UNIVERSIDADE DO VALE DO RIO DOS SINOS – UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS GRADUAÇÃO
MBA EM GESTÃO ESTRATÉGICA EM *BANKING*

DIEISON ANTONELLO DEPRÁ

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA CONSTRUÇÃO DE MODELOS DE RISCO DE CRÉDITO

Porto Alegre

2023

DIEISON ANTONELLO DEPRÁ

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA CONSTRUÇÃO DE MODELOS DE RISCO DE CRÉDITO

Artigo apresentado como requisito parcial para obtenção do título de Especialista em Gestão Estratégica, pelo Curso de MBA em Gestão Estratégica em *Banking* da Universidade do Vale do Rio dos Sinos – UNISINOS

Orientadora: Profa. Ms. Graziela Molling

Porto Alegre

2023

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA CONSTRUÇÃO DE MODELOS DE RISCO DE CRÉDITO:

APPLICATION OF MACHINE LEARNING TECHNIQUES IN CONSTRUCTION OF CREDIT RISK MODELS

Dieison Antonello Deprá [*]

Graziela Molling [**]

Resumo: A capacidade de avaliar o risco é fundamental às instituições do setor financeiro e a existência de casos não capturados por um modelo traz incerteza ao cenário. Neste contexto o presente trabalho explora técnicas de transformação de dados, reconhecimento de padrões e aprendizado de máquina para construção de um modelo de risco de crédito que possa melhorar a avaliação e mitigar o risco na concessão do crédito. A metodologia utilizada empregou uma mescla de pesquisas quanti-quali, bibliográfica e exploratória que culminou em uma proposta de rede neural convolucional 2D capaz de classificar um conjunto de imagens, que representam dados financeiros em séries temporais de uma base de clientes, entre pagadores “bons” e “ruins”. Os resultados experimentais produzidos pela arquitetura proposta são apresentados para diversas métricas como, por exemplo, acurácia de 97,96%, precisão de 98,32%, especifidade de 98,18%, sensibilidade de 97,77%, taxa de falsos positivos de 1,82%, F1 de 98,14% e AUC de 97,09%, sendo que tais resultados competem com seus pares definindo o estado da arte.

Palavras-chave: Modelagem do Risco de Crédito. Aprendizado de Máquina. Regressão Logística (LR). Redes Neurais (NN). Redes Neurais Convolucionais (CNN).

Abstract: The ability to assess risk is fundamental to financial sector institutions and the existence of cases not captured by a model brings uncertainty to the scenario. In this context, the present work explores data transformation techniques, pattern

[*] Dieison Antonello Depra – Mestre em Ciência da Computação pela Universidade Federal do Rio Grande do Sul (UFRGS – 2009). Bacharel em Ciência da Computação pela Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS – 2004). Possui interesse e publicações acadêmicas nas áreas de estatística, jogos 2D, integração de *hardware* e *software* e arquiteturas de *hardware* especializadas para compressão vídeo. Atuação profissional com mais de 20 anos de experiência no desenvolvimento de sistemas. Contato: dieisondepra@gmail.com.

[**] Graziela Molling – doutoranda e Mestre em Administração na Universidade do Vale do Rio dos Sinos. Graduada em Bacharelado Ciências Contábeis pela Universidade Feevale (2012) e Sistemas de Informação (2014) pela Universidade do Vale do Rio dos Sinos (UNISINOS). Pós-Graduação em nível de Especialização de Governança em TI baseado em Padrões Internacionais pela Universidade do Vale do Rio dos Sinos (UNISINOS). Consultora de Negócios e TI, com mais de 10 anos de experiência na área de TI. Contato: gmolling@edu.unisinos.br.

recognition and machine learning to build a credit risk model that can improve the assessment and mitigate the uncertainty in the granting of credit. The process was guided by a mixture of quantitative-qualitative, bibliographic and exploratory research that culminated in a proposal for a 2D convolutional neural network capable of classifying a set of images, which represent financial data in time series of a customer base, among payers “good” and “bad”. The experimental results produced by the proposed architecture are presented for several metrics such as accuracy of 97.96%, precision of 98.32%, specificity of 98.18%, sensitivity of 97.77%, false positive rate of 1.82%, F1 of 98.14% and AUC of 97.08%. with such results competing with their peers making the state of the art.

Keywords, Mots-clès e Palabras clave: *Credit Risk Modeling. Machine Learning. Logistic Regression (LR). Neural Networks (NN). Convolutional Neural Networks (CNN).*

1 INTRODUÇÃO

No setor financeiro classificar clientes quanto ao seu risco de crédito é essencial e, neste sentido, Castro Neto (2009, p. 21) diz que: “[…] Para avaliar o risco do tomador de empréstimo, a instituição financeira deverá possuir funcionários capacitados, além de softwares e aplicativos específicos para estimar o risco de cada cliente”.

Além disso, conforme citado por Kammoun e Triki (2016, p. 1) “[…] a análise de risco de crédito ganha ênfase com base nas recentes crises financeiras e regulamentações originadas nos acordos de Basiléia II e III […]”[1]. Tais acordos estabeleceram, dentre outros deveres, a obrigatoriedade das instituições financeiras definirem e aplicarem modelos matemáticos consistentes para avaliação do risco de crédito de forma a permitir o acompanhamento e mitigação sistemática deste risco.

Segundo Mays (2001, p. 91)[1] “[…] a modelagem matemática através das técnicas estatísticas, dentre as quais pode-se citar, em especial, o modelo de regressão logística, é normalmente aplicada para construção de modelos de risco”. Entretanto, para alguns cenários de uso específico, onde certos clientes possuem especificidades (os chamados “*outliers*”), este tipo de modelagem pode não ser adequada para capturar casos extremos. Assim, para mitigar esse problema, a utilização de técnicas de aprendizado de máquina tem sido explorada, pois segundo Gollapudi (2016, p. 34)[1]: “[…] é um mecanismo para pesquisa de padrões e construção de inteligência em uma máquina, para que mesma possa apreender e fazer melhor no futuro a partir de sua própria experiência.”.

[1] Em tradução livre.

A existência de casos que não sejam capturados adequadamente por um determinado modelo pode representar um risco desconhecido para as instituições financeiras e, portanto, explorar técnicas que possibilitem minimizar a incidência de tais riscos são desejáveis. Nesse sentido, Kammoun e Triki (2016, p. 2)[1] apontam que “[…] métodos mais sofisticados como inteligência artificial, em especial redes neurais, são comumente empregados como alternativa a técnica de regressão logística.”

Nesse contexto, a temática do trabalho envolve a aplicação de técnicas de aprendizado de máquina para construção de um modelo de risco de crédito para responder a questão: “a utilização de técnicas de aprendizado de máquina poderia oferecer ganhos aos modelos de risco de crédito?”. Logo, o objetivo principal passa pela construção de uma solução para classificação de risco de crédito baseado em técnicas de aprendizado de máquina que possa ser comparável as propostas da literatura. Como objetivos específicos o trabalho visa identificar algum caso de uso direto para a aplicação da técnica; mapear quais as variáveis e fontes de dados devam ser utilizadas; construir um modelo de risco com a utilização de técnicas de aprendizado de máquina; coletar os resultados experimentais para comparação e discussão frente a literatura.

A estrutura do restante do texto traz, inicialmente, a fundação teórica do assunto e uma revisão da literatura sobre trabalhos relacionados, a seguir passa pela metodologia aplicada para construção do modelo e, na sequência, a apresentação dos resultados obtidos, passando, posteriormente, pela discussão sobre as vantagens e desvantagens de cada abordagem e fechando com algumas considerações sobre o trabalho realizado.

2 FUNDAMENTAÇÃO TEÓRICA

Neste trabalho estamos interessados na capacidade de predição de risco de crédito através de técnicas de aprendizado de máquina, mais especificamente com uso de redes neurais. Antes, porém, vamos visitar algumas definições sobre esses conceitos e, na sequência, apresentar um resumo das principais propostas encontradas na literatura.

2.1 Risco de Crédito

Classificado pelo comitê de supervisão bancária (Basiléia) como risco de contraparte o risco de crédito é definido por Kammoun e Triki (2016, p. 62) como "[…] risco do mutuário ou contraparte não honrar com suas obrigações conforme os termos acordados […]" ou ainda como "[…] risco de calote ou redução do valor de mercado das garantias ou avalistas envolvidos na operação […]", conforme Duffie e Singleton (2003 apaud Kammoun e Triki, 2016).

O risco de crédito na intermediação financeira deriva, segundo Kammoun e Triki (2016, p. 63), da assimetria da informação entre as partes, dado que "[…] nem todo o participante tem a mesma informação, ao mesmo tempo, sobre o objeto do contrato […]". Ainda, baseado em Kammoun e Triki (2016, p. 63), podemos inferir que a criação de sistemas de *credit score* é uma abordagem que visa enfrentar os riscos de crédito na intermediação financeira, uma vez que permite diferenciar entre bons e maus pagadores, a partir de características dos mutuários, gerando classes ou grupos que podem ser ponderados pelo risco.

A seguir serão apresentadas as duas principais técnicas que têm sido empregadas para construção de sistemas de *credit score*, ou seja, a modelagem estatística através de regressão logística e os modelos baseados em técnicas de aprendizado de máquina, inteligência artificial, especialmente em redes neurais.

2.2 Regressão Logística

A regressão logística (LR[2]) é comumente utilizada na construção de modelos de risco de crédito (MAYS, 2001, p. 72) e pode ser descrita, segundo Alice (2015a), como "[…] uma técnica que permitir classificar uma variável Y baseada na computação de um conjunto X, de variáveis interdependentes, que podem ser contínuas, categóricas ou uma mistura de ambas."[3]. Assim, como exemplo, pode-se obter como resposta (variável dependente) se um determinado cliente é bom ou mal pagador baseado em outros dados do seu perfil financeiro (variáveis interdependentes).

A LR também pode ser descrita, conforme Hari et al (2010, p. 315), como uma forma especializada de regressão que permite classificar e explicar uma variável

[2] LR – acrônimo a partir do nome em inglês: *Logistic Regression*.

[3] Em tradução livre.

categórica binária. Além disso, segundo Gupta e Goyal (2018, p. 10-11), "[…] a grande aplicação da LR para sistemas de *credit score* se deve ao fato de sua simplicidade de aplicação e bons resultados nas tarefas de classificação para respostas binárias (0 ou 1, sim ou não, verdadeiro ou falso)."[3].

Uma vantagem da LR, segundo Kammoun e Triki (2016, p. 64), é que ela não requer restrições ou normalização das variáveis tampouco da distribuição da população entre variância e covariância. Entretanto, possui a desvantagem de que pequenas amostras podem resultar em estimativas inadequadas. Além disso, autores como Correa et al. (2011, p. 6) apresentam experimentos que demonstram que a precisão dos resultados obtidos com a LR pode ser inferior aos resultados obtidos com redes neurais, que será apresentada a seguir.

2.3 Redes Neurais Artificiais

As redes neurais artificiais são, segundo Gollupadi (2016), um subconjunto das técnicas de aprendizado profundo (*Deep Learning*) que, por sua vez, é subconjunto de técnicas de aprendizado de máquina (*Machine Learning*) as quais são um subconjunto de técnicas de inteligência artificial (*Artificial Intelligence – AI*). As técnicas de aprendizado de máquina têm sido, especialmente, aplicadas aos problemas de classificação de dados e, neste sentido, há, segundo o Golluapdi (2016, p. 34), uma similaridade com a LR.

A definição, segundo Napitupulu e Triana (2019, p. 2), de uma rede neural artificial (RNA) é:

> […] um sistema de processamento de informação que possui características similares a uma rede neural biológica (RNB). Sua habilidade de reconhecimento de padrões e capacidade de aprendizado permite que seja aplicada para resolução de diversos problemas que seriam complexos ou impossíveis de serem resolvidos através de técnicas computacionais ou métodos estatísticos convencionais.

Além disso, segundo Fausett (1994 apaud Napitupulu e Triana 2019), considerando as similaridades das RNAs com as RNBs pode-se dizer que as RNAs são compostas por "neurônios" interconectados em camadas, os quais são formados por três tipos de componentes: i) dendritos ($w_{1,2,n}$) que recebem informação de outros neurônios; ii) função de ativação; e iii) um axônio (g) que leva informação de saída para outros neurônios.

Conforme Gupta e Goyal (2018, p. 11) uma RNA é composta por um conjunto de neurônios organizado em uma topologia específica que, no modelo clássico (MLP[4]), contempla três camadas: i) uma camada de entrada; ii) uma ou mais camadas ocultas; e iii) uma camada de saída. Além disso, as redes neurais (NN[5]) apresentam variações, seja na organização topológica em camadas, nas funções de ativação utilizadas, nas técnicas de retroalimentação empregadas ou quanto aos tipos de problemas de aprendizado pretendem resolver (GOLLUPADI, 2016).

Em um estudo de caso, aplicado por Alice (2015b), foram comparados, usando o erro médio quadrático (MSE), os resultados obtidos com um modelo de LR frente aos obtidos com uma NN no qual concluiu-se que, para aquele caso de uso específico, às predições do modelo gerado pela NN apresentavam maior acurácia. Portanto, a seguir, serão discutidas algumas aplicações de aprendizado de máquina para o contexto de dados financeiros, as quais foram selecionadas a partir da revisão bibliográfica.

2.4 Detecção de risco de crédito através de NN baseada em árvores

No estudo realizado por Su (2009, p. 81-106) apresentou-se o uso combinado de várias técnicas de aprendizado de máquina, aplicando múltiplas florestas de decisão aleatórias (RF[6]) com objetivo de melhorar a precisão dos eventos de identificação de risco em operações de crédito. O trabalho se destaca por aplicar técnicas de impulsionamento conhecidas como *Boosting*, tais como AdaBoost, XGBoost e LightGBM visando obter ganhos em efetividade e sensibilidade, dentre os quais a técnica XGBoost mostrou os melhores resultados.

No trabalho desenvolvido por Sayjadah (2018, p. 1-4) é realizada a avaliação de performance com diferentes métodos, dentre os quais pode-se citar: a LR, as árvores de decisão particionadas (PDT[7]) e as RF; visando medir o comportamento para variável de teste e, assim, predizer o risco de calote em cartões de crédito. Os resultados apresentados demonstram que o método RF atingiu precisão superior a 80%, com um percentual de área dentro curva (AUC[8]) acima de 77%. A conclusão

---

4 MLP – acrônimo a partir do nome em inglês: *Multi Layer Perceptron*.
5 NN – acrônimo a partir do nome em inglês: *Neural Networks*.
6 RF – acrônimo a partir do nome em inglês: *Random Forest*.
7 PDT – acrônimo a partir do nome em inglês: *Partition Decision Tree*.
8 AUC – acrônimo a partir do nome em inglês: *Area Under the Curve.*

dos autores sugere que a utilização de técnicas de aprendizado de máquina poderia melhorar a precisão das previsões sobre o risco de crédito para o setor financeiro.

2.5 Adaptação de CNN para modelos de predição do risco de crédito

As redes neurais convolucionais (CNN[9]) são um tipo de NN que combina elementos de exploração de similaridade espacial com MLP, tendo aplicação mais comum no processamento de imagens e, portanto, esperam que os dados estejam no formato de matrizes, que podem ser bidimensionais (2D), com as dimensões definidas por altura e largura, para imagens monocromáticas (tons de cinza), ou tridimensionais (3D), com dimensões definidas por altura, largura e profundidade, sendo a profundidade dada pelos componentes de cor (SHARMA, 2019).

As redes do tipo CNN apresentam vantagens sobre as redes do tipo MLP, conforme apontado por Sharma (2019, p. 2), como, por exemplo: i) não requererem seleção prévia das características mais importantes; ii) capacidade de encontrar relações estatísticas de ordem superior ou correlações não lineares; iii) demandam menor quantidade de neurônios, por usar o princípio da localidade, ao passo que melhoram a profundidade do aprendizado com poucos parâmetros; e iv) compartilhamento de componentes como pesos, bias e filtros fato que propicia redução do uso de memória.

Entretanto, para permitir a exploração das vantagens de redes do tipo CNN em outros contextos, onde os dados de entrada não estão no formato de imagens, é necessário realizar um processo de transformação de forma que os dados de entrada possam ser “enxergados” pela rede como um conjunto de imagens.

Uma estratégia de organização de dados financeiros (não imagens) para transformação em “imagens” é apresentada no trabalho desenvolvido por Yan (2018), no qual uma CNN é aplicada para tentar prever quais produtos bancários serão utilizados por cliente no próximo mês, baseado no seu histórico passado e no comportamento dos perfis similares. No estudo foram empregados dados históricos sobre o uso de 24 produtos (mais suas variações, num total de 47 opções) do banco Santader num período de 16 meses. A organização proposta por Yan (2018) consiste em criar uma imagem de 16 linhas e 47 colunas por cliente, sendo que cada coluna

[9] CNN – acrônimo a partir do nome em inglês: *Convolutional Neural Network*

representa a utilização de um dos produtos e cada linha representa a posição do cliente num dado mês.

No trabalho realizado por Tai e Huyen (2019) são apresentadas duas aplicações de NN profunda para o contexto de crédito *score*, sendo uma do tipo CNN. Vale destacar outro ponto interessante deste que é a utilização de três bases de dados públicas para as aplicações propostas, fato que facilita a comparação dos resultados apresentados pelos autores com outras propostas encontradas na literatura. Além disso, a estrutura das redes propostas é apresentada e discutida. E, por fim, os experimentos são detalhados e os resultados apresentados são comparados, para as três bases de dados citadas, através de quatro parâmetros distintos, em 6 aplicações diferentes (2 dos próprios, mais 4 da literatura), indicando que, em geral, as redes do tipo CNN funcionam melhor que os outros tipos de redes discutidos.

Em Golbayani (2020) quatro tipos de redes neurais são aplicadas para tentar prever o risco de crédito associado a cada uma das empresas do S&P500[10]. O artigo se propõe a responder algumas questões interessantes, dentre as quais destacam-se: i) que a utilização de todas as variáveis disponíveis produz melhores resultados que utilizar apenas um subconjunto delas; ii) que a utilização de dados em séries temporais melhora a eficiência dos modelos; e iii) que os melhores resultados em termos de eficiência, considerando os dados financeiros, foram obtidos com variações de redes do tipo CNN ou LSTM[11].

A seguir, após a breve revisão conceitual e do estado da arte aqui discutida, a próxima seção apresentará os materiais e o método aplicado neste trabalho.

3 MATERIAL(IS) E MÉTODOS

A primeira etapa para realização do estudo proposto foi identificar dentro do Banrisul quais os colabores responsáveis pela modelagem de risco, sendo que tal questionamento foi formalizado junto à Universidade Corporativa, através de fluxo interno (*workflow*).

Assim, na segunda etapa, elaborou-se um roteiro de entrevista semi-estruturada, o qual foi aplicado junto ao colaborador indicado pela instituição (ver

[10] Trata-se de um índice composto por quinhentos ativos (ações) cotados nas bolsas de NYSE ou NASDAQ, sendo um acrônimo do inglês para: *Standard & Poor`s 500*.

[11] LSTM – acrônimo a partir do nome em inglês: *Long Short Term Memory*.

apêndice A), visando obter informações para compressão sobre o modelo de risco de crédito utilizado atualmente, como ele foi construído, quais variáveis de entrada são utilizadas nele, como estas variáveis foram selecionadas, qual o peso de cada variável, quais os resultados obtidos com o modelo atual em termos de precisão de efetividade, como esses resultados são medidos, qual a base de dados utilizada nessa medição, quais as limitações conhecidas e onde o modelo poderia ser melhorado.

Na terceira etapa realizou-se a análise de documentos, considerando as respostas do questionário (apêndice A), bem como avaliação das bases de dados utilizadas pela instituição, buscando identificar quais os dados que estão adequadamente estruturados e, a critério da administração, se poderão ser disponibilizados para utilização na formulação de um modelo neural.

Nesta etapa foram identificados os seguintes complicadores: i) a base de dados utilizada no modelo é fragmentada, sendo composta por dados de diversas fontes e requer autorizações de uso com diversos gestores; ii) o volume de dados utilizado aumenta a complexidade de processamento de forma significativa; iii) há uma série de questões relacionadas ao sigilo de informações (LGPD[12]) e concorrência de mercado que dificultariam a publicação aberta dos resultados; e iv) conjunto de dados num formato próprio da instituição, fato que dificultaria a comparação com outros trabalhos. Assim, para contornar tais problemas, definiu-se o uso de uma base de dados pública, fato que será discutido na seção 3.1.

As próximas etapas na realização deste trabalho envolvem e a seleção e preparação da base de dados a ser utilizada seguida pela construção de uma proposta de modelo de NN a ser treinada para posterior avaliação de seus dados experimentais, passando, por fim, à discussão dos mesmos frente aos resultados encontrados na literatura, sendo que cada um destes tópicos será apresentado em seção própria a seguir.

3.1 Análise e seleção da base de dados

Considerando os fatores “complicadores”, apontados na seção anterior, optou-se por buscar na literatura uma base de dados pública que pudesse ser

[12] LGPD – acrônimo para Lei Geral de Proteção de Dados Pessoais — LEI Nº 13.709, DE 14 DE AGOSTO DE 2018.

utilizada, chegando-se a base de dados de crédito de um banco do sul da Alemanha (South German Credit), a qual é citada nos trabalhos desenvolvidos por: West (2000), Oreski (2012), Doori (2014), Gante (2015) e Tai e Huyen (2019). Tal base de dados, criada pelo Dr. Hans Hofmann, está disponível publicamente[13] no repositório de dados de aprendizado de máquina da UCI[14] (GANTE, 2015), sendo esta a base de dados selecionada para uso no presente trabalho.

A base de dados selecionada possui 1000 registros com dados (reais) de crédito concedidos, os quais foram classificados, quanto a sua adimplência, entre pagadores “bons” (700 registros) e “ruins” (300 registros). Além disso, cada registro da base contém 20 variáveis (características dos clientes) que estão organizadas em dois grupos, sendo: i) o primeiro com 7 variáveis quantitativas contínuas; e ii) o segundo com 13 variáveis qualitativas ordinárias. A Tabela 1 contempla a relação destas variáveis e sua organização, conforme descrito por Oreski (2012).

Tabela 1: Relação de variáveis da base de dados

Variável	Significado	Tipo
C01	Situação da conta com cheque especial	Qualitativa ordinal
C02	Tempo de duração do empréstimo	Quantitativa contínua
C03	Histórico de crédito	Qualitativa ordinal
C04	Propósito do crédito	Qualitativa ordinal
C05	Valor do crédito	Quantitativa contínua
C06	Saldo na poupança	Qualitativa ordinal
C07	Tempo no emprego	Qualitativa ordinal
C08	Percentual da renda comprometida	Quantitativa contínua
C09	Estado cível por gênero	Qualitativa ordinal
C10	Outros débitos ou avalista	Qualitativa ordinal
C11	Tempo residência no local	Quantitativa contínua
C12	Propriedades	Qualitativa ordinal
C13	Idade	Quantitativa contínua
C14	Outros planos de crédito	Qualitativa ordinal
C15	Habitação	Qualitativa ordinal
C16	Número de empréstimo neste banco	Quantitativa contínua
C17	Emprego	Qualitativa ordinal
C18	Dependentes	Quantitativa contínua
C19	Telefone	Qualitativa ordinal
C20	Estrangeiro	Qualitativa ordinal

Fonte: Elaborada pelo autor.

A utilização de uma determinada base de dados por um modelo de NN demanda que a base de dados passe por um conjunto transformações específicas

[13] Link para base de dados: https://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29#

[14] UCI – acrônimo a partir do nome em inglês: Acrônimo para University of California, Irvine.

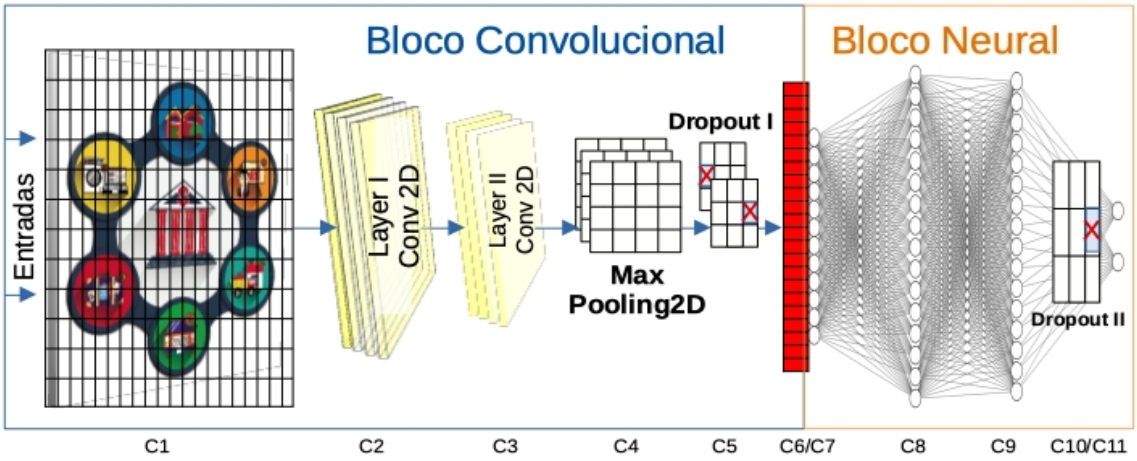
(PROVENZANO, 2020), conforme o tipo de NN aplicado. Na próxima seção, portanto, o modelo proposto será apresentado.

4 MODELO DE REDE NEURAL PROPOSTO

A construção de modelos para redes neurais apresenta variação de características a depender do contexto de uso, conforme apresentado na revisão de literatura (ver sub-seções 2.4 e 2.5). Portanto, no contexto deste trabalho, selecionamos dentre as propostas encontradas na literatura aquelas que pudessem tratar dados financeiros, preferencialmente àquelas com capacidade de incluir amostras temporárias, classificando-as pela métrica de desempenho, em termos de precisão das predições realizadas, como fator decisivo para orientar as escolhas.

Assim, combinando características encontradas nos trabalhos apresentados por Yan (2018), Kvammea (2018), Tai (2019) e Golbayani (2020), foi construída neste trabalho, uma proposta de NN com arquitetura do tipo CNN, organizada em um conjunto sequencial de camadas empilhadas, a qual oferece capacidade de tratar amostras temporais de dados financeiros. A visão estrutural desta arquitetura é apresentada na Figura 1, a qual será detalhada a seguir.

Figura 1: Modelo de NN proposto usando arquitetura CNN



Fonte: Elaborada pelo autor.

Na Figura 1 pode-se observar que há uma divisão estrutural na arquitetura proposta entre dois blocos, conforme descrições apresentadas na parte superior da figura, sendo eles: i) o bloco Convolucional responsável pela extração das características mais relevante da imagem; e ii) o bloco Neural que usa as

características extraídas pelo bloco anterior como entrada para fazer a classificação das amostras, quanto ao seu risco de crédito, entre clientes “bons” e clientes “ruins”. Além disso, dentre esses dois blocos temos 11 componentes identificados, na área inferior da imagem, pelas siglas C1…C11, os quais serão descritos a seguir.

O primeiro componente (C1) são as imagens de entrada que devem estar organizadas como matrizes bidimensionais, com dimensões 12x20 (detalhamento do significado das dimensões na seção 4.1), na qual cada linha representa uma “foto”, num dado momento do tempo, contendo uma coluna para cada variável uma das 20 variáveis descritas na Tabela 1, as quais traduzem a situação financeira do cliente, para fins de avaliação de risco de crédito, naquele instante.

O segundo e o terceiro componentes (C2 e C3) são camadas de convolução em duas dimensões (conv2d), as quais, conforme Kvammea (2018, p. 11), aplicam uma série de 20 filtros paralelos (um para cada característica analisada) que fazem a soma de produtos entre as entradas e pesos ponderados em cada vetor interno. O processamento destas camadas segmenta a matriz de entrada aplicando uma janela de busca com dimensões próprias, sendo 6x4 para C2 e 4x4 para C3, com o objetivo de fazer extração das características mais relevantes da imagem e transformando a matriz de saída para uma nova configuração com dimensões 7x17 para C2 e 4x14 para C3. Na saída de cada um dos componentes (C2 e C3) são aplicadas funções de ativação do tipo ReLU (TAI, 2019, p. 94-95).

O quarto componente (C4) traz uma camada de subamostragem, ou redução da amostra, obtida através da aplicação da função de máximo sobre a janela de busca (dimensões 1x2) que visa reduzir a quantidade de características obtidas nas camadas anteriores, preservando apenas as importantes (TAI, 2019, p. 95). O quinto componente (C5) introduz uma função de descarte aleatório de parte do resultado produzido pelo componente C4 de forma a reduzir o sobre-treino da rede, ou seja, generalizar os resultados da rede para que não sejam tão dependentes das amostras de treino (GOLBAYANI, 2020, p. 4).

O sexto e sétimo componentes (C6, C7) fazem a fronteira entre os blocos convolucional e neural, sendo que no componente C6 as características relevantes, extraídas pelas semelhanças espaciais entre as regiões da imagem, são transpostas de uma matriz para um vetor, o qual servirá de entrada para a primeira camada densa do modelo neural (C7) (camada de entrada). A seguir são aplicadas duas camadas densas ocultas, representadas pelos componentes oito e nove (C8 e C9)

os quais diferem entre si pelo: i) número de neurônios empregados, sendo 128 e 96, respectivamente; ii) pela função de ativação aplicada, sendo ReLU e Sigmoid, respectivamente; iii) pela forma de regularização do núcleo, sendo L2 e L1, respectivamente. O trabalho de Tai e Huyen (2019, p. 94-95) oferece maiores detalhes sobre as funções de ativação ReLU e Sigmoid, assim como as técnicas de regularização de núcleo podem ser encontradas em Srivastavak (2014), Kvammea (2018, p. 16) e Provenzano (2020, p. 4, 6-8).

Os componentes dez e onze (C10 e C11) finalizam a estrutura e, aqui, temos uma variação no bloco neural da arquitetura em relação às MLP tradicionais, pois o componente C10 (que é similar ao C5), responsável por descartar algumas das saídas produzidas pelo componente anterior, foi introduzido após as camadas densas ocultas e antes da última camada densa, camada de saída, representada por C11, com objetivo de melhorar a regularização e generalização da NN, minimizando o sobre-treino. Uma descrição mais aprofundada sobre as técnicas de descarte empregada pelos componentes C5 e C10 pode ser encontrada em Srivastavak (2014).

Outros aspectos importantes do modelo proposto, porém não representados visualmente na Figura 1 são: i) a função de predição de perdas; ii) o algoritmo de otimização de aprendizado; iii) a métrica de avaliação dos resultados; e iv) o método de treino da rede. A função de predição de perdas aplicada, como estamos interessados na classificação dos dados em um de dois conjuntos possíveis, foi a *binary-crossentropy*, conforme citado por Kvamme (2018, p. 10). O algoritmo de otimização de aprendizado empregado foi *ADAM*[15] pois apresenta a melhor relação custo-beneficio entre taxa de aprendizado da rede e consumo de recursos computacionais, para os mais diversos cenários de uso, conforme Muhammad (2020). A métrica de avaliação do resultado foi a acurácia por ser uma das métricas que melhor atende ao cenário de classificação binária conforme discussões em Gudelek, Boluk e Ozbayoglu (2017) e Provenzano (2020). O método de treino empregado no modelo é conhecido como *backpropagation* o qual consiste em ajustar os pesos internos de cada nó da rede com base na comparação entre os resultados produzidos e os esperados, fazendo com que a cada iteração os resultados produzidos tendam a se aproximar dos resultados esperados.

---

[15] ADAM – acrônimo a partir do nome em inglês: *Adaptive Momentum*.

Finalizado o detalhamento da arquitetura de CNN aqui proposto passaremos, na próxima seção, a descrever as transformações aplicadas sobre a base de dados selecionada para que suas amostras pudessem ser processadas.

4.1 Preparação e transformação dos dados

O modelo de NN aqui proposto emprega arquitetura do tipo CNN para a qual os dados de entradas devem estar no formato de imagem bidimensional, tipicamente uma matriz com altura e largura e, além disso, cada dado (valor das variáveis) é tratado como um *pixel*, os quais devem ser dimensionadas para a escala de 8 bits, ou seja, devem estar dentro do intervalo de valores que varia entre 0 e 255 (KVANMME, 2018).

Na Tabela 1 pode-se observar que parte das variáveis da base de dados seleciona são qualitativas, entrando um passo preliminar ao treino das redes neurais do tipo CNN envolve processos de transformação conforme citados por Provenzano (2020, p. 3-4), portanto a primeira transformação aplicada foi converter todas as variáveis de cada uma das amostras para o formato numérico, ou seja, para números inteiros sequenciais começando no número 1.

A segunda transformação aplicada tratou a questão do dimensionamento das amostras, convertendo todos os valores para o intervalo [0-255] de acordo com os valores máximos de cada variável em todas as amostras da base de dados. A seguir, na terceira transformação, todas as amostras foram normalizadas, conforme discutido por Kammoun e Triki (2015, p. 64-73), para melhorar sua distribuição e proporcionar melhores condições de aprendizado para a NN.

Neste ponto, após aplicar as três primeiras transformações, a base de dados estaria apta ao processamento, porém há outro detalhe fundamental. A base de dados selecionada possui amostras em séries individuais, ou seja, apenas uma amostra ou “foto” das características (variáveis) de cada cliente, num dado momento do tempo. Entretanto, no escopo deste trabalho, um dos objetivos é explorar a habilidade do modelo proposto em processar séries temporais, logo, se faz necessário obter (ou gerar) mais amostras de cada cliente.

Assim, na quarta transformação, a base de dados expandida através da geração de novas amostras temporárias para cliente, representando observações em diferentes pontos do tempo. A quantidade de observações foi definida com base

no estudo apresentado por Kvamme (2018, p. 18), que estabeleceu janelas de observações de um ano (365 dias), em 12 períodos (um a cada mês do ano). Logo, para cada cliente, se fez necessário gerar 11 observações adicionais, visto que a base original só possui uma amostra por cliente.

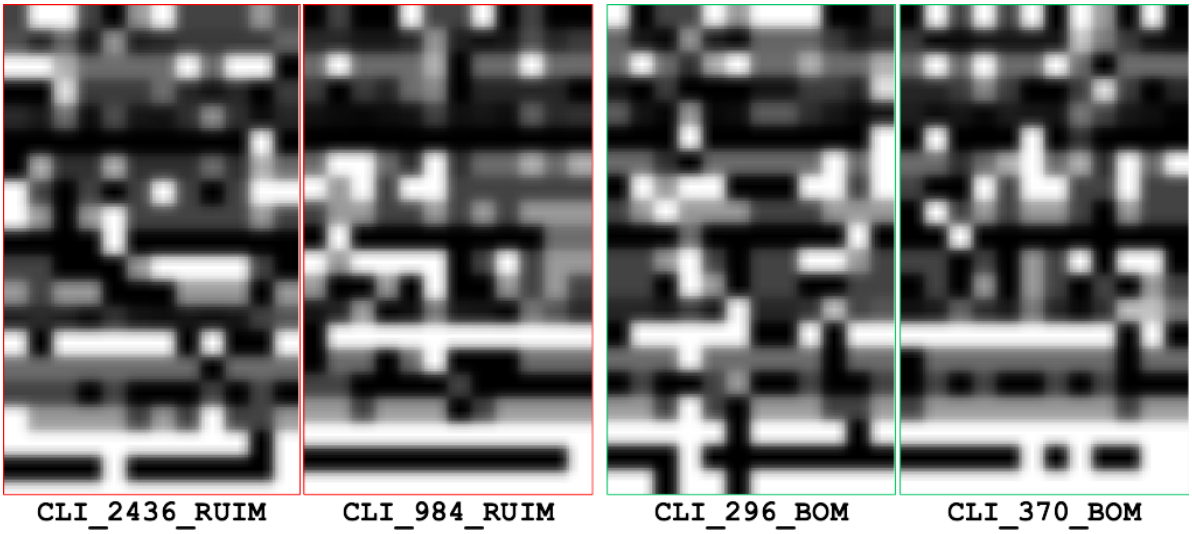
O método empregado para produzir as amostras adicionais foi a composição com amostras de outros clientes, observando a classe de cada um. Assim, separando-se os clientes entre as classes de pagadores (“bons” e “ruins), foram escolhidas, de forma aleatória, amostras de outros 11 clientes da mesma classe, resultando, para cada cliente, num conjunto de 12 amostras temporais com 20 variáveis cada, ou dito de outra forma, produzindo como resultado, para cada cliente, uma matriz bidimensional, com dimensões dadas por 12 linhas e 20 colunas, as quais representariam as características (dados financeiros) do cliente num dado instante de tempo.

No apêndice B são introduzidos dois quadros (1 e 2) os quais apresentam as matrizes geradas para dois clientes distintos, sendo um de cada conjunto de pagadores (“bons” e “ruins”). Além disso, os quadros estão divididos em dois blocos, sendo: i) dados finais das amostras após todas as transformações citadas; e ii) dados das amostras apenas com as transformações 1 e 4.

A Figura 2 visa ilustrar o resultado final do processo de transformação de dados financeiros dos clientes, após todas as transformações citadas anteriormente, em imagens bidimensionais, usando tons de cinza. Nela são apresentadas 4 imagens (com rotação de 90º), que representam dados financeiros de 4 clientes distintos, sendo que tal representação demonstra a forma como modelo proposto “enxerga” as amostras.

Importante observar na Figura 2 que tanto as bordas vermelhas, que envolvem as duas imagens à esquerda, quanto as bordas verdes, que envolvem as duas imagens à direita, não fazem parte dos dados financeiros. Tais contornos foram inseridos propositalmente, para gerar uma separação visual que delimite o agrupamento dos clientes, quanto a sua classificação entre pagadores “ruins” e “bons”, respectivamente. Além disso, há, na parte inferior da Figura 2, para cada imagem, um texto descritivo com a identificação do cliente bem como sua classificação (“RUIM” ou “BOM”).

Figura 2: Representação dos dados financeiros como imagens



Fonte: Elaborada pelo autor.

Considerando os dados contidos na Figura 2 não é simples, a partir de uma observação a olho “nu”, deduzir, diretamente, quais características definem que um cliente deva ser classificado como pagador “ruim” ou “bom”, porém é justamente nesse aspecto que as redes do tipo CNN se destacam, pois conseguem explorar as semelhanças espaciais de cada imagem através do bloco convolucional e, com isso, identificam e extraem as características relevantes que servirão de entrada para o bloco neural, o qual classificará os clientes como pagadores “bons” ou “ruins”.

A finalização desta etapa de transformações precisa ainda considerar dois outros aspectos que são relativos a quantidade e distribuição das amostras, visto que as 1000 amostras presentas base de dados estão organizadas em dois conjuntos, contendo 700 e 300 unidades cada. Entretanto, as redes neurais aprendem melhor quando treinadas em conjuntos de dados maiores e, além disso, é importante que as amostras possuam boa distribuição entre as classes (conjuntos) possíveis.

A estratégia utilizada, para contornar as limitações relativas a quantidade de amostras e sua distribuição, foi de expandir os conjuntos por fatores distintos, sendo 4 (“bons”) e 7 (“ruins”) vezes cada, respectivamente, de forma que número total de amostras passou de 1000 para 5600 amostras, aumento de 560%. A quantidade de amostras para cada classe passou a ser de 2800 e 2400, “bons” e “ruins”,

respectivamente, alterando a distribuição entre elas de 70%-30% para 54%-46%, respectivamente.

A formatação final das bases de dados (de treino e de teste) para ser submetida ao modelo neural proposto precisou ainda passar pela etapa de reorganização como matrizes 3D, sendo a altura representada pelos diferentes clientes, a largura representada pelas diferentes amostras temporais do cliente e a profundidade representada pelas variáveis.

Assim, considerando que a base de dados selecionada, que foi abordada na sub-seção 3.1, e o modelo de arquitetura proposto, que foi discutido na seção 4 e as transformações aplicadas sobre a base, que foram apresentadas nesta sub-seção, passaremos, a seguir, a detalhar os resultados experimentais da aplicação prática destes elementos essenciais, bem como sua comparação com resultados encontrados na literatura.

5 RESULTADOS

O modelo de NN proposto na seção 4 foi implementado na linguagem R[16] a qual oferece, além da linguagem de programação de propósito geral, um ambiente completo para computação voltada à análise estatística e as demonstrações gráficas. A versão 4.0.3 compilador/interpretador da linguagem R foi utilizada combinada ao IDE[17] Rstudio, na sua versão 1.4.1103, a qual provê acesso facilitado aos recursos do ecossistema R.

A estação de trabalho utilizada no desenvolvimento e coleta de resultados é caracterizada como um notebook da linha Macbook PRO equipado com processador Intel i7 Quad-Core de 2.8Ghz, 16GB memórias do tipo DDR3 de 1600Mhz, armazenamento de 512GB em SSD e placa de vídeo Intel Iris Pro 1.5 GB integrada e uma AMD Radeon R9 M370x via PCIe.

Uma das partes essenciais para realização do presente trabalho foi o uso de um vasto conjunto de bibliotecas R e Python[18] (com *wrappers*/cascas para R) providas pelo ecossistema R. Assim, as principais bibliotecas e suas respectivas finalidades estão relacionadas na Tabela 2.

---

16 Link para o site principal do projeto "*The R project for Statistical Compute*" - https://www.r-project.org/

17 IDE – acrônimo a partir do nome em inglês: *Integrated Development Environment*.

18 Linguagem de programação de alto nível, multi-paradigma, amplamente utilizada na área aprendizado de máquina.

A construção do modelo proposto se assenta sobre uma estrutura sequencial, oferecida pela biblioteca Keras (CHOLLET, 2017), listada na Tabela 2, a qual oferece flexibilidade e capacidade de expressão computacional robusta, mesmo para cenários complexos. A flexibilidade citada advém do fato que as camadas da CNN podem ser mescladas de forma arbitrária e cada camada possui um conjunto de parâmetros que podem ser ajustados dentre inúmeras possibilidades, a fim de adequar os modelos convolucionais construídos ao caso de uso específico.

Tabela 2: Relação de bibliotecas e suas finalidades

Biblioteca	Conjunto de funções para:
Caret	Treino e plotagem em modelos de regressão e classificação
Corrplot	Organização e plotagem de matrizes de correlação
Corrr	Explorar correlações entre variáveis
Ebimage	Processamento e análises de imagens
InformationValue	Análise de performance e classificação de modelos
Keras	API[19] de alto nível para criação de NN tanto RNN[20] quanto CNN
Lime	Explicar as decisões do modelo de NN para cada classificação
Proc	Visualização e comparação de curvas ROC[21]
Rocr	Visualização de performance de pontual para classificadores
Tensorflow	Computação numérica usando gráficos de fluxo de dados
Tfestimator	API para diferentes modelos de NN (neural e profunda)
Tfruns	Gerenciar o treinamento e otimização dos *hyper-parameters*
Tidyverse	Tratamento de diversos: gráficos, tipos de dados, matrizes, vetores
Yardstick	Qualificar o quão bem o modelo se adapta ao conjunto de dados

Fonte: Elaborada pelo autor.

Entretanto, fazer a escolha do valor a ser atribuído para cada parâmetro livre, de forma a otimizar a performance da CNN, para o caso de uso específico, é uma tarefa complexa visto que arranjo entre todas as combinações possíveis pode aumentar o desafio numa escala de ordens de grandeza. Assim, para realizar a otimização dos parâmetros livres de forma que fosse computacionalmente viável definiu-se, para cada parâmetro, um intervalo de 3 a 5 valores elegíveis, os quais são relacionados na Tabela 3, que foram enumerados a partir da literatura discutida na seção 2.

A permutação entre todas as possibilitadas foi realizada utilizando-se da biblioteca TFRUNS, a qual exercitou o modelo com cada uma das combinações de valores para os parâmetros livres, coletando seu resultado para, posteriormente, permitir a ordenação dos mesmos segundo critério de avaliação escolhido que, no

[19] API – acrônimo a partir do nome em inglês: *Application Programming Interface*

[20] RNN – acrônimo a partir do nome em inglês: *Recurrent Neural Network*

[21] ROC – acrônimo a partir do nome em inglês: *Receiver Operating Characteristic*

caso, foi a taxa de acertos (predições corretas) atingida pelo modelo. Ao final das permutações o melhor resultado geral para o modelo de NN proposto foi obido pelos valores: i) plDrop1 = 0.35; ii) plDrop2 = 0.35; iii) plUnDense1 = 128, iv) plUniDense = 96; v) plKnReg1 = 0.01; e vi) plKnReg2 = 0.01, conforme destacados na Tabela 3.

Tabela 3: Relação de parâmetros livres, finalidade e valores possíveis

Parâmetro	Finalidade	Valores
plDrop1	% de descarte do componente C5	[0.25, **0.35,** 0.40, 0.45]
plDrop2	% de descarte do componente C10	[0.20, 0.30, **0.35,** 0.40]
plUnDense1	Neurônios no componente C8	[60, 96, **128**, 196]
plUnDense2	Neurônios no componente C9	[32, 64, **96**, 128]
plKnReg1	% de regularização do componente C2	[**0.01,** 0.02, 0.04]
plKnReg2	% de regularização do componente C8	[**0.01**, 0.02, 0.03]

Fonte: Elaborada pelo autor.

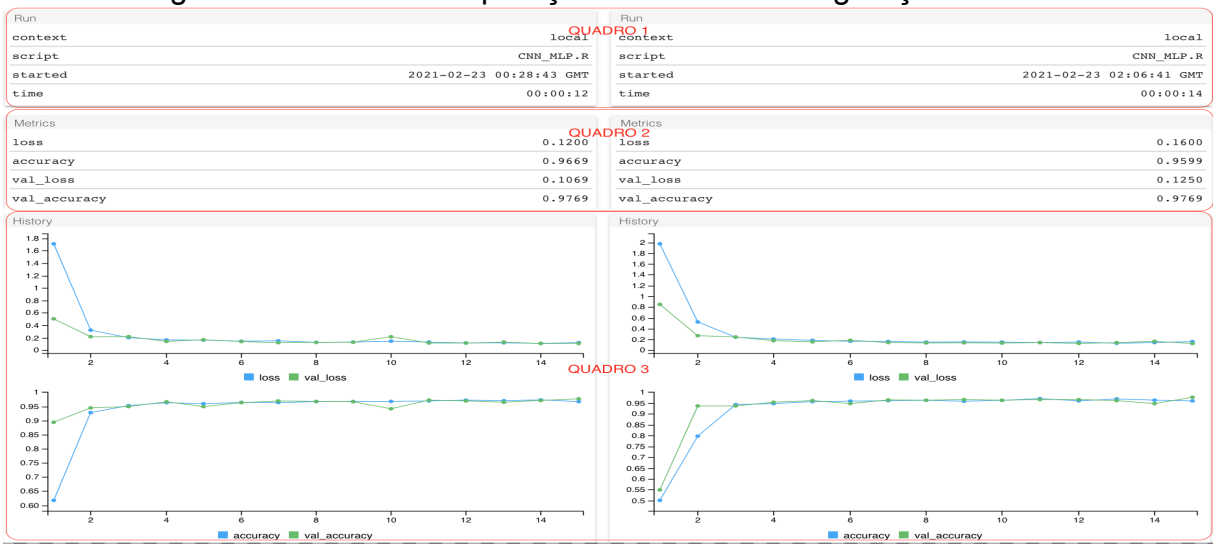
A biblioteca TFRUNS oferece, além disso, a possibilidade de exibir um painel comparativo entre duas instâncias de treino, executadas com diferentes valores para os parâmetros livres, de forma que as diferenças entre ambos possam ser percebidas visualmente. A Figura 3 apresenta um exemplo deste painel comparativo entre duas das 2304 permutações realizadas, para 15 épocas de treino, no qual pode-se observar a separação em três quadros (1, 2 e 3) os quais informam: i) dados do script executado; ii) resultados para as métricas de desempenho avaliadas; e iii) histograma para evolução do treino, separado em redução da perda (parte superior do quadro 3) e aumento da acurácia (parte inferior do quadro 3).

A base de dados utilizada (5600 casos) foi segmentada em dois conjuntos, um de treino e o outro de teste, com a distribuição de 60% a 40% entre eles, respectivamente, sendo que tal proporção foi definida com base no estudo apresentado por Onal e Karacuha (2018) e considerando adaptações necessárias para o tamanho e contexto dos dados aqui tratados. Além disso, para o processo de treinamento supervisionado da rede, que adota o mecanismo de *backpropagation*, foi utilizada a taxa de 21% das amostras (base de validação), parâmetro este definido através de processo experimental.

No processo de treinamento da rede foi utilizada a técnica conhecida como interrupção antecipada (referenciada em inglês como *early stopping* ou *early prunning*) na qual o treinamento é interrompido após um determinado número de épocas sem que a variável escolhida para medir a evolução do treinamento apresente melhoria. A varável escolhida para medir a evolução do treinamento foi a

“acurácia” (que será discutida nesta seção) e o modo indicativo de evolução foi o “max”, ou seja, seu valor deveria subir entre as épocas. A taxa de aprendizado também aplicou uma função monitoramento para que fosse alterada a cada 10 épocas sem evolução das métricas de avaliação, enquanto os demais parâmetros não foram alterados, assumindo os valores padrão.

Figura 3: Painel de comparação entre duas configuração de CNN



Fonte: Elaborada pelo autor.

Considerando as repetições realizadas para o processo de treinamento, embora houvessem algumas variações entre elas, em geral, após a 45 épocas o processo era interrompido. Assim, ao final do treinamento, foi possível obter os valores para as métricas de erro, o erro médio quadrático (RSME) e acurácia para os conjuntos de treino e de validação, conforme Tabela 4.

Tabela 4: Resultados para conjuntos de treinamento, validação e testes

Conjunto	Erro	RSME	Acurácia
Treino	5,58	1,03%	98,96%
Validação	10,14	-	97,08%
Teste	8,81	2,03%	97,96%

Fonte: Elaborada pelo autor.

Na próxima etapa foi realizada a avaliação do modelo proposto, para o conjunto de amostras de teste, quanto sua performance segundo as métricas apresentadas na literatura e, além disso, considerando o contexto no qual está inserido, ou seja, classificação binária para dados financeiros em séries temporais.

Logo, o primeiro passo tratou de capturar as mesmas métricas coletadas na etapa de treinamento, as quais também constam na Tabela 4.

A seguir optou-se pela métrica da matriz de confusão (**confmat**), pois com base nela é possível calcular diversas outras métricas estatísticas comumente encontradas na literatura (que serão discutidas nesta seção), visto que ela correlacionada os resultados previstos pelo modelo proposto frente aos dados reais apresentados. A Tabela 5 apresenta as combinações de valores possíveis para a matriz de confusão.

Tabela 5: Matriz de confusão

	Dados Reais	
Dados previstos	Verdadeiro (TP) Falso (TN)	Falso (FN) Verdadeiro (FP)

Fonte: Elaborada pelo autor.

A matriz apresentada na Tabela 5 permite contabilizar os casos reais verdadeiros (**TP**), os casos reais falsos (**TN**), os casos verdadeiros que foram incorretamente classificados como falsos (**FN**) e os casos falsos que foram incorretamente classificados como verdadeiros (**FP**). Além disso, se faz necessário contabilizar, dentro dos dados reais, número total de casos verdadeiros (**CV**) e casos falsos (**CF**). A matriz de confusão apresentada na Tabela 6 foi gerada a partir dos resultados obtido pelo modelo proposto, considerando parte da base de dados que compõe o conjunto de dados de teste. Nela é possível observar que o número total de casos “ruins” (CF) foi de 993, frente a 1069 casos “bons” (CV). O percentual TN foi de 97,58%, contra 98,31% para o TP e as taxas de erros FP e FN ficaram em 1,81% e 2,24%, respectivamente.

Tabela 6: Matriz de confusão para dados de teste

		Real	
		Ruins	**Bons**
Previsto	**Ruins**	969 (97,58%)	24 (2,24%)
	Bons	18 (1,81%)	1051 (98,31%)

Fonte: Elaborada pelo autor.

A partir destes elementos é possível avaliar algumas métricas encontradas na literatura como, por exemplo em Tai e Huyen (2018) e Provenzano (2020), para as

quais suas respectivas fórmulas e valores obtidos são apresentadas na Tabela 7. A métrica de precisão indica a taxa de acertos do modelo, considerando ambas as classes possíveis (“ruins” e “bons”). A acurácia indica qual a proporção dos “bons” que o modelo consegue prever. A especifidade indica qual a proporção de “ruins” que o modelo captura. A sensibilidade, que também é encontrada na literatura como “RECALL” ou “TPR”, indica a proporção dos “bons” descontando-se os falsos negativos, enquanto a FPR indica a proporção dos ‘ruins” descontando-se os falsos positivos e, por fim, a métrica F1 indica a média harmônica entre a capacidade de acertar as predições pra cada uma das classes.

Tabela 7: Métricas, fórmulas e os resultados obtidos com o modelo proposto

Métrica	Termo	Fórmula	Resultado
Precisão	**PREC**	= TP / (TP + FP)	**98,32%**
Acurácia	**ACC**	= (TP +TN) / (CV + CF)	**97,96%**
Especificidade	**SPEC**	= TN / (TN + FP)	**98,18%**
Sensitividade	**SENS**	= TP / (TP + FN)	**97,77%**
Recordação	**RECALL**	= SENS	**97,77%**
Taxa de positivos verdadeiros	**TPR**	= SENS	**97,77%**
Taxa de positivos falsos	**FPR**	= FP / (TN + FP)	**01,82%**
Média harmônica de PREC e SENS	**F1**	= (2 * PREC * ACC) / (PREC + ACC)	**98,14%**

Fonte: Elaborada pelo autor.

Outra métrica utilizada para avaliar os resultados de um modelo é conhecida como AUROC[22] a qual, segundo Provenzano (2020, p. 9), é composta por duas partes: i) ROC que representa uma curva de probabilidades; e ii) AUC que representa o grau de separabilidade entre as amostras de cada conjunto. O resultado de AUROC pode variar entre [0; 1] e indica o quanto o modelo em questão é capaz de diferenciar elementos entre as classes. Assim, quanto mais próximo a 1 for o resultado melhor será o modelo. Para calcular a curva ROC é necessário calcular a **TPR** e a **FPR**. O resultado obtido pelo modelo, para o conjunto de teste, foi **AUC = 0.9708 (97.08%)**.

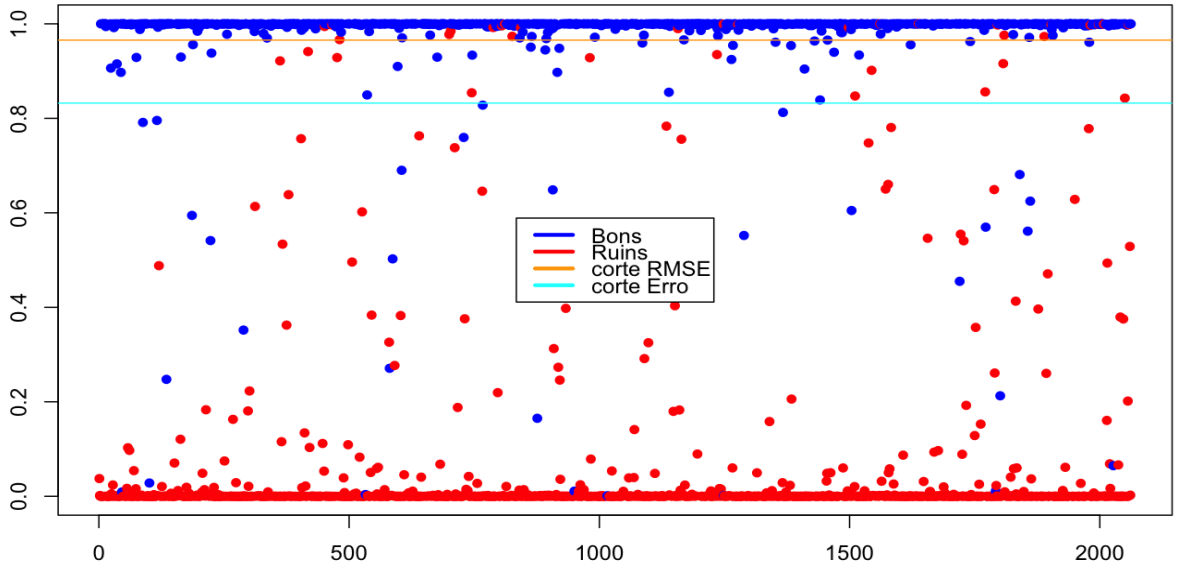
A capacidade do modelo proposto em classificar os clientes entre pagadores “bons” e “ruins” pode ser visualizada através de um gráfico que contemple a distribuição de probabilidades estimadas pelo modelo. A Figura 4 traz esse tipo informação para a qual os clientes corretamente classificados como pagadores “bons” são representados pelos pontos em azul, os quais possuem probabilidades próximas a 1 e, portanto, ficam na parte superior, enquanto os clientes classificados

[22] AUROC – acrônimo a partir do nome em inglês: *Area Under the Receiver Operating Characteristics*.

como pagadores “ruins”, representados pelos pontos em vermelho, e possuem probabilidades mais baixas, tendendo a ficar na parte inferior da imagem.

Além disso, há, na Figura 4, duas linhas horizontais que representam possíveis pontos de corte (*threshold)* para o classificador, sendo o primeiro (na cor laranja) dado do pela taxa de erro do modelo e o segundo (na cor ciano) pelo erro médio quadrático (EMQ). Essas linhas servem como referência sobre onde seria um ponto adequado para fazer a separação entre os conjuntos e, conforme os dados apresentados, denota-se que a utilização do EMQ pode oferecer um número menor de falsos positivos (FP), mediante um aumento no número de falsos negativos (FN).

Figura 4: Classificação das amostras do conjunto de teste



Fonte: Elaborada pelo autor.

A pontuação de KAPPA[23] é outra métrica encontrada na literatura, conforme trabalhos de Su (2019) e Kulkarni e Dhage (2019), porém parece haver uma divergência entre as variantes utilizadas em cada um destes, sendo que o primeiro parece utilizar a definição clássica, enquanto o segundo parece utilizar KAPPA máxima (K_{max}). A medição de KAPPA, segundo Chen (2019)[1], fornece uma “[...] indicação sobre a confiabilidade da avaliação realizada, ou seja, até que ponto os dados coletados no estudo são representações corretas das variáveis medidas”. No presente trabalho o valor de KAPPA foi calculado pelo pacote YARDSTICK, citado na Tabela 2, o qual adota o K_{max}, tendo resultado no valor de **0.9592 (95,92%)**, o qual,

[23] Cohen Kappa conforme https://en.wikipedia.org/wiki/Cohen%27s_kappa

segundo parâmetros de referência indicados por Chen (2019), está dentro do intervalo [0.81-1.0] que é classificado como “perfeito”.

Uma outra questão a ser abordada diz respeito a avaliação de distribuição de probabilidades para as variáveis da base de dados e qual sua significância estática para o teste realizado. Essas duas questões podem ser respondidas através da análise do KS[24] e dos P-values. O teste de KS, conforme Mosse (2020), permite avaliar o quão próximas são as distribuições de probabilidades para os valores de um conjunto de variáveis, enquanto os P-values indicam o quão confiável, ou estatisticamente significantes, são esses dados para o teste em questão. A escala de valores para ambos é definida no intervalo de [0-1], sendo que para o KS quanto maior o número mais diferentes serão as probabilidades enquanto que para os P-values quanto menor o número mais confiável serão os resultados. Os resultados obtidos pelo modelo aqui proposto foram de **KS_{max}=0,9608 (96,08%),** enquanto a máxima variação entre TPR e o FPR, definida **KS_{cutoff}**, foi de 0.4108 (**41,08%)** e o **P-values=2^{e-16}**, os quais denotam que as variáveis possuem elevada variação de probabilidades entre si (KS próximo a 1), ou seja, não são redundantes e que estes dados são estatisticamente significantes (P-values tendendo a zero).

Assim, concluímos a parte de apresentação dos resultados e, dessa forma, na próxima seção, será realizada uma discussão destes bem com sua comparação frente aos demais trabalhos encontrados na literatura.

6 DISCUSSÃO

Começaremos a discussão sobre os resultados por um dos pontos apontados na literatura como principal desvantagem das redes neurais profundas como, por exemplo, as redes com arquiteturas do tipo CNN que, segundo Provenzano (2020, p. 15), diz respeito a falta de clareza quanto ao caminho percorrido pela NN para cada classificação realizada, ou seja, sua árvore de decisões, olhando pela perspectiva humana, pode ser comparada a uma caixa-preta (KVAMMEA, 2018, p. 5). Entretanto, já existem alternativas promissoras para o tema, conforme discussões apresentadas em Molnar (2019, cap. 5), com a introdução de novos algoritmos, dentre os quais pode-se destacar: i) LIME[25]; ii) SHAP; e o iii) ALE. Os

[24] Teste Kolmogorov-Smirnov conforme de https://pt.wikipedia.org/wiki/Teste_Kolmogorov-Smirnov

[25] LIME – acrônimo a partir do nome em inglês: *Local Interpretable Model-agnostic Explanations (MOLNAR, 2019, cap 5.7)*

algoritmos LIME e SHAP são utilizados em Provenzano (2020, p. 16-20) para justificar os resultados do modelo lá proposto, visto que ambos oferecem explicações para as predições realizadas pelo modelo, caso a caso, baseando-se nas características avaliadas e nos valores da amostra de entrada.

No presente trabalho, aplicou-se o algoritmo de LIME, visando avaliar sua adequação ao explicar as decisões tomadas pelo modelo proposto. Entretanto, um dos principais desafios enfrentados para utilização do LIME, assim como das ferramentas para medir a correlações entre as variáveis em geral, recaiu sobre o modelo de dados utilizado, pois as bases de dados em questão, tanto a de treino como a de teste, foram formatadas como matrizes de 3D (ver seção 4.1), enquanto tais ferramentas trabalham com matrizes 2D (ou quadro de dados – *dataframes*). A alternativa encontrada para contornar tal situação passou por reformatar os conjuntos de dados selecionados para o formato 2D antes de fornece-los as ferramentas de análise, como o LIME. Na matriz 2D gerada neste processo cada cliente foi representado por uma única linha contendo todos as 20 variáveis multiplicadas por cada uma das 12 amostras temporais, ou seja, 240 “variáveis” para análise.

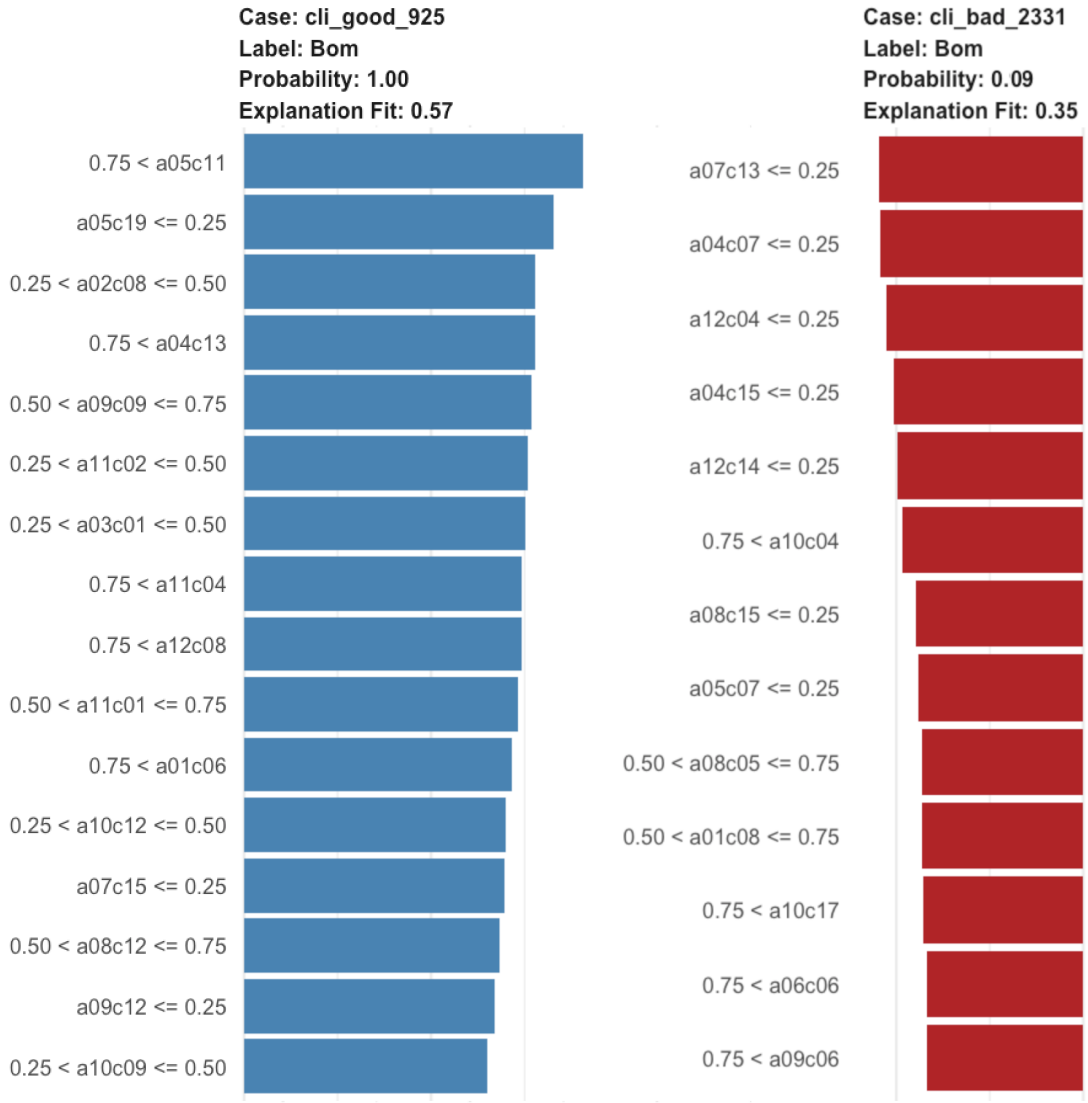
Uma vez superada a transformação dos dados para o formato aceito pelo LIME foi então realizada a avaliação considerando duas abordagens distintas oferecidas pelo LIME: i) características decisivas; e ii) mapa de calor; sendo que, para atender cada uma das abordagens, foram selecionados, de forma aleatória, a partir da base de teste, dois conjuntos distintos de clientes, um para cada avaliação.

No caso da abordagem que avalia as características decisivas foram utilizados, como conjunto de entrada, dois clientes, os quais foram submetidos ao LIME com o pedido para que fossem avaliadas as 16 características mais relevantes para classificar cada um dos clientes como pagador “BOM”. O resultado é apresentado na Figura 5, a qual está organizada em duas colunas, sendo que na coluna da esquerda constam os dados do cliente identificado como “cli_good_925”, o qual como o nome sugere, está classificado na base original como sendo um pagador “BOM”, enquanto na coluna da direta constam os dados do cliente identificado como “cli_bad_2331”, o qual está classificado na base original como sendo um pagador “RUIM”. Abaixo da identificação dos clientes constam outras três informações importantes: i) a classificação que está sendo avaliada, neste caso de ser “BOM” pagador; ii) a probabilidade do cliente em questão ser classificado

daquela forma; iii) o percentual que as características selecionadas impactam na decisão.

Na Figura 5, à esquerda das barras em azul e em vermelho, constam as características selecionadas para cada um dos clientes, assim como o condicional de decisão utilizado, enquanto que a largura das barras indica a escala de peso ponderado para cada característica. As barras na cor azul indicam que aquela característica oferece suporte a decisão pela classificação proposta, enquanto as barras na cor vermelha indicam oposição a classificação proposta.

Figura 5: Características mais relevantes para classificação



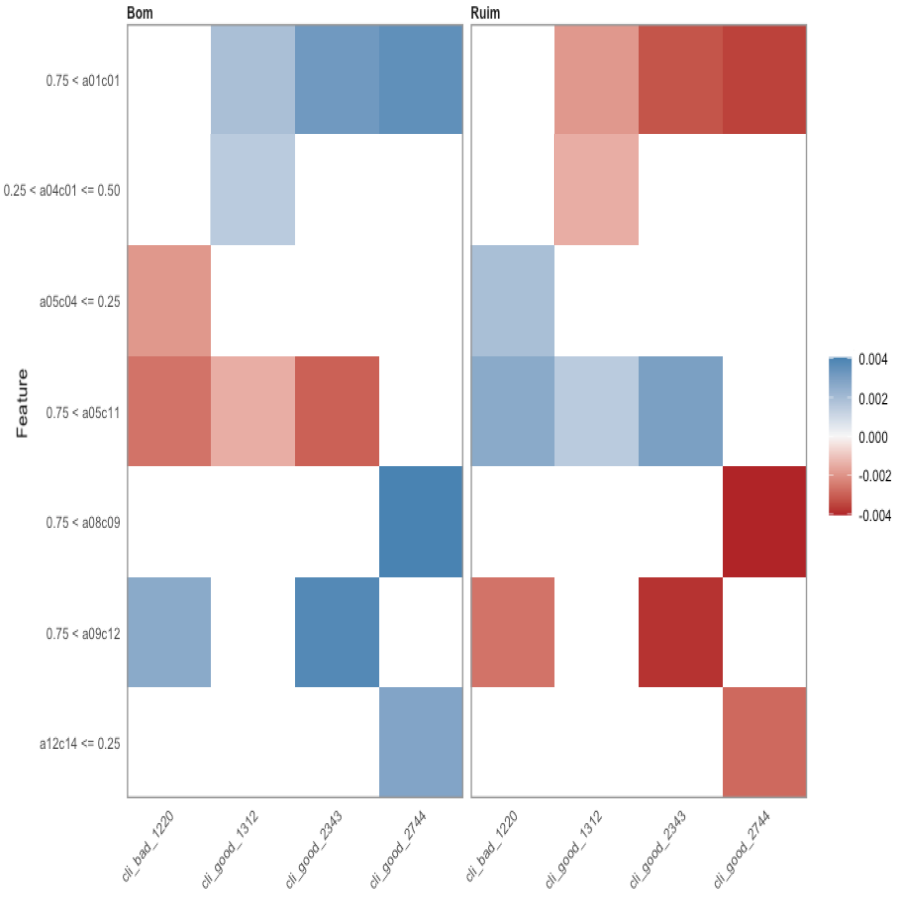
Fonte: Elaborada pelo autor.

Os dados apresentados na Figura 5 denotam que o cliente identificado como “cli_good_925” possui 100% de probabilidade de ser classificado como bom pagador e as 16 características ali apresentadas são responsáveis por 57% do peso total

nesta decisão, enquanto que, para o cliente identificado como “cli_bad_2131”, as respostas apontam para 0.09% de probabilidade do mesmo ser classificado como pagador “BOM” e as 16 características selecionadas são responsáveis por 37% do peso total nesta decisão.

No caso da abordagem do mapa de calor o LIME oferece uma visão que correlaciona o peso das características com sua capacidade de orientar suporte ou oposição a cada classificação possível que, neste caso, se restringem entre pagadores “bons” e “ruins”. Esta avaliação foi realizada para um conjunto de 4 clientes da base de testes, sendo que três destes constam classificados como pagador “BOM”, enquanto o outro está classificado como pagador “RUIM”. Além disso, foram selecionadas apenas 3 características para cada cliente.

Figura 6: Mapa de calor com classificação do pagador



Fonte: Elaborada pelo autor.

O resultado da avaliação, usando a abordagem do mapa de calor, é apresentado na Figura 6, a qual está organizada em dois blocos, sendo um à esquerda e outro à direita. O primeiro bloco traz as evidências que suportam ou

fazem oposição ao cliente ser classificado como pagador “BOM”, enquanto o segundo traz evidências que suportam ou fazem oposição ao cliente ser classificado como pagador “RUIM”. No eixo X da Figura 6 são apresentados os clientes com as respectivas identificações na parte inferior, enquanto no eixo Y são relacionadas as características selecionadas para cada cliente com sua condicional de decisão.

A cor azul nos retângulos da Figura 6 indica que a característica em questão está oferecendo suporte para classificação do bloco, enquanto os retângulos na cor vermelha indicam oposição a classificação do bloco. A intensidade das cores, azul e vermelho, representa o peso ponderado daquela característica.

É importante destacar que as características apresentadas no eixo Y das Figuras 5 e 6 estão identificadas com um padrão de nomes dado por: “aAAcCC”. Neste padrão a letra “a” indica a amostra temporal considerada, a qual que pode ter 12 variações representadas pelo trecho “AA”, com valores possíveis no intervalo de [01-12]; continuando com a letra “c” que indica a variável tratada, a qual pode assumir 20 valores no intervalo de [01-20], representadas pelo trecho “CC”. Esse padrão de nomes de variável é decorrente da transformação que foi aplicada nas matrizes 3D para 2D, descritas anteriormente. O texto adicional apresentado junto ao nome da varável denota os condicionais que foram utilizados para a tomada de decisão e, a combinação destes elementos, gera a característica analisada.

A utilização do LIME demonstra que existem ferramentas que conseguem abrir a “caixa-preta” dos modelos gerados em redes neurais profundas. Contudo, durante a fase experimental foram realizadas diversas iterações com o LIME, as quais expuseram variações significativas entre as explicações fornecidas e, além disso, algumas explicações fornecidas podem não ser adequadas como, por exemplo, no caso da Figura 6 que os clientes identificados como “cli_bad_1220” e “cli_good_1312” para os quais, num conjunto de apenas três características avaliadas uma delas apresentou oposição a classificação proposta pelo modelo, sendo elas: “a09c12” no primeiro caso e “a05c11” para o segundo. Esta falta de consistência do LIME é uma de suas deficiências, conforme citado em Molnar (2019, cap. 5.7.5).

Os resultados experimentais introduzidos na seção 5 parecem promissores, porém é fundamental coloca-lós em perspectiva, através de uma comparação com os resultados obtidos por outros trabalhos encontrados na literatura, para determinar as vantagens e desvantagens em cada métrica avaliada. A base de comparação

pesquisada contempla 18 trabalhos e cada trabalho traz uma série de variações que vão desde a base de dados utilizada até as métricas de resultado escolhidas, fato que dificulta a comparação direta dos resultados.

Assim, a Tabela 8 relaciona cada um dos trabalhos comparados, o setor de aplicação dos estudos e a base de dados adotada. Além disso, demostra que os trabalhos selecionados para comparação são, em sua absoluta maioria, voltados ao setor financeiro e utilizam bases de dados de instituições financeiras de diversos lugares do mundo.

Tabela 8: Trabalhos selecionados, setor de aplicação e origem dos dados

Autor	Setor	Base de dados
West (2000)	Financeiro	German Bank
Šušteršicˇ (2009)	Financeiro	Sloveniam Bank
Correa et al. (2011)	Financeiro	Colpatria Bank
Oreski (2012)	Financeiro	Croatian Bank
Doori (2014)	Financeiro	Australian Bank
Gante (2015)	Financeiro	German Bank
Kammoun e Triki (2015)	Financeiro	Tunisian Bank
Gudelek (2017)	Financeiro	Forecast – StockPrice[26]
Sayjadah (2018)	Financeiro	Credit Card
Kvamme (2018)	Financeiro	Noruega Bank – DNB
Tang (2018)	Financeiro	Australian Bank
Gupta e Goyal (2018)	Financeiro	Kaggle Lends
Kulkarni (2018)	Financeiro	Bank + Social Networks
Tai e Huyen (2019)	Financeiro	German Bank
Napitupulu (2019)	Financeiro	Similar German Bank
Su (2019)	Financeiro	MCC – Categorias de serviços[27]
Provenzano (2020)	Contábil	Credit Research Database (Moody's)
Golbayani (2020)	Contábil	Credit Corporativo (Standard & Poor's)

Fonte: Elaborada pelo autor.

Contudo, existem algumas exceções, como, por exemplo, o trabalho de Gudelek (2017) que utiliza como base de dados a movimentação de preços de

[26] Base de dados utiliza séries temporais de movimentação do preço de ativos e moedas

[27] Base de dados utilizada e enfoque da pesquisa diverge das demais (baixa similaridade)

ativos e moedas no mercado de capitais; o trabalho de Su (2019) que avalia a qualidade dos créditos com base no tipo de serviço financeiro prestado através de uma tabela internacional padronizada (MCC); os trabalhos de Provenzano (2020) e Golbayani (2020), por sua vez, utilizam-se de demonstrações contábeis de grandes empresas a partir das bases de dados de agências de classificação de risco (Standard & Poor´s e Moody´s).

Uma vez qualificados os trabalhos a serem comparados passaremos as métricas de avaliação de desempenho. Nesse aspecto não foi possível encontrar um consenso ou forte tendência, pois cada autor utiliza um subconjunto de métricas distintas. O presente trabalhou coletou 17 métricas de desempenho, que foram abordas na seção 5, as quais cobrem a maioria absoluta das métricas encontradas na literatura. Contudo, a quantidade média de métricas apresentadas pelos trabalhos citados na Tabela 8 fica em torno de 4 métricas, ou seja, uma avaliação e comparação profunda entre a presente proposta e as demais fica prejudicada.

Outro aspecto que deve ser destacado é que alguns dos trabalhos citados na Tabela 8 discutem mais do que uma proposta arquitetural e, portanto, trazem dados de desempenho para cada uma delas. Assim, se consideradas todas as alternativas, seriam necessárias 54 comparações. Contudo, os autores já fazem a comparação entre suas propostas; então optou-se por apresentar as duas melhores propostas de cada trabalho, visto que uma proposta pode ser melhor em determinada métrica mas pior em outras e vice-versa.

Considerando que, ainda assim, seria necessário uma tabela de 17 colunas, uma para cada métrica, por 34 linhas, uma pra cada trabalho/proposta, tal formato prejudicaria a interpretação dos resultados, tanto pela sobrecarga cognitiva quanto em razão da disparidade, visto que parte significativa das colunas não teria resultados. Logo, as comparações foram segmentadas em três grupos de tabelas, os quais foram definidos com base na similaridade das métricas apresentadas. Assim, o primeiro grupo de comparação, retratado na Tabela 9, abrange apenas duas propostas mencionadas na Tabela 7, as duas mais antigas, para as quais foi possível extrair apenas número absoluto de falsos positivos (FP), o número total de falsos negativos (FN) e o indicador de acurácia (ACC), todos normalizados para termos percentuais.

Conforme métricas apresentadas na Tabela 9 fica evidente a superioridade da presente proposta, nas 3 métricas exibidas. Além disso, destaca-se os elevados

percentuais de FN e FP em ambas as propostas quanto comparadas ao presente trabalho, sendo até 9 vezes pior para o FP e até 21 vezes pior para o caso do FN, enquanto o indicador de ACC mostra divergência de quase 20 pontos percentuais.

Tabela 9: Comparação de resultados – Grupo I

Autor	Técnica	FP[28]	FN[28]	ACC[29]
Esta proposta[30]	**CNN2D**	**1,82**	**2,24**	**97,96**
West (2000)	RBF	14,24	48,21	75,30
West (2000)	LR	11,86	51,33	76,30
Šušteršicˇ (2009)	MLP (EB04)	17,80	29,90	79,30
Šušteršicˇ (2009)	LR (LOGIT01)	13,30	37,70	76,10

Fonte: Elaborada pelo autor.

No agrupamento seguinte foram reunidas, conforme Tabela 10, as propostas que informaram dados relacionadas a tabela de confusão (**confmat**) ou diretamente as métricas derivadas desta, as quais são mencionadas na Tabela 7, bem como a métrica **AUC**, presente na maioria destas prospostas. Sobre os dados apresentados na Tabela 10 cabe, primeiro, destacar que os mesmos foram normalizados em percentuais, visto que a escala de valores é a mesma [0; 1] para todos. O segundo ponto a ser ressaltado é que os resultados do presente trabalho superam todas as demais propostas, em todas 7 métricas apresentadas nesta tabela, excetuando-se a proposta de Kvamme (2018) que é superior em relação a métrica **SPEC**, porém é inferior nas demais.

O terceiro ponto que merece atenção é que o presente trabalho traz um nítido equilíbrio entre os resultados de cada métrica, enquanto os demais possuem variações significativas entre elas, o que implica que a implementação daquelas propostas teve que fazer uma escolha (*tradeoff*) para melhorar uma determinada métrica impactando negativamente em outra como, por exemplo, melhorar a assertividade do TP relaxando o ponto de corte, escolha que resulta em piora no FP.

Ainda, sobre o terceiro ponto supracitado, dito de outra forma, para garantir que os clientes que são bons pagadores sejam reconhecidos como tal o modelo relaxa sua seletividade e, em contra-partida, começa a classificar erroneamente (métrica FPR – erro tipo 1) clientes que são maus pagadores como se bons fossem.

[28] Quanto menor melhor
[29] Quanto maior melhor

Aliás, não por acaso, é incomum encontrar na literatura trabalhos que tragam essa métrica e, nesse sentido, destaque para as propostas de Oreski (2012) e Provenzano (2020) que informam FPR de 23,20%, 15% e 9,00%, respectivamente, porém todas acima do valor de **1,82%** (FPR quanto menor melhor) obtido no presente trabalho.

Tabela 10: Comparação de métricas – Grupo II

Autor	Técnica	PREC	ACC	SPEC	SENS	FPR	F1	AUC
Esta proposta [30]	CNN2D	**98,32**	**97,96**	98,18	**97,77**	**1,82**	**98,14**	**97,08**
Tai e Huyen (2019)	DSNN	81,00	75,90		82,00		81,50	
	CNN	79,00	79,00		80,00		79,50	
Provenzano[30] (2020)	MCONF1	85,71	87,50	85,00	90,00	15,00	86,60	95,00
	MCONF2	90,11	86,50	91,00	82,00	9,00	88,27	95,00
Kvamme (2018)	CNN		95,40	**98,50**	37,40			91,50
	LR		91,00	93,50	49,00			86,40
Oreski[30] (2012)	MLP	90,41	73,90	76,80	72,93	23,20	81,33	
Kammoun e Triki (2015)	LR		72,00	20,50	93,80			
	MLP		71,67	78,41	68,87			
Tang (2018)	PNN		85,64	91,10	94,84			94,11
	MLP		84,23	77,89	90,63			89,76
Su (2019)	RF		88,03	96,50	76,93		84,50	
	XGBoost		88,27	96,72	77,77		85,18	

Fonte: Elaborada pelo autor.

O último grupo de comparação a ser discutido, conforme Tabela 11, reune todas as demais propostas que não foram selecionadas para os grupos anteriores. Um fato curioso sobre este último grupo é que dentre as 6 métricas apresentadas não há nenhuma que seja comum a todas as propostas. Logo, tal característica evidencia e confirma a diversidade entre as métricas utilizadas. Contudo, se consideras todas as propostas citadas, nos 3 grupos de comparativos (Tabelas 9, 10 e 11), pode-se observar que a métrica **ACC** está presente em todas, exceto no trabalho de Correa et al. (2011).

Novamente, considerados os dados comparáveis, ou seja, as colunas preenchidas, os resultados do presente trabalho superam os demais. A proposta

[30] Resultados apresentados pela proposta trazem a matriz de confusão (confmat) completa, possibilitando o cálculo e conferência dos resultados.

apresentada por Gupta e Goyal (2018) também merece destaque, pois foi a segunda melhor, ficando muito próxima à primeira. Infelizmente o autor deste só oferece duas métricas comparáveis, sendo elas ACC e RSME, e há pouco detalhamento sobre a arquitetura desenvolvida, além de ser o único a utilizar a base de dados Kaggle.

Tabela 11: Comparação de resultados – Grupo III

Autor	Técnica	ACC	SPEC	SENS	AUC	KAPPA	RSME
Esta proposta[30]	CNN2D	**97,96**	**98,18**	**97,77**	**97,08**	**95,92**	**2,04**
Correa et al. (2011)	MLP(GA)				71,25		
	LR				65,92		
Doori (2014)	MLP	90,62					
Kammoun e Triki (2015)	LR	72,00	20,50	93,80			
	MLP	71,67	78,41	68,87			
Gante (2015)	NN3D	80,00					4,00
Tang (2018)	PNN	85,64	91,10	94,84	94,11		
	MLP	84,23	77,89	90,63	89,76		
Gupta e Goyal (2018)	LR	97,69					2,27
	ANN	97,67					2,20
Golbayani (2020)	LSTM	89,84					
	CNN2D	90,13					
Gudelek (2017)	LR	71,72					23,61
	MLP	78,46					
Sayjadah (2018)	LR	82,00			75,00		
	RDT	82,06			64,00		
Napitupulu (2019)	MLP	83,60					
Kulkarni (2018)	NB	84,64				19,00	36,00
	RF	84,47				19,00	37,00

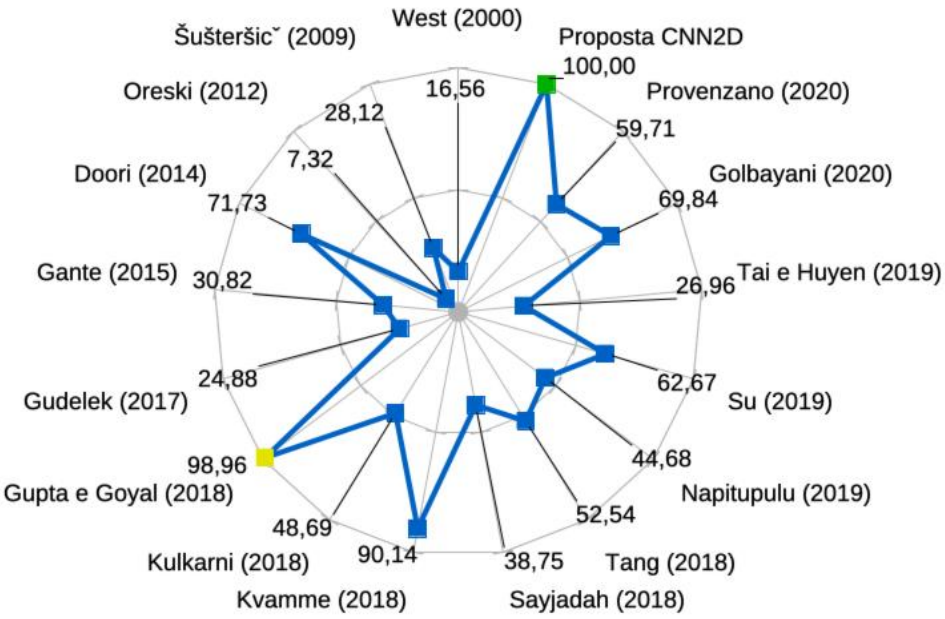
Fonte: Elaborada pelo autor.

Considerando que a representação gráfica dos resultados facilita, em geral, a visualização das diferenças entre os objetos de comparação foi elaborado o gráfico introduzido pela Figura 7, o qual foi construído a partir da métrica **ACC**, por ser está a métrica disponível em todas as propostas avaliadas – exceto Correa et al. (2018). Além disso, para cada trabalho, foi selecionado apenas o melhor resultado da métrica e, para destacar a representatividade da variação desta, num conjunto de amostras onde o valor oscila entre [72; 97,96], foi aplicado um processo de

normalização pela diferença entre os dois extremos, de forma que os trabalhos tiveram seus valores escalonados dentro da variação percentual deste intervalo.

Na Figura 7 a proposta desenvolvida neste trabalho está indicada como "**proposta CNN2D**", qual obteve a melhor pontuação quanto a métrica ACC, tendo sido destacada através de um retângulo verde, enquanto a segunda foi destacada com um retângulo amarelo.

Figura 7: Resultados métrica ACC normalizados para o intervalo de variação



Fonte: Elaborada pelo autor.

A discussão realizada teve por objetivo apresentar os prós e contras da solução proposta, bem como posicionar a mesma frente aos seus pares. Na próxima, será realizado o fechamento deste trabalho com as considerações finais e sugestões para trabalhos futuros..

6.1 Posicionando a realidade interna na discussão

Seção omitida da versão pública para manter o sigilo das informações relativas a fonte interna.

7 CONSIDERAÇÕES FINAIS

Este trabalho navegou através do oceano de nuances presentes no estado da arte das técnicas de transformação de dados e aprendizado de máquina sendo aplicadas a tarefa de classificação de risco de crédito. A metodologia de desenvolvimento mesclou pesquisa qualitativa, através de entrevista semi-estruturada, pesquisa quantitativa através de análise documental, extensa revisão bibliográfica, que direcionou a seleção da base de dados e ajudou a responder alguns objetivos específicos, além de ter direcionado a pesquisa exploratória, última etapa, na qual foi construído um modelo de CNN que se propõe a responder algumas das questões suscitadas.

O primeiro objetivo específico, que buscava responder se era melhor utilizar todas variáveis disponíveis ou apenas um subconjunto delas, foi respondido na fase de revisão da literatura, pois trabalhos anteriores já haviam tratado desta questão e apontado que para as técnicas de aprendizados de máquina a utilização de todas as variáveis disponíveis tenderia a oferecer os melhores resultados.

O segundo objetivo específico, que questionava qual seria a técnica mais efetiva, entre as alternativas de regressão logística e técnicas de aprendizado de máquina, também foi respondida durante a fase de revisão da literatura onde outros autores já apontavam que, apesar de similares, para o cenário de dados financeiros em séries temporais as técnicas de aprendizados de máquina ofereceriam resultados superiores.

A base de dados selecionada para uso neste trabalho precisou passar por série de transformações para se adequar ao formato esperado pela arquitetura de CNN ora proposto, sendo que, para tanto, foram exploradas diversas técnicas como, por exemplo: sanitização, normalização, equalização do tamanho de amostras entre os conjuntos, a transformação em séries temporais e, por fim, a conversão em pseudo-arquivos de imagens no formato de matrizes bidimensionais (2D).

A arquitetura de NN proposta foi construída a partir das observações oriundas da fase de revisão bibliográfica e refinada através de um processo iteração e refinamento incremental. Além disso, seus principais parâmetros de operação foram selecionados após execução de 2304 permutações, entre os valores possíveis, de forma automatizada, utilizando-se da biblioteca TFRUNS.

Na etapa de teste e avaliação foram coletadas 17 métricas, as quais foram descritas e sua aplicabilidade foi exemplificada, sendo que esse conjunto tão abrangente foi escolhido visando possibilitar comparação objetiva frente aos pares encontrados na literatura. A comparação direta frente aos pares ficou prejudica pelo limitado número de métricas apresentadas pelos demais, contudo, mesmo assim, o presente trabalho ficou no topo de todas as métricas de avaliação. O grande destaque da presente proposta foi o equilíbrio observado entre as diferentes métricas, fato incomum na literatura e serve como indicativo que as transformações aplicadas sobre a base foram efetivas, a ponto de evitar os problemas de sub-treinamento e supertreinamento, minimizando efeitos de viés em relação ao conjunto de dados de treino.

Assim, pode-se considerar que abordagem para transformação de dados financeiros, relativos ao histórico de crédito de uma base de clientes, em imagens bidimensionais que foram fornecidas como fonte de entrada para a CNN2D, a qual realizou a classificação dos mesmos entre pagadores bons e ruins, foi exitosa, considerando o resultado das métricas a seguir: **PREC**=98,32%, **ACC**=97,96%, **SPEC**=98,18%, **SENS**=97,77%, **FPR**=1,82%, **F1**=98,14% e **AUC**=97,08%, **KAPPA**=95,92, **KS**=96,08, **RSME**=2,04 e **GINI**=94,16.

Além disso, os resultados do modelo proposto, frente a classificação de um conjunto de clientes, foram explicados através de uma ferramenta denominada LIME, a qual demonstrou ser possível abrir a “caixa-preta” dos modelos neurais profundos e explicar suas decisões. Contudo, o estágio de desenvolvimento do LIME ainda precisará amadurecer para gerar resultados consistentes e melhorar sua confiabilidade.

Contudo, como nem tudo são flores, algumas questões ficarão em aberto poderão servir como guia de continuidade ou trabalhos futuros. Um ponto que pode ser lapidado diz respeito as técnicas de explicação das decisões da NN e, para isso, uma das alternativas seria avaliar algoritmos alternativos como SHAP ou ALE.

REFERÊNCIAS

ALICE, Michy. How to perform a Logistic Regression in R. 2015a. Disponível em: https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r. Acesso em: 5/12/2019.

ALICE, Michy. Fitting a Neural Network in R: 2015b. Disponível em: https://rpubs.com/adriensieg/311842. Acesso em: 5/12/2019.

CASTRO NETO, J. L. de; SÉRGIO, R. S. G. **Análise de Risco e Crédito** [s. l.]: Curitiba: IESDE Brasil S.A., 2009. ISBN: 978-85-387-1997. Disponível em: https://www.passeidireto.com/arquivo/22247659/livro-analise-de-risco-e-credito. Acesso em: 5/12/2019. ISBN: 978-85-387-1997-7.

CHEN, Yingting Sherry. Towards Data Science: A Medium publication sharing concepts, ideas and codes. In: **Interpretation of Kappa Values**: Evaluate the agreement level with condition. [S. l.]: Ben Huberman, 6 jul. 2019. Disponível em: https://towardsdatascience.com/interpretation-of-kappa-values-2acd1ca7b18f. Acesso em: 14 jan. 2021

CHOLLET, François *et al*. **R Interface to Keras**. 1.4. [S. l.]: GitHub, 2017. Disponível em: https://github.com/rstudio/keras. Acesso em: 3 nov. 2020.

CORREA, A. B.; GONZALEZ, A. M.; COLPATRIA, Banco. **Evolutionary algorithms for selecting the architecture of a MLP Neural Network**: A Credit Scoring Case. The IEEE International Conference on Data Mining series (ICDM), Vancouver, Canada. 2011.

DOORI, Mulhim Al; BEYROUTI, Bassam. Credit Scoring Model Based on Back Propagation Neural Network Using Various Activation and Error Function: Mulhim Al Doori and Bassam Beyrouti. **JCSNS International Journal of Computer Science and Network Security**, [s. l.], v. 14, ed. 3, p. 16-24, 20 mar. 2014. ISSN: 1738-7906. Disponível em: http://paper.ijcsns.org/07_book/201403/20140303.pdf. Acesso em: 24 jun. 2020.

GANTE, Dionicio D.; GERARDO, Bobby D.; TANGUILIG, Bartolome T. NEURAL NETWORK MODEL USING BACK PROPAGATION ALGORITHM FOR CREDIT RISK EVALUATION. **Proceeding of the 3rd International Conference on Artificial Intelligence and Computer Science (AICS2015)**, Penang, MALAYSIA, ano 1, v. 1, n. 3, p. 93-104, 13/10/2015. 3rd International Conference on Artificial Intelligence and Computer Science, 2015, Penang, MALAYSIA. e-ISBN: 978-967-0792-06-4.

GOLBAYANI, Parisa; WANG, Dan; e FLORESCU, Ionut. Application of Deep Neural Networks to assess corporate Credit Rating. Preprint. [arXiv:2003.02334 [q-fin.RM]]. **arXiv:2003.02334 [q-fin.RM]**, Cornell University Library, 2020.

GOLLAPUDI, S. **Practical Machine Learning**. Birmingham, UK. 2016. ISBN: 978-1-78439-968-9.

GUDELEK, Ugur; BOLUK, Arda; e OZBAYOGLU, Murat. A deep learning based stock trading model with 2-D CNN trend detection. E**EE Symposium Series on Computational Intelligence (SSCI)**, Honolulu, HI, 2017, pp. 1-8, DOI: 10.1109/SSCI.2017.8285188.

GUPTA, D. K; GOYAL, S. **Credit Risk Prediction Using Artificial Neural Network Algorithm**. I.J. Modern Education and Computer Science, p. 5-9. 2018, 5, 9-16. Disponível em http://www.mecs-press.net/ijmecs/ijmecs-v10-n5/IJMECS-V10-N5-2.pdf. Acesso: 15/01/2020. DOI: 10.5815/ijmecs.2018.05.02. 2018.

KAMMOUN, A.; TRIKI, I. Credit Scoring Models for a Tunisian Microfinance Institution: Comparison between Artificial Neural Network and Logistic Regression. **Review of Economics & Finance**. Better Advances Press, Canada, Fevereiro 2016, vol. 6, p 61-78. Disponível em: http://www.bapress.ca/ref/ref-article/1923-7529-2016-01-61-18.pdf. Acesso em: 12/12/2019.

KVAMME, Håvard et al, Predicting mortgage default using convolutional neural networks. **Expert Systems with Applications**, v. 102, pp. 207-217, 2018. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2018.02.029.

MAYS, E. **Handbook of Credit Scoring**. Chicago, New York: The Glenlake Publishing Company Ltd, 2001.

MOLNAR, Christoph. ***Interpretable Machine Learning****: A Guide for Making Black Box Models Explainable*. 1. ed. Internet: Lulu.com, 2019. 318 p. v. 1. ISBN 9780244768522. Disponível em: https://christophm.github.io/interpretable-ml-book/. Acesso em: 17 fev. 2021.

MOSSE, Billy. Towards Data Science: A Medium publication sharing concepts, ideas and codes. In: **Why you may be getting low test accuracy: try Kolmogorov-Smirnov**. [S. l.]: Ben Huberman, 18 maio 2020. Disponível em: https://towardsdatascience.com/why-you-may-be-getting-low-test-accuracy-try-this-simpstatistical-tests-30585b7ee4fa. Acesso em: 31 out. 2020.

MUHAMMAD, Yaqub et al. State-of-the-Art CNN Optimizer for Brain Tumor Segmentation in Magnetic Resonance Images. **Brain Sciences**, [s. l.], ano 2020, v. 10, n. 427, ed. 7, p. 1-20, 3 jul. 2020. ISSN: 2076-3425. DOI 10.3390/brainsci10070427. Disponível em: https://www.mdpi.com/2076-3425/10/7/427. Acesso em: 16 nov. 2020.

NAPITUPULU, T.; TRIANA, D. **Measuring Credit Risk Of New Customer Using Artificial Neural Network Model**: A Case Of Multi Finance In Indonesia. International Journal Of Scientific & Technology Research vol. 8, ed. 10, Outubro 2019.

ONAL, Nisa Ozge; KARACUHA, Ertugrul. Novel Approaches on Sovereign Credit Ratings. **EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS**, New York, ano 2018, v. 11, n. 4, 23 jul. 2018. Mathematics, p. 1014-1026. DOI 10.29020/nybg.ejpam.v11i4.3333. Disponível em: https://www.ejpam.com/index.php/ejpam/article/view/3333. Acesso em: 17 jun. 2020.

ORESKI, Stjepan; ORESKI, Dijana; ORESKI, Goran. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. **Expert Systems with Applications**, [s. l.], v. 39, ed. 16, p. 12605-12617, 15/11/2012. ISSN 0957-4174. DOI 10.1016/j.eswa.2012.05.023. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S095741741200721X?via%3Dihub. Acesso em: 7 set. 2020.

PROVENZANO, Angela et al. Machine Learning approach for Credit Scoring. Preprint. [arXiv:2008.01687v1 [q-fin.ST]]. **arXiv:2008.01687v1 [q-fin.ST**], Cornell University Library, 2020.

SAYJADAH, Yashna. et al, Credit Card Default Prediction using Machine Learning Techniques. 2018 **Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)**, Subang Jaya, Malaysia. 2018. pp. 1-4. DOI: 10.1109/ICACCAF.2018.8776802.

SHARMA, Alok et al. Deepinsight: A methodology to transform a non-image data to an image for convolution neural network architecture. **Nature**: Scientific Reports, United Kingdom, ano 2019, v. 1, n. 11399, ed. 9, 6 ago. 2019. DOI 10.1038/s41598-019-47765-6. Disponível em: https://www.nature.com/articles/s41598-019-47765-6. Acesso em: 26 jul. 2020.

SRIVASTAVA, Nitish *et al*. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. **Journal of Machine Learning Research**, [s. l.], ano 15, v. 15, n. 56, p. 1929-1958, 15 jun. 2014. ISSN: 1929-1958. Disponível em: http://jmlr.org/papers/v15/srivastava14a.html. Acesso em: 14 set. 2020.

SU, C-H. et al. A *Ensemble Machine Learning Based System for Merchant Credit Risk Detection in Merchant MCC Misuse*. **Journal of Data Science**, P. 81 – 106. 2019. DOI:10.6339/JDS.201901_17(1).0004.

ŠUŠTERŠICˇ, Maja; MRAMOR, Dušan; ZUPAN, Jure. Consumer credit scoring models with limited data. **Expert Systems with Applications**, [s. l.], ano 2009, v. 36, n. 1, ed. 3, p. 4736–4744, 13 abr. 2009. DOI:10.1016/j.eswa.2008.06.016. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0957417408002996. Acesso em: 26 set. 2020.

TAI, Le Quy e HUYEN, Thi Thu. *Deep Learning Techniques for Credit Scoring*. **Journal of Economics, Business and Management**, v. 7, n. 3, pp. 93-96, 2019. ISSN: 2301-3567. DOI: 10.18178/joebm.2019.7.3.588.

TANG, Yajiao et al. A Pruning Neural Network Model in Credit Classification Analysis. **Computational Intelligence and Neuroscience**, [s. l.], ano 1, v. 2018, n. 9390410, p. 1-22, 11 fev. 2018. DOI 10.1155/2018/9390410. Disponível em: https://www.hindawi.com/journals/cin/2018/9390410/. Acesso em: 8 nov. 2020.

WEST, D. *Neural network credit scoring models*. **Computers & Operations Research**. Greenville, USA. 2000. Disponível em: https://www.researchgate.net/publication/223425357_Neural_Network_Credit_Scoring_Models. Acesso em: 5/03/2020.

YAN, Carson. Convolutional Neural Network on a structured bank customer data. **Towards Data Science**. Set. 2018. Disponível em: https://towardsdatascience.com/convolutional-neural-network-on-a-structured-bank-customer-data-358e6b8aa759, Acesso em: 17/08/2020.

APÊNDICE A – ENTREVISTA SEMI-ESTRUTURADA

Projeto Aplicado	Aplicação de técnicas de aprendizado de máquina na construção de modelos de risco de crédito
Finalidade	Conhecer o modelo de risco utilizado atualmente no Banrisul e identificar pontos de melhoria
Data aplicação	17 de Dezembro de 2019
Entrevistado	---

Questionário omitido para preservar a fonte na versão pública.

APÊNDICE B – DADOS FINANCEIROS POR CLIENTE

Quadro 1: Cliente 2436 (classe Ruim)

Dados normalizados:

```
 0.33 0.38 1.00 0.0 0.22 0 0.00 1.00 1.00 0 0.33 0.67 0.16 1 0.5 0.33 1.00 1 0 1
 1.00 0.21 1.00 0.0 0.14 0 0.75 0.33 0.67 0 0.33 0.33 0.21 0 0.5 0.33 0.67 1 0 1
 1.00 0.47 0.75 0.9 0.51 0 0.25 0.00 0.00 0 0.00 0.67 0.16 1 0.5 0.33 0.67 1 0 1
 0.33 0.12 0.50 0.3 0.06 0 0.25 0.33 0.67 0 0.00 0.67 0.05 1 0.5 0.00 0.67 1 0 1
 0.00 0.65 0.50 0.3 0.37 0 0.75 0.00 1.00 1 0.00 0.00 0.26 1 0.5 0.33 0.67 1 1 1
 0.67 0.17 0.50 0.3 0.11 0 0.25 0.33 0.33 0 0.67 0.00 0.10 1 0.5 0.00 0.33 1 0 1
 1.00 0.21 0.50 0.5 0.09 0 0.25 1.00 0.33 0 1.00 0.00 0.07 1 0.5 0.00 0.67 1 0 1
 0.33 0.29 1.00 0.2 0.25 0 0.25 0.33 0.33 0 1.00 0.67 0.10 0 0.5 0.00 0.33 1 0 1
 0.33 0.29 0.50 0.1 0.62 0 0.50 0.00 0.33 0 1.00 0.67 0.07 1 0.0 0.33 1.00 1 0 1
 0.33 0.29 1.00 0.2 0.25 0 0.25 0.33 0.33 0 1.00 0.67 0.10 0 0.5 0.00 0.33 1 0 1
 0.67 0.29 1.00 0.0 0.06 1 0.75 1.00 0.67 0 0.33 0.00 0.32 0 0.5 0.33 0.33 0 0 1
 0.33 0.47 0.00 0.3 0.20 0 0.50 1.00 0.33 0 0.00 0.67 0.41 1 0.5 0.00 0.67 1 1 1
```

Dados Originais:

```
 2 30 4 0  4249 1 1 4 4 1 2 3 28 3 2 2 4 2 1 2
 4 18 4 0  2775 1 4 2 3 1 2 2 31 1 2 2 3 2 1 2
 4 36 3 9  9572 1 2 1 1 1 1 3 28 3 2 2 3 2 1 2
 2 12 2 3  1331 1 2 2 3 1 1 3 22 3 2 1 3 2 1 2
 1 48 2 3  6999 1 4 1 4 3 1 1 34 3 2 2 3 2 2 2
 3 15 2 3  2327 1 2 2 2 1 3 1 25 3 2 1 2 2 1 2
 4 18 2 5  1943 1 2 4 2 1 4 1 23 3 2 1 3 2 1 2
 2 24 4 2  4736 1 2 2 2 1 4 3 25 1 2 1 2 2 1 2
 2 24 2 1 11560 1 3 1 2 1 4 3 23 3 1 2 4 2 1 2
 2 24 4 2  4736 1 2 2 2 1 4 3 25 1 2 1 2 2 1 2
 3 24 4 0  1344 5 4 4 3 1 2 1 37 1 2 2 2 1 1 2
 2 36 0 3  3804 1 3 4 2 1 1 3 42 3 2 1 3 2 2 2
```

Quadro 2: Cliente 296 (classe Bom)

Dados normalizados:

```
 1.00 0.47 0.50 0.1 0.43 0 0.50 0.00 0.33 0.00 0.33 0.33 0.20 0 0.5 0.00 0.67 1 0 1
 0.00 0.18 1.00 0.0 0.10 0 0.50 1.00 0.67 0.00 0.33 0.33 0.20 1 0.5 0.33 0.67 1 0 1
 0.33 0.03 0.50 0.3 0.02 0 0.25 0.67 1.00 0.00 0.67 0.00 0.13 1 0.5 0.00 0.33 1 0 0
 0.33 0.65 0.25 0.0 0.66 1 0.00 1.00 0.67 0.50 1.00 1.00 0.30 1 1.0 0.00 1.00 1 1 1
 1.00 0.15 0.75 0.0 0.03 0 0.50 1.00 0.67 0.00 0.33 0.67 0.14 1 0.5 0.33 0.33 1 0 1
 0.67 0.03 1.00 0.0 0.06 0 0.50 0.00 0.67 0.00 0.00 0.00 0.98 1 0.5 0.67 0.00 0 0 0
 1.00 0.47 0.50 0.1 0.43 0 0.50 0.00 0.33 0.00 0.33 0.33 0.20 0 0.5 0.00 0.67 1 0 1
 1.00 0.47 0.50 0.1 0.43 0 0.50 0.00 0.33 0.00 0.33 0.33 0.20 0 0.5 0.00 0.67 1 0 1
 1.00 0.47 0.75 0.3 0.23 0 0.50 1.00 0.33 0.00 1.00 0.00 0.27 1 0.5 0.33 0.67 1 0 1
 0.00 0.29 0.50 0.3 0.12 0 1.00 1.00 0.67 0.00 1.00 0.00 0.80 0 0.0 0.00 0.33 1 0 1
 0.00 0.12 0.50 0.2 0.03 0 0.50 0.33 0.67 1.00 0.67 0.33 0.34 1 0.5 0.00 0.33 0 0 1
 0.33 0.21 0.75 0.9 0.12 1 1.00 1.00 0.67 0.00 0.33 0.33 0.41 1 0.5 0.33 0.67 1 0 1
```

Dados Originais:

```
 4 36 2 1  8133 1 3 1 2 1 2 2 30 1 2 1 3 2 1 2
 1 12 4 0  2121 1 3 4 3 1 2 2 30 3 2 2 3 2 1 2
 2  6 2 3   590 1 2 3 4 1 3 1 26 3 2 1 2 2 1 1
 2 48 1 0 12169 5 1 4 3 2 4 4 36 3 3 1 4 2 2 2
 4 14 3 0   802 1 3 4 3 1 2 3 27 3 2 2 2 2 1 2
 3  6 4 0  1299 1 3 1 3 1 1 1 74 3 2 3 1 1 1 1
 4 36 2 1  8133 1 3 1 2 1 2 2 30 1 2 1 3 2 1 2
 4 36 2 1  8133 1 3 1 2 1 2 2 30 1 2 1 3 2 1 2
 4 36 3 3  4454 1 3 4 2 1 4 1 34 3 2 2 3 2 1 2
 1 24 2 3  2384 1 5 4 3 1 4 1 64 1 1 1 2 2 1 2
 1 12 2 2   708 1 3 2 3 3 3 2 38 3 2 1 2 1 1 2
 2 18 3 9  2427 5 5 4 3 1 2 2 42 3 2 2 3 2 1 2
```