complexo, possui aproximadamente 10^{11} unidades chamadas neurônios. Estes neurônios são interconectados, formando aproximadamente 10^{15} conexões (MUNAKATA, 2008; HAYKIN, 2009).

Similarmente ao cérebro biológico, as redes neurais artificiais são compostas de neurônios artificiais e interconexões (MUKANATA, 2008). Para Schmidhuber (2014), as RNA são constituídas de simples processadores conectados, os neurônios, cada qual produzindo uma sequência de ativações de valores reais. Analisando uma representação de RNA, os neurônios são representados por nós ou vértices, enquanto as interconexões por arestas (MUKANATA, 2008).

2.2.1 Neurônio

De acordo com Módolo (2016), o neurônio é a unidade fundamental de processamento do cérebro. As principais partes que compõem um neurônio biológico são:

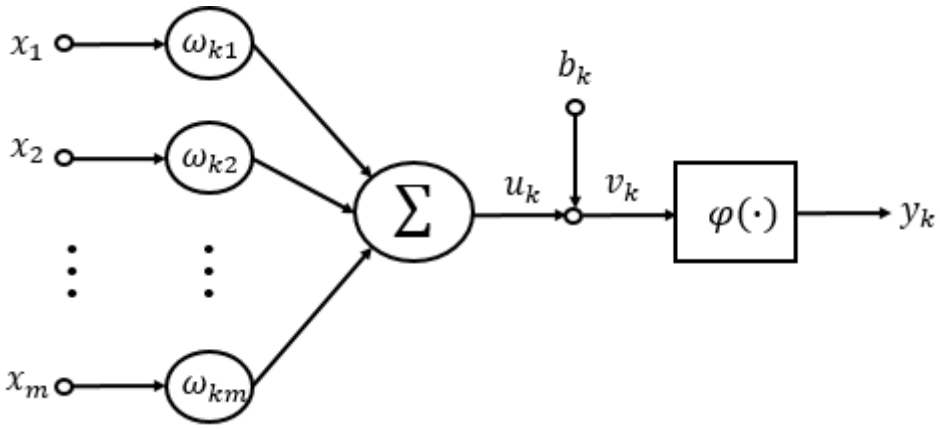
- Dendritos: responsáveis pela captação de informações;
- Corpo celular: responsável pelo processamento das informações;
- Axônio: responsável pela transmissão de informações até a extremidade do neurônio;
- Sinapses: transmitem as informações para outras células.

2.2.2 Modelos de redes neurais artificias

McCulloch e Pitts (1943) foram os primeiros a apresentar um modelo de redes neurais, de acordo com Másson e Wang (1990). O modelo criado por eles ficou conhecido como neurônio MCP, e foi a base para o desenvolvimento dos modelos

seguintes, como o modelo de neurônio não-linear que será utilizado neste trabalho (ASSEF, 2018; MÁSSON e WANG, 1990). O neurônio não-linear mais empregado na atualidade pela comunidade de RNA é representado na Figura 1 (MÓDOLO, 2016).

Figura 1 – Representação de um neurônio não-linear



Fonte: Adaptado de Haykin (2009)

As variáveis $x_1, x_2, \ldots, x_m$ são as entradas do neurônio. Em um neurônio biológico, essas entradas correspondem aos níveis de estimulação. Cada entrada é multiplicada pelo peso $w_1, w_2, \ldots, w_{km}$ do neurônio k. Em neurônios biológicos, os pesos correspondem às forças de sinapse (MUNAKATA, 2008). A variável u_k é a saída do combinador linear e pode ser calculada pela Equação 1.

$$u_k = \sum_{j=1}^{m} w_{kj}\, x_j \tag{1}$$

A variável b_k é o *bias*, $\varphi(\cdot)$ é a função de ativação e y_k é a saída do neurônio, que pode ser calculada pela Equação 2.

$$y_k = \varphi(u_k + b_k) \tag{2}$$

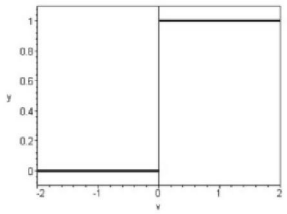
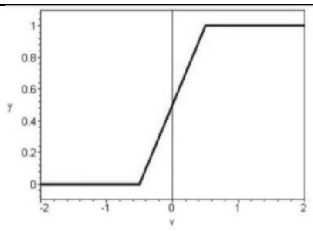
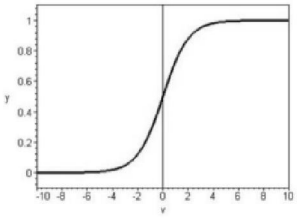
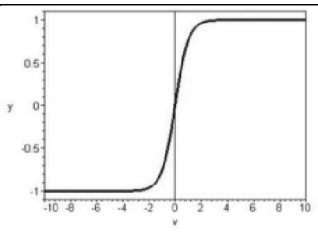
O *bias* é uma entrada externa adicionada à função de ativação. Ela tem a finalidade de aumentar ou diminuir a entrada líquida da função de ativação, servindo como uma transformação afim na função, como apresentado na Equação 3.

$$v_k = u_k + b_k \tag{3}$$

Em comparação com o neurônio MCP, o neurônio não-linear possui os pesos (w_{kj}) ajustáveis, a função de ativação $\varphi(u_k)$ pode ser não-linear e é possível que o *bias* b_k seja acrescido ao somatório u_k. Enquanto isso, o neurônio MCP possui apenas pesos fixos, a função de ativação é linear e não existe a adição de *bias* (MÓDOLO, 2016; REZENDE, 2003).

Sobre as funções de ativação, diferentes tipos podem ser utilizadas, de acordo com as características das aplicações (MUNAKATA, 2008). No Quadro 1 são apresentados alguns exemplos de funções de ativação descritos por Haykin (2009) e Quiles (2004).

Quadro 1 – Funções de ativação

Limiar	$\varphi(v) = \begin{cases} 1 \; se \; v \geq 0 \\ 0 \; se \; v < 0 \end{cases}$	
Linear por partes	$\varphi(v) = \begin{cases} 1 \; se \; v \geq \frac{1}{2} \\ v \; se - \frac{1}{2} < v < \frac{1}{2} \\ 0 \; se \; v \leq -\frac{1}{2} \end{cases}$	
Logística	$\varphi(v) = \frac{1}{1 + \exp(-av)}$	
Tangente hiperbólica	$\varphi(v) = tangh(v)$	

Fonte: Adaptado de Haykin (2009); Quiles (2004)

2.2.3 Arquitetura

De acordo com Munakata (2008), os padrões de conexões entre os neurônios formam a arquitetura da RNA, também conhecida como sua estrutura. As arquiteturas das redes neurais podem ser classificadas em dois grandes grupos: estáticas ou não recorrentes e dinâmicas ou recorrentes (GRZEIDAK, 2016).

No primeiro grupo a informação percorre uma única direção – da entrada para a saída – não havendo nenhum *loop* na rede. No segundo grupo, entretanto, a informação percorre dois caminhos, tendo ao menos um segundo fornecimento de dados na rede (GRZEIDAK, 2016).

As RNA recorrentes, por armazenarem informações temporais, são bem empregadas para tarefas em que memória associativa é necessária, como séries temporais e tarefas sequenciais. Redes não recorrentes, por sua vez, são adequadas para problemas de mapeamento, em que é possível analisar como as variáveis de entrada afetam as saídas (GRZEIDAK, 2016).

Existem diversas arquiteturas de redes dentro da classe de não recorrentes, como *perceptron* de múltiplas camadas (MLP), funções de bases radiais (RBF) e redes difusas. Neste trabalho, será utilizado a MLP, o modelo de rede mais estudado e empregado atualmente (GRZEIDAK, 2016).

2.2.4 Processos de aprendizagem

A maneira como os neurônios são arranjados, ou seja, a arquitetura escolhida para a RNA, está estreitamente relacionada com o algoritmo utilizado para treiná-la (HAYKIN, 2009). O treinamento (ou aprendizado) da rede, por conseguinte, é o processo em que os parâmetros (ou pesos) da rede são ajustados por meio de mecanismos de estímulos externos – as entradas (CASTRO, 2006).

As redes neurais podem ser treinadas a partir de três tipos de algoritmos: aprendizagem supervisionada, não-supervisionada e de reforço. No aprendizado supervisionado, os pesos são ajustados a partir da comparação entre a entrada e a saída correspondente. Neste caso, portanto, é necessário que um supervisor forneça à rede a saída desejada. Existem diversos métodos para este tipo de aprendizado, dentre eles pode-se citar o método dos mínimos quadrados, *backpropagation* e matriz Hessiana (GRZEIDAK, 2016).

No aprendizado não-supervisionado, no entanto, não existe um supervisor e, como consequência, não é dada a saída desejada. Neste caso, a rede se adapta às regularidades estatísticas das entradas e gera classes automaticamente (CASTRO, 2006).

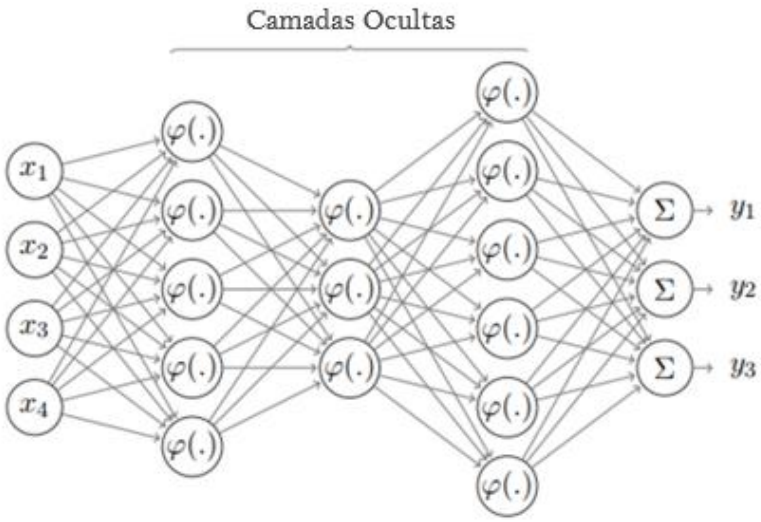
Por fim, no aprendizado por reforço, a rede não recebe as saídas desejadas como no aprendizado supervisionado, mas recebe uma avaliação da performance da rede, por meio de recompensas ou penalidades. O algoritmo neste tipo de

aprendizado tem como objetivo maximizar as saídas com recompensas (CASTRO, 2006). Neste trabalho, será utilizado o algoritmo de aprendizado supervisionado *backpropagation*, detalhado na seção 2.2.5.

2.2.5 *Perceptron* de Múltiplas Camadas

O *perceptron* de múltiplas camadas é caracterizado por possuir uma ou mais camadas ocultas entre a camada de entrada e a de saída. A camada de entrada tem o papel de fornecer as informações para a camada oculta, e a saída da camada oculta corresponde à entrada da próxima camada (GONÇALVES, 2005). Na Figura 2 é possível observar a representação gráfica da rede neural MLP com camadas ocultas.

Figura 2 – Representação de uma rede MLP



Fonte: Adaptado de Grzeidak (2016)

A Figura 2 apresenta as camadas de entrada representadas pelos quatro neurônios $x_1, x_2, x_{3,} x_4$. A rede é seguida por três camadas ocultas com cinco, três e seis neurônios cada uma, respectivamente. A última camada é composta por três

neurônios e é responsável por gerar as saídas da rede. Neste caso, três saídas possíveis são: y_1, y_2, y_3.

Para o treinamento da rede MLP, será utilizado o método de *backpropagation* com três camadas (entrada, oculta e saída). Este método é composto pelas etapas de propagação *foward* e propagação *backward*. A seguir será apresentado o algoritmo da rede, conforme descrito por Assef (2018).

Propagação *Forward*

1 Inicialização

Defina aleatoriamente valores entre [-1,1] para os pesos sinápticos da camada oculta (w_{kj} e b_k) e para a camada de saída (w_{lj} e b_l).

2 Ativação – camada oculta

Efetue o somatório apresentado na Equação 4 para cada neurônio k da camada oculta, no qual n é a iteração atual do treinamento.

$$v_{k(n)} = \sum_{j=1}^{m} w_{kj(n)}.x_{j(n)} + b_{k(n)} \quad (4)$$

Após, calcule o valor da função de ativação apresentado na Equação 5 para cada um dos neurônios da camada oculta, no qual $\varphi(.)$ é uma função logística.

$$y_{k(n)} = \varphi\left(v_{k(n)}\right) \quad (5)$$

3 Ativação – camada de saída

Efetue o somatório apresentado na Equação 6 para cada neurônio l da camada de saída, em que y_k é entrada da camada de saída.

$$v_{l(n)} = \sum_{j=1}^{m} w_{lj(n)}.y_{k(n)} + b_{\mathrm{l(n)}} \quad (6)$$

Em seguida, calcule o valor da função de ativação apresentado na Equação 7 para cada neurônio presente na camada de saída.

$$y_l(n) = \varphi\left(v_{l(n)}\right) \tag{7}$$

Propagação *Backward*

1 Ajuste dos pesos – camada de saída para camada oculta

Primeiramente, calcule o erro da camada de saída a partir da Equação 8, no qual dj é o valor esperado de saída e $\varphi'(.)$ é a derivada da função de ativação do neurônio de saída.

$$\delta_{l(n)} = \left(d_j - y_l\right).\varphi_l'\left(u_{l(n)}\right) \tag{8}$$

A seguir, calcule o ajuste dos pesos pela Equação 9 e atualize os pesos sinápticos da camada de saída pela Equação 10, em que η é a taxa de aprendizagem e α é a taxa de *momentum*.

$$\Delta\left(w_{lj}\right) = a\left(w_{lj(n-1)}\right) + \eta\delta_{l(n)}y_{l(n)} \tag{9}$$

$$w_{lj(n+1)} = w_{lj(n)} + \Delta w_{lj} \tag{10}$$

2 Ajuste de pesos – camada oculta para a camada de entrada

Inicialmente, calcule o erro da camada oculta a partir da Equação 11.

$$\delta_{k(n)} = \varphi'\left(y_{k(n)}\right)\sum_{l}\delta_{l(n)}w_{lj(n)} \tag{11}$$

E, de forma análoga ao passo 1, calcule o ajuste de pesos pela Equação 12, a fim de atualizá-los a partir da Equação 13.

$$\Delta(w_{kj}) = a(w_{kj(n-1)}) + \eta\delta_{k(n)}y_{k(n)} \quad (12)$$

$$w_{kj(n+1)} = w_{kj(n)} + \Delta w_{kj} \quad (13)$$

3 Cálculo do erro global

Conclui-se, então, a primeira iteração do algoritmo de aprendizado. A seguir, calcule o erro global a partir da Equação 14, no qual N representa o número de elementos de treinamento.

$$\varepsilon(n) = \frac{1}{N}\sum_{l=1}^{N} e_l^2(n) \quad (14)$$

2.3 Regressão linear múltipla

A regressão linear múltipla pode ser entendida como uma extensão da regressão linear simples. Ela permite compreender qual a influência que múltiplas variáveis independentes exercem sobre uma variável resposta dependente (NATHANS, OSWALD, NIMON, 2012).

Pode-se citar como outro objetivo da RLM a previsão de variáveis dependentes. De acordo com Gazola (2002), a RLM tem como finalidade o desenvolvimento de um modelo matemático linear que permita a estimativa de respostas de determinada variável, levando em consideração as múltiplas variáveis explicativas (variáveis independentes).

Como um modelo que tem sido bem estudado nas últimas décadas, a RLM tem muitas aplicações, como previsões metereológicas, avaliação na relação de variáveis mercadológicas com preços de bens de consumo e imobiliários, avaliação de solos, entre outros (SU-FEN, 2013; BARRETO, BATISTELA e GAIOTTO, 2016; PRUNZEL *et al.,* 2016; COUTINHO, SILVA e DELGADO, 2016; YILMAZ e KAYNAR, 2010).

De acordo com Barreto, Batistela e Gaiotto (2016), o modelo estatístico de uma RLM com k variáveis explicativas e n observações é descrito conforme a Equação 15.

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + \varepsilon_j \quad (15)$$

Y_j é a variável resposta, para $j = 1,2, \ldots, n$;

X_{ij} é a variável explicativa, para $i = 0,1, \ldots, k$;

β_i é o parâmetro do modelo (ou coeficientes de regressão) a serem estimados para cada X_{ij};

ε_j é o erro aleatório do modelo, pressuposto independente e normalmente distribuído com variância σ^2 e média zero.

O parâmetro β_i indica a variação na resposta Y_j por unidade de variação em X_{ij}, quando forem mantidas constantes todas as demais variáveis explicativas (MONTGOMERY e RUNGER, 2008).

3 METODOLOGIA

Este trabalho iniciou-se com o levantamento de dados históricos de clientes de crédito. Então, estes dados foram selecionados e pré-processados. Para utilização do método de classificação de RNA, foram primeiramente definidos, a partir da revisão de literatura, o tipo de arquitetura da rede, o algoritmo utilizado e a função de ativação.

A partir disto, a RNA recebeu os dados históricos dos clientes para ser treinada e testada, o que permitiu escolher de forma empírica os demais parâmetros que definiram sua arquitetura final.

A RLM utilizou a mesma amostra como dados de entrada do modelo, e, seus parâmetros, como pesos e o erro aleatório, foram calculados pelo próprio modelo.

Por fim, os dois métodos foram testados a partir da amostra de testes (1/3 da amostra total) e então avaliados por meio da matriz de confusão e cálculo de assertividade. A partir disto, foi possível comparar os resultados e avaliar o

desempenho dos métodos para a aplicação proposta. A seguir será apresentado cada passo detalhamente.

3.1 DADOS

Primeiramente, foram coletados dados históricos de clientes mutuários de crédito de uma empresa de *peer to peer lending*. Como o dinheiro concedido aos tomadores de crédito provém de investidores individuais, as informações sobre os mutuários são disponíveis no site da empresa (LENDING CLUB, 2018). Assim, os investidores podem avaliar e decidir entre emprestar dinheiro ou não a determinada pessoa.

A partir da coleta dos dados, fez-se uma seleção da amostra a ser estudada, além de um pré-processamento dos dados. A seguir serão detalhadas essas duas etapas.

3.1.1 Definições da amostra de dados

Os dados disponibilizados contêm mais de cinquenta variáveis, de mais de oitocentos mil clientes. Para este trabalho, foram selecionadas dezessete variáveis consideradas mais relevantes de uma amostra aleatória de 1800 empréstimos diferentes. No Quadro 2 são apresentadas as variáveis utilizadas e seus respectivos conceitos.

As variáveis selecionadas são, com excessão do "*status* do empréstimo", as entradas na primeira camada da RNA e as variáveis explicativas na RLM. O "*status* do empréstimo", por sua vez, corresponde à saída desejada e a variável resposta para a RNA e RLM, respectivamente.

Quadro 2 – Variáveis históricas escolhidas

Variáveis	**Conceitos**
Quantia do empréstimo	Quantia total destinada ao empréstimo até o momento.
Prazo	O número de pagamentos do empréstimo. Valores estão em meses e podem ser 36 ou 60.
Taxa de juros	Taxa de juros cobrada no empréstimo. Valores em %.
Prestação	A parcela a ser paga mensalmente.
Nota	Nota do possível empréstimo atribuída pela empresa. Valores de A a G, em que quanto mais próximo de A, melhor, e quanto mais próximo de G, pior.
Subnota	Subnota do possível empréstimo atribuída pela empresa. Valores de 1 a 5, em que quanto mais próximo de 1, melhor, e quanto mais próximo de 5, pior.
Duração do emprego	Duração do emprego do mutuário. Valores variam de 0 a 10 (números inteiros), no qual 0 significa menos de 1 ano e 10 significa mais de 10 anos.
Propriedade domiciliar	*Status* de propriedade domiciliar fornecido pelo mutuário. Valores possíveis são: aluguel, casa própria e hipoteca.
Renda anual	Renda anual fornecida pelo mutuário no momento de registro.
Status do empréstimo	*Status* do empréstimo. Valores podem ser: adimplentes, inadimplentes ou temporariamente inadimplentes.
Finalidade	A finalidade para qual o empréstimo será utilizado, com 13 opções de respostas fechadas.
Inadimplência 2 anos	O número de vezes em que o mutuário foi inadimplente por mais de 30 dias nos últimos 2 anos.
Linha de crédito mais antiga	O mês da mais antiga linha de crédito aberta pelo mutuário.
Última inadimplência	O número de meses desde a última inadimplência do mutuário.
Saldo rotativo	Total de crédito rotativo disponível.
Crédito rotativo utilizado	Quantia de crédito rotativo utilizada em relação ao total de crédito rotativo disponível.
Limite crédito rotativo	Limite de crédito rotativo total.

Fonte: Adaptado de Lending Club (2018).

3.1.2 Pré-processamento de dados

Para aplicação dos dados na RNA e RLM, é recomendado realizar a normalização dos dados, a fim de se trabalhar com dados mais consistentes e

homogêneos (ROBINSON e OSHLACK, 2010). A normalização aplicada converte os valores da variável para valores entre 0 e 1, a partir da Equação 16.

$$y = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{16}$$

Sendo y a variável normalizada e x a variável original.

A aplicação da Equação 16 é possível para todas as variáveis quantitativas. Porém, para as variáveis qualitativas, fez-se uma escala entre as variáveis, atribuindo valores entre 0 e 1. No quadro 3 é apresentado um exemplo de normalização de dados para a variável “Propriedade domiciliar”.

Quadro 3 – Normalização de variáveis qualitativas

Valor qualitativo	**Valor quantitativo atribuído**
Casa própria	1
Aluguel	0,5
Hipoteca	0

Fonte: A autora (2018).

A variável “linha de crédito mais antiga”, por ser uma data, foi primeiramente transformada em números de série sequenciais. Por definição, 1 de janeiro de 1990 é o número 01 da série. Logo, o dia 02/01/1990 é o número 02 da série, e assim sucessivamente. Após a transformação da variável em questão, aplicou-se, então, a Equação 16 para a normalização dos dados.

3.2 Parâmetros da RNA

O primeiro passo para a construção de uma RNA é definir qual o tipo de arquitetura que ela terá. A partir da revisão de literatura, determinou-se que a arquitetura será *perceptron* de multi camadas. Esta arquitetura permite, além da camada de entrada e saída da rede, camadas ocultas ou intermediárias.

Estudos comprovam que os *perceptrons* de multi camadas são aproximadores universais, e as redes MLP com apenas uma camada oculta são suficientes para aproximar qualquer função contínua, dado o número necessário de neurônios (CYBENKO, 1989; HORNIK, STINCHCOMBE e WHITE, 1989). Portanto, a RNA aqui proposta apresenta apenas uma camada oculta, além das camadas de entrada e de saída.

Os dados do Quadro 2, com exceção de "*status* do empréstimo", são os dados de entrada da RNA. Ou seja, como foram consideradas dezesseis variáveis na rede, ela foi composta por uma camada de entrada com dezesseis neurônios, cada qual correspondendo a uma das variáveis supracitadas.

Na camada seguinte, denominada camada oculta ou intermediária, foram testadas arquiteturas com quantidades de um a trinta neurônios, de forma a verificar qual quantidade de neurônios gera os menores erros na camada de saída.

Na camada de saída, os valores gerados são, assim como os dados de entrada, entre 0 e 1. Assim, realizou-se um arredondamento entre os valores de saída para as três classes de clientes, conforme Quadro 4. Como são três classes na camada de saída, a RNA contempla três neurônios na última camada.

Quadro 4 – Valores da camada de saída

Valor de saída	**Valor de saída arredondado**	**Classe**
$y < 0{,}33$	0	Inadimplentes
$0{,}34 < y < 0{,}66$	0,5	Temporariamente inadimplentes
$y > 0{,}66$	1	Adimplentes

Fonte: A autora (2018).

Outro parâmetro a ser considerado é a função de ativação. Com o papel de limitar a amplitude da saída do neurônio, a função de ativação escolhida foi a logística. Esta função pertence à classe das funções sigmoides, que são as mais utilizadas para redes neurais. Sendo uma função diferenciável, os valores de saída da função logística variam de 0 a +1 (HAYKIN, 2009).

Como critério de parada, adotou-se o valor de 0,6 para o parâmetro *threshold*, o que especifica o limite para as derivadas parciais da função de erro. O número de neurônios utilizados na camada oculta foi 11. Por fim, a taxa de aprendizagem para a RNA foi definida como 0,01. O Quadro 5 resume os parâmetros adotados na RNA.

Todos estes parâmetros foram escolhidos por meio de exaustivos testes computacionais, com exceção da arquitetura, função de ativação e algoritmo, que foram definidos a partir da revisão de literatura. Assim, os testes permitiram escolher os demais parâmetros que acarretaram no melhor resultado para a RNA.

Quadro 5 – Parâmetros da RNA

Parâmetro	**Definição**
Arquitetura	*Perceptron* de multi camadas
Função de ativação	Logística
Algoritmo	*Backpropagation*
Neurônios na camada de entrada	16
Número de camadas ocultas	1
Neurônios na camada oculta	11
Neurônios na camada de saída	3
Critério de parada	0,6
Taxa de aprendizagem	0,01

Fonte: A autora (2018).

3.3 Treinamento da RNA

O *software* utilizado para toda a programação, treinamento e obtenção de resultados foi o *software* livre R (R CORE TEAM, 2018). Existem diversas funções do R que podem ser utilizadas para redes neurais. Neste trabalho, utilizou-se a função *neuralnet()* do pacote *neuralnet*.

A partir da parametrização da RNA descrita no seção 3.2, o treinamento da rede pode ser iniciado. Nesta etapa, a rede foi alimentada com 2/3 da amostra com os dados históricos dos clientes. Assim, o restante poderia ser utilizado posteriormente para validação da configuração final da rede.

A partir das dezesseis entradas, a RNA teve os pesos ajustados conforme o algoritmo de treinamento do tipo *backpropagation.* Este treinamento foi supervisionado, ou seja, foi fornecido o "*status* do empréstimo" para que a rede aprendesse com a resposta.

Para a etapa de validação, conforme feito durante o treinamento, os dados dos clientes foram inseridos na rede como entradas, porém, não foram fornecidas as saídas desejadas. Desta forma, a RNA calculou as saídas (valor previsto), que foram então comparadas com os *status* do empréstimo (valor real).

3.4 Regressão linear múltipla

A aplicação da RLM fez-se a partir do *software* R, pela função *Fitting Linear Models (lm)* do pacote *stats*. Assim como na RNA, a aplicação do modelo se divide em fornecimento de dados e validação do modelo. Para fornecimento dos dados, utilizou-se a mesma amostra de treinamento da RNA e, para validação do modelo (teste com as variáveis), utilizou-se a amostra de testes utilizada na RNA.

Sendo assim, é possível confrontar os resultados dos dois modelos, a partir da mesma amostra de dados, e verificar qual método é mais assertivo na sua previsão de risco de crédito.

3.5 Método de avaliação de desempenho

Para a avaliação do desempenho dos métodos, utilizou-se a matriz de confusão e o cálculo de assertividade.

A matriz de confusão é comumente utilizada para avaliar o desempenho de sistemas de classificação. Ela é uma matriz que contém os dados sobre a classificação real e prevista de um sistema de classificação, sendo comumente utilizado para avaliar a performance desses sistemas (PATIL e SHEREKAR, 2013).

Já a assertividade é definida, neste artigo, como sendo o número de acertos na classificação dos clientes sobre o número total de classificações (previsões) feitas, conforme a Equação (17).

$$assertividade = \frac{nº\ acertos}{nº\ total\ previsões} \quad (17)$$

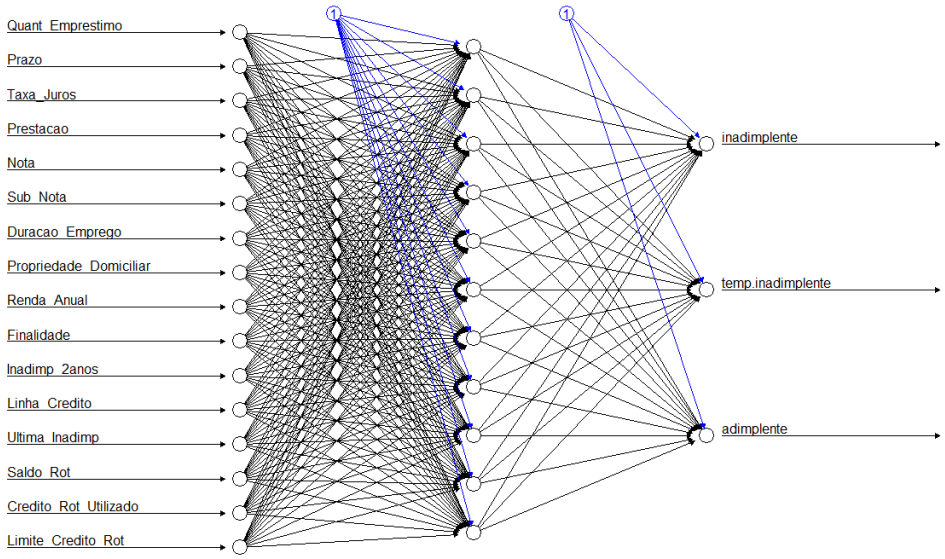
4 RESULTADOS

A seguir serão apresentadas a RNA com sua arquitetura final e suas saídas, a equação de RLM com seus respectivos parâmetros e classificações e uma análise da correlação entre as variáveis, a fim de se proceder com as considerações dos resultados adquiridos.

4.1 Resultados RNA

A RNA, definida a partir dos parâmetros do Quadro 5, pode ser ilustrada na Figura 3.

Figura 3 – RNA final



Fonte: A autora (2018).

A classificação de clientes da rede pode ser observada a partir de uma matriz de confusão, apresentada no Quadro 6.

Quadro 6 – Matriz de confusão RNA

		Real		
		Adimplentes	Temporariamente Inadimplentes	Inadimplentes
Previsto	Adimplentes	351	91	120
	Temporariamente Inadimplentes	0	1	0
	Inadimplentes	15	9	13

Fonte: A autora (2018).

Realizando o cálculo de assertividade, a RNA obteve um desempenho de 60,8%. Apesar da RNA acertar mais do que a metade para o conjunto de teste, os resultados observados no Quadro 6 indicam que a RNA classifica corretamente um número considerável de clientes adimplentes, porém nas outras duas classes a rede não é uma boa classificadora.

4.2 Resultados RLM

A aplicação do método de RLM gerou uma equação linear com os coeficientes de regressão β_i definidos pela Equação 18, no qual $\beta_0 = 0{,}23$ e $\varepsilon_j = 037$. As variáveis explicativas X_{ij} são descritas no Quadro 8, no qual é apresentado o valor de correlação entre as variáveis.

$$\begin{aligned} Y_j = {} & 0{,}23 + 0{,}27X_{1j} - 0{,}1X_{2j} + 0{,}1X_{3j} - 0{,}52X_{4j} - 0{,}1X_{5j} + 0{,}6X_{6j} + 0{,}1X_{7j} \\ & - 0{,}05X_{8j} + 6{,}49X_{9j} + 0{,}09X_{10j} + 0X_{11j} + 0X_{12j} + 0{,}18X_{13j} \\ & - 2{,}24X_{14j} + 0{,}07X_{15j} + 1{,}84X_{16j} + 0{,}37 \end{aligned} \tag{18}$$

Após a aplicação do método de RLM com a mesma amostra de treinamento da RNA, gerou-se uma matriz de confusão, a partir dos testes realizados com a

amostra de testes. A classificação dos clientes em adimplentes, temporariamente inadimplentes e inadimplentes pode ser visualizada na matriz de confusão do Quadro 7.

Quadro 7 – Matriz de confusão RLM

		Real		
		Adimplentes	Temporariamente Inadimplentes	Inadimplentes
Previsto	Adimplentes	240	43	51
	Temporariamente Inadimplentes	126	58	82
	Inadimplentes	0	0	0

Fonte: A autora (2018).

Realizando o cálculo de assertividade, a RLM obteve um desempenho de 49,7%.

4.3 Análise de Resultados

Para análise dos resultados, utilizou-se o método de correlação entre as variáveis. A análise de correlação busca determinar o grau de relacionamento entre duas variáveis. O coeficiente de correlação r varia entre -1 e 1 e, quando positivo, as variáveis analisadas possuem variações no mesmo sentido. Em contrapartida, quando o coeficiente é negativo, as variáveis tendem a variar em sentidos opostos (MUKAKA, 2012).

Sendo $r = 1$ a correlação positiva perfeita e $r = -1$ a correlação negativa perfeita, existem também casos em que não há nenhuma correlação entre as variáveis, ou seja, $r = 0$. Quanto mais próximos a zero, menor é o grau de relacionamento linear entre as variáveis (MUKAKA, 2012).

Um dos métodos mais utilizados para cálculo de correlação de variáveis é pelo coeficiente linear de Pearson (MUKAKA, 2012). Este método foi utilizado para comparar a relação entre as variáveis dos clientes de *peer-to-peer lending* com a variável resposta *"status* do empréstimo" e seus coeficientes apresentados no Quadro 8.

Quadro 8 – Correlação entre a variável resposta e todas as variáveis previsoras.

Descrição	**Variável**	**Correlação**
Quantia do empréstimo	X_1	-0.097
Prazo	X_2	-0.146
Taxa de juros	X_3	-0.231
Prestação	X_4	-0.087
Nota	X_5	0,227
Subnota	X_6	0,233
Duração do emprego	X_7	0,012
Propriedade domiciliar	X_8	-0,035
Renda anual	X_9	-0,036
Finalidade	X_{10}	0,040
Inadimplência 2 anos	X_{11}	-0,028
Linha de crédito mais antiga	X_{12}	-0,043
Última inadimplência	X_{13}	0,029
Saldo rotativo	X_{14}	-0,003
Crédito rotativo utilizado	X_{15}	-0,067
Limite crédito rotativo	X_{16}	0,040

Fonte: A autora (2018).

Pode-se observar que a correlação entre todas as variáveis de entrada com a variável resposta "*status* do empréstimo" é baixa. Isso permite inferir que, pelo fato da variável resposta não ter relação forte com as variáveis preditoras, os métodos de previsão podem não convergir para resultados satisfatórios.

Essa teoria é corroborada pelos pesos das variáveis explicativas da equação de RLM, que são coeficientes baixos e indicam que as variáveis explicativas possuem vínculos de dependência fracos em relação à variável resposta.

5 CONSIDERAÇÕES FINAIS

Apesar de a RNA possuir assertividade maior que a RLM (60,8% e 49,7%, respectivamente) ambos os casos são considerados ineficientes na classificação de risco de crédito para esta amostra de dados. A RNA, por exemplo, classificou 120 clientes como adimplentes, enquanto eram, na verdade, inadimplentes.

Por outro lado, a RLM não classificou nenhum dos 133 clientes inadimplentes como tal. Ambas situações poderiam causar grandes prejuízos para a concessão de crédito baseada nos modelos apresentados, seja pela concessão inapropriada ou por deixar de conceder crédito a clientes em potencial.

Para futuros trabalhos, sugere-se a utilização de outros algoritmos para o treinamento da RNA, como Função de Base Radial (RBF) e *Backpropagation* Resiliente (RB). Para o pré-processamento de dados, é possível trabalhar com números binários. Sugere-se, também, a utilização de um número menor de variáveis e com maiores correlações com a variável resposta.

REFERÊNCIAS

AMARAL Jr, J. B.; TÁVORA Jr, J. L. Uma análise do uso de redes neurais para a avaliação do risco de crédito de empresas. Revista do BNDES, v. 34, p. 134, 2010.

ASSEF F.M. Algoritmos de classificação em aplicação financeira: avaliação de risco de crédito para pessoa jurídica. Tese (Mestrado em Engenharia de Produção) – Universidade Federal do Paraná. Curitiba. 2018.

BARRETO, V. C. S.; BATISTELA, G. C.; GAIOTTO, M. R.; SIMÕES, D. Regressão linear múltipla aplicada ao preço do leite. CQD Revista Eletrônica Paulista de Matemática, v. 7, p. 109 – 118, 2016.

CASTRO L. N. Fundamentals of natural computing: basic concepts, algorithms, and applications. Nova Iorque: Chapman & Hall/CRC, 2006.

CENTA, S. A. Análise de crédito. 3 ed. Curitiba: IBPEX, 2005.

COUTINHO, E. R.; SILVA, R. M.; DELGADO, A. R. S. Utilização de técnicas de inteligência computacional na predição de dados metereológicos. Revista Brasileira de Meteorologia, v. 31, n. 1, p. 21 – 36, 2016.

CYBENKO, G. Approximation by superpositions of a sigmoidal function. Math. Control Signal Systems, v. 2, n. 4, p. 303–314, 1989.

DRAPER, N. R.; SMITH, H. Applied regression analysis. 3 ed. Nova Iorque: John Wiley & Sons, Inc., 2014.

GAZOLA, S. Construção de um modelo de regressão para avaliação de imóveis. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Catarina. Florianópolis, 2002.

GONÇALVES E. B. Análise de risco de crédito com o uso de modelos de regressão logística, redes neurais e algoritmos genéricos. Tese (Mestrado em Administração) – Universidade de São Paulo. São Paulo, 2005.

GRZEIDAK E. Identification of nonlinear systems based on extreme learning machine and multilayer neural network. Tese (Mestrado em Engenharia Mecânica) – Universidade de Brasília. Brasília, 2016.

HAYKIN, S. Neural networks and learning machines. 3 ed. Hamilton, Canadá: Pearson Prentice Hall, 2009.

HOJI, M. Administração financeira e orçamentária. 9 ed. São Paulo: Atlas, 2010.

HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. Neural Networks, v. 2, n. 5, p. 359–366, 1989.

LEMES Jr, A. B.; RIGO, C. M.; CHEROBIM, A. P. M. S. Administração financeira: Princípios, fundamentos e práticas brasileiras. 3 ed. Rio de Janeiro: Elsevier, 2010.

LENDING CLUB. Disponível em: <https://www.lendingclub.com>. Acesso em: 28 abr. 2018.

MÁSSON, E.; WANG, Y. Introduction to computational and learning in artificial neural networks. European Journal of Operational Research, Dinamarca, p. 1-28, 1990.

MATEESCU A. Peer-to-peer lending. Nova Iorque: Data & society research institute. 2015.

MCCULLOCH, W.S.; PITTS, W.H. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, p. 115-133, 1943.

MÓDOLO, M. Classificação automática de supernovas usando redes neurais artificiais. 2016. Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais. São José dos Campos. 2016.

MONTGOMERY, D. C.; RUNGER, G. C. Estatística aplicada e probabilidade para engenheiros. 2. ed. Rio de Janeiro: LTC, 2008.

MUKAKA M. M. A guide to appropriate use of Correlation coefficient in medical research. Malawi Medical Journal, v. 24, n.3, p.69 – 71, 2012.

MUNAKATA, T. Fundamentals of the new artificial intelligence: neural, evolution, fuzzy and more. 2 ed. Cleveland: Springer, 2008.

NATHANS, L. L.; OSWALD, F. L.; NIMON, K. Interpreting Multiple Linear Regression: A Guidebook of Variable Importance. Practical Assessment, Research & Evaluation, v. 17, n. 9, 2012.

PATIL, T. R.; SHEREKAR, S. S. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal of Computer Science and Applications, v. 6, n. 2, 2013.

PRUNZEL, J.; TOEBE, M.; LOPES, A. B.; MOREIRA, V. S. Modelos de regressão linear múltipla aplicados à avaliação de terrenos urbanos - caso de município de Itaqui-RS. Boletim de Ciências Geodésicas, v. 22, n. 4, p. 651 – 664, 2016.

QUILES, M. G. Sistema de Visão Baseado em Redes Neurais para o Controle de Robôs Móveis. 2004. Dissertação (Mestrado em Clência da Computação e Matemática Computacional) - Universidade de São Paulo. São Paulo. 2004.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018. Disponível em: www.R-project.org.

REZENDE, S. O. Sistemas Inteligentes: Fundamentos e Aplicações. 1. ed. Barueri: Manole, 2003. ISBN 85-204-1683-7.

ROBINSON M. D.; OSHLACK, A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 2010.

SCHMIDHUBER, J. Deep learning in neural networks: an overview. Neural networks, Suíça, p. 85-117, 2014.

SEHN, C. F.; CARLINI Jr., R. J. Inadimplência no sistema financeiro de habitação: um estudo junto à Caixa Econômica Federal. Revista de Administração Mackenzie, v. 8, n. 2, 2007.

SU-FEN, C. Dynamic Population Structure based PSO with granular computing for unified multiple linear regression. Information technology journal. Nanchang Jiangxi, China. 2013

YILMAZ, I.; KAYNAR, O. Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. Expert Systems with Applications, v. 38, n. 5, p. 5958 – 5966, 2010.

ZHANG, F.; TADIKAMALLA, P. R.; SHANG, J. Corporate credit-risk evaluation system: Integrating explicit and implicit financial performances. International Journal of Production Economics. v. 177, p. 77 – 100, 2016.