UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

FACULDADE DE CIÊNCIAS ECONÔMICAS

DEPARTAMENTO DE ECONOMIA E RELAÇÕES INTERNACIONAIS

MATHEUS GOMBOSKI

A UTILIZAÇÃO DE ALGORITMOS DE MACHINE LEARNING NA ANÁLISE ECONÔMICA

Porto Alegre

2019

MATHEUS GOMBOSKI

A UTILIZAÇÃO DE ALGORITMOS DE MACHINE LEARNING NA ANÁLISE ECONÔMICA

Trabalho de conclusão submetido ao Curso de Graduação em Ciências Econômicos da Faculdade de Ciências Econômicas da UFRGS, como requisito parcial para obtenção do título de Bacharel em Economia.

Orientador: Prof. Dr. Carlos Eduardo Schönerwald da Silva

Porto Alegre

2019

CIP - Catalogação na Publicação

Gomboski, Matheus
A Utilização de Algoritmos de Machine Learning na Análise Econômica / Matheus Gomboski. -- 2019.
65 f.
Orientador: Carlos Schönerwald da Silva.

Trabalho de conclusão de curso (Graduação) -- Universidade Federal do Rio Grande do Sul, Faculdade de Ciências Econômicas, Curso de Ciências Econômicas, Porto Alegre, BR-RS, 2019.

1. Machine Learning. 2. Economia Aplicada. 3. Falência Bancária. 4. C5.0. 5. Árvore de Decisão. I. da Silva, Carlos Schönerwald, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

MATHEUS GOMBOSKI

A UTILIZAÇÃO DE ALGORITMOS DE MACHINE LEARNING NA ANÁLISE ECONÔMICA

Trabalho de conclusão submetido ao Curso de Graduação em Ciências Econômicos da Faculdade de Ciências Econômicas da UFRGS, como requisito parcial para obtenção do título de Bacharel em Economia.

Aprovado em: Porto Alegre, _____ de _____ de 2019.

BANCA EXAMINADORA:

Prof. Dr. Carlos Eduardo Schönerwald da Silva – Orientador

UFRGS

Prof. Dr. Nelson Seixas dos Santos

UFRGS

Prof. Dr. Sabino da Silva Porto Júnior

UFRGS

AGRADECIMENTOS

Primeiramente, agradeço aos meus pais, pela amizade e pelo apoio incondicional dispensado em todas as situações. Todas as conversas, conselhos e exemplos foram essenciais nestes anos. Sem eles, nada teria sido possível.

À minha família, por todo o carinho e apoio e, em especial, à minha avó, com quem tive o privilégio de discutir e compartilhar diversas etapas deste trabalho. Seus conhecimentos foram cruciais para o bom desenvolvimento do meu estudo.

À minha esposa, por todo companheirismo, carinho e paciência nos momentos mais delicados desta caminhada. Ao seu lado, sonhos são arquitetados, atribuindo um novo significado aos estudos. Seus apontamentos e releituras foram capitais à realização desta pesquisa.

A todos os professores da Economia e da Engenharia, com quem tive o privilégio de conviver, mesmo que por poucas semanas. São eles os responsáveis pelo conhecimento que adquiri nesta graduação e que hão de nortear meus passos na vida profissional e em futuros aprimoramentos.

Ao meu orientador, Carlos Schönerwald da Silva, sempre muito atento e solícito em todos os momentos. Graças às nossas conversas, termino esta etapa de minha formação acadêmica com uma visão reformulada a respeito da economia e com ganas de buscar novos conhecimentos.

Aos amigos engenheiros e economistas que me acompanharam durante estes anos. As longas conversas deram um significado a mais aos estudos, além de tornarem os dias mais interessantes e divertidos.

Por fim, agradeço a todos os contribuintes brasileiros, responsáveis pelo financiamento desta universidade.

RESUMO

Este estudo investiga a possibilidade de utilização de algoritmos de *machine learning* em problemas econômicos aplicados. Mediante revisão da literatura vigente sobre o tema, verificaram-se os benefícios que esses algoritmos podem trazer ao debate econômico, além de se vislumbrarem as vantagens e as desvantagens que podem oferecer em relação aos métodos tradicionais. A pesquisa ancora-se na representação de árvores de decisão que, habitualmente, são enfatizadas por acadêmicos como ideais para abordar problemas de projeção. Além de serem facilmente interpretáveis, costumam apresentar alta performance de predição. Visto que estudos aplicado abordando casos brasileiros ainda são escassos, propôs-se a abordagem de um caso doméstico. Assim, com o emprego do algoritmo C5.0, constrói-se uma árvore de decisão destinada a projetar a falência de uma instituição financeira brasileira. Isso torna possível averiguar os principais indicadores que levam um banco à bancarrota, bem como seus valores. Todo o processo de construção, bem como a árvore de decisão propriamente dita, são apresentados, explicados e discutidos ao longo do trabalho.

Palavras-chave: Machine Learning. Economia Aplicada. Falência Bancária. C5.0. Árvore de Decisão.

ABSTRACT

This study investigates the possibility of using algorithms of machine learning in problems of applied economics. Through a review of current literature about the theme, it was possible to verify the benefits that these algorithms can bring to the economic debate, as well as showing the advantages and disadvantages they can offer when compared to traditional methods. The present research is based in the representation of decision trees which, usually, are emphasized by academics as the ideal model to approach projection problems. In addition, they are easily interpreted and usually show high prediction performance. Since applied studies involving Brazilian cases are still scarce, an approach of a domestic case was purposed. Therefore, with the implementation of C5.0 algorithm, a decision tree was developed to foresee the bankruptcy of a financial Brazilian institution. This makes possible to evaluate the main indicators that can make a bank bankrupt, as well as its numbers. All the development process, as well as the decision tree itself, are presented, explained and discussed throughout this study.

Keywords: Machine Learning. Applied Economics. Bank Failure. C5.0. Decision Tree.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

B3	Bolsa Brasil Balcão
BACEN	Banco Central do Brasil
CMN	Conselho Monetário Nacional
DRE	demonstração do resultado do exercício
EBITDA	*earnings before interest, taxes, depreciation and amortization*
ML	*machine learning*
MQO	mínimos quadrados ordinários
PIB	Produto Interno Bruto
PL	patrimônio líquido
Q1	primeiro quartil
Q3	terceiro quartil
TVM	títulos e valores mobiliários

SUMÁRIO

1 INTRODUÇÃO

Cada vez mais, economistas lidam com problemas complexos e que exigem ferramentas poderosas para serem resolvidos. Visto que os métodos tradicionais abordados na econometria, como o mínimos quadrados ordinários (MQO), apresentam desempenho satisfatório ao realizar inferências causais, mas não exibem a mesma assertividade ao lidar com problemas de previsão (VARIAN, 2014), é importante a tentativa de investigar a eficiência de outros métodos. A bibliografia aponta limitações do MQO ao realizar inferências preditivas. Dessa forma, estudos tratando de situações complexas e próximas à realidade são comprometidos. Percebe-se que as pesquisas econômicas vêm apresentando dificuldades em prever situações de maior estresse, como, por exemplo, a crise de 2008 (NYMAN; ORMEROD, 2017), ou períodos de expansão econômica (FILDES; STEKLER, 2002). Com isso, as ferramentas utilizadas atualmente, a fim de realizar projeções econômicas, devem ser revistas.

Uma das sugestões da literatura para lidar com problemas preditivos é a utilização de algoritmos de *machine learning*. Mesmo apresentando limitações que não garantam sua total eficácia em qualquer problema preditivo, em se tratando de casos complexos, essa abordagem aparenta ser a mais satisfatória (FERNÁNDEZ-DELGADO *et al*., 2014). Dessa forma, questionam-se quais algoritmos, exatamente, são os melhores para problemas de previsão e em quais casos podem ser aplicados a fim de gerarem bons resultados. Além disso, é importante investigar as possíveis falhas que podem decorrer da abordagem escolhida.

Assim, o presente trabalho propõe-se a responder às seguintes questões: “algoritmos de *machine learning* são eficientes para realizar projeções econômicas?” “Caso positivo, qual o algoritmo mais indicado e qual a sua precisão?”. À vista disso, a pesquisa será exploratória, já que procurará entender melhor a abordagem perante problemas econômicos, mediante uma análise bibliográfica e estudo de caso. Espera-se que as conclusões do trabalho confirmem o que indica a literatura vigente, apontando a eficiência do método investigado.

2 ALGORITMOS DE MACHINE LEARNING

A busca por métodos capazes de realizar previsões econômicas adequadas e assertivas vem desafiando cada vez mais economistas e profissionais de áreas afins. Atualmente, a econometria oferece ferramentas amplamente difundidas, com ênfase na estimação de parâmetros. Entretanto, ao lidar com problemas que requerem inferência preditiva, acadêmicos apontam para algoritmos de *machine learning* como os mais aconselháveis. Dessa forma, o presente trabalho investiga a utilização desses algoritmos a fim de melhorar a compreensão, sob uma nova abordagem, a respeito de previsões econômicas. Visto isso, é de alta relevância, inicialmente, entender os principais conceitos e a visão de acadêmicos sobre suas características, seus pontos positivos, negativos e em quais campos da economia sua contribuição pode ser mais significativa.

2.1 CONCEITOS E CARACTERÍSTICAS GERAIS

Lantz (2015) define o *machine learning* (ML) como "o campo de estudo interessado no desenvolvimento de algoritmos computacionais capazes de transformar dados em ação". Ainda, afirma que seu foco é ensinar máquinas[1] (computadores) sobre como usar os dados disponíveis para resolver determinados problemas. Como campo de estudo, é possível situá-lo na fronteira da ciência da computação, estatística e outras disciplinas envolvidas com inferências e decisões em cenários de incerteza, como a economia e a neurociência (JORDAN; MITCHELL, 2015). Segundo Lantz (2015), sua origem está relacionada com o surgimento de bancos de dados cada vez maiores e mais complexos, aliado ao desenvolvimento de métodos estatísticos e computacionais. Entretanto, esse autor afirma que, por mais que máquinas já tenham atingido feitos impressionantes, elas ainda possuem um alto grau de limitação para entender e interpretar problemas. As máquinas podem ter uma capacidade superior à dos seres humanos para resolver tais problemas, mas necessitam de direcionamento para atingirem resultados satisfatórios.

[1] Mitchell (1997) define que um programa aprende por meio de uma experiência *E*, a respeito de uma tarefa *T* e uma medida de performance *P*, quando sua performance na tarefa *T*, medida por *P*, melhora conforme aumenta sua experiência em *E*.

Outra descrição também é concebida por Géron (2019). Segundo ele, o "*machine learning* é a ciência de programar computadores para que eles possam aprender por meio de dados". Além disso, esse acadêmico também elenca os problemas ideais para serem resolvidos com tais sistemas. São eles: 1) problemas que necessitam de grande quantidade de ajustes, visto que um único algoritmo de *machine learning* pode simplificar o código e apresentar um desempenho mais satisfatório; 2) problemas complexos aos quais os métodos tradicionais não oferecem uma solução adequada; 3) problemas inseridos em ambientes dinâmicos, uma vez que os processos em questão podem, facilmente, ser adaptados a um novo conjunto de dados. Ainda, o autor afirma que, esses sistemas são, igualmente, propícios para se obterem informações sobre problemas complexos e grandes bancos de dados.

Na prática, o processo para implantação do *machine learning* pode ser dividido em cinco etapas: coleta de dados, preparação dos dados, treinamento do modelo, avaliação do modelo e melhoria do modelo (LANTZ, 2015). A primeira etapa – coleta de dados – envolve, como o nome já sugere, o recolhimento de dados. Em alguns casos, esses dados devem ser agrupados sob um único formato. A segunda etapa – preparação dos dados –, por sua vez, é crucial para a qualidade do resultado final. Assim, é importante conhecer as nuances dos dados, bem como eliminar aqueles desnecessários e confusos. O terceiro ponto – treinamento do modelo – envolve a seleção de um algoritmo apropriado para representar os dados já estabelecidos. Tal algoritmo representará os dados mediante um modelo que será usado posteriormente, para a realização de projeções. O quarto item – avaliação do modelo – envolve verificar a sua acurácia, visto que a solução encontrada por meio do algoritmo pode ser tendenciosa. Tal problema afetaria a qualidade do resultado futuro, pois o modelo seria eficiente apenas para a amostra já selecionada e apresentaria desempenho ruim para os dados fora dela. Por fim, o quinto item – a melhoria do modelo – é aplicada apenas em alguns casos, quando a verificação do quarto ponto indica tal necessidade. Nessas situações, é essencial, em geral, adicionar novos dados ou prepará-los de maneira distinta. Assim que todas as etapas estiverem concluídas com sucesso, o modelo estará pronto para ser implantado ao problema desejado. Para exemplificar o uso do *machine learning* atualmente, Jordan e Mitchell (2015) citam diversos casos em que o sistema é aplicado com grande sucesso. Entre eles, está a detecção de fraudes em pagamentos feitos por intermédio de cartões de crédito, a segmentação de pacientes médicos para direcioná-los aos tratamentos adequados, a redução de congestionamentos de trânsito mediante análise do histórico de tráfico, a alocação de forças policiais de acordo com o histórico de crimes de determinada cidade, o reconhecimento

facial por meio de imagens, entre outros. Apesar disso, os algoritmos não podem ser vistos de maneira alguma como substitutos humanos, visto que não possuem flexibilidade para extrapolar as restrições previamente definidas, nem são dotados de senso comum (LANTZ, 2015).

2.2 CATEGORIAS

Definidas as características gerais do *machine learning*, é conveniente discorrer-se um pouco mais sobre os tipos de algoritmos, sobretudo aqueles direcionados para os estudos de projeções, essenciais na análise de diversos problemas econômicos. Nos últimos anos, uma grande variedade de sistemas de *machine learning* foi desenvolvida com a finalidade de se abordarem os mais diversos problemas e tipos de dados (JORDAN; MITCHELL, 2015). Assim, é de grande importância classificar tais sistemas, para que haja um melhor entendimento sobre suas características e diferenças. De acordo com Géron (2019), essa categorização pode ser feita atendendo-se a três critérios de verificação: primeiro, se o sistema é preparado com a supervisão humana; segundo, se é capaz de aprender de maneira gradual; e, por último, se funciona mediante a comparação entre novos dados e dados já conhecidos, ou, se, em vez disso, detecta padrões no conjunto de dados e, a partir daí, constrói um modelo preditivo.

Assim, a primeira classificação leva em conta o tipo de supervisão que é realizado durante o ajuste do sistema. Há quatro categorias: aprendizado supervisionado, aprendizado não supervisionado, aprendizado semissupervisionado e aprendizado reforçado (GÉRON, 2019). Na primeira categoria – aprendizado supervisionado –, os dados usados para ajustar o modelo (os quais são chamados de "identificados") já incluem as soluções desejadas. Um exemplo prático de tal processo são tarefas de classificação ou predição, como classificadores de *spans* ou programas de reconhecimento facial. Nessa categoria, entre os algoritmos mais populares, situam-se os SVMs (*support vector machine*), redes neurais, regressões lineares, árvores de decisões e *random forests* (JORDAN; MITCHELL, 2015). Já na segunda categoria – aprendizado não supervisionado –, para Géron (2019), não há indicação sobre a solução desejada, não existindo, portanto, qualquer dado identificado. Dessa forma, o sistema deve aprender sem nenhum auxílio humano. A terceira categoria – aprendizado semissupervisionado – apresenta algoritmos que lidam com ambos os dados: identificados e

não identificados. Como exemplos, podem-se citar alguns serviços de hospedagem *on-line* de fotos, nos quais o usuário indica ao sistema o nome de determinada pessoa em apenas uma das fotografias. A última categoria – sistema de aprendizado reforçado – é completamente distinta das demais. Aí, o sistema observa o ambiente e seleciona e executa ações. Com base nisso, o sistema é recompensado ou penalizado. Como exemplo, é possível apontarem-se algoritmos que aprendem a caminhar ou a jogar algum jogo.

A segunda classificação proposta por Géron (2019) é baseada na capacidade de aprendizagem do sistema. Basicamente, os sistemas podem ser divididos em duas categorias: aprendizagem em lotes e aprendizagem *on-line*. Os sistemas de aprendizagem em lotes são incapazes de aprender de maneira incremental. Hoi *et al*. (2018) indicam que, nesse caso, é necessário que o ajuste seja feito utilizando todos os dados disponíveis. Assim que o ajuste é feito, o sistema passa a funcionar sem que haja um novo processo de aprendizagem. Segundo os autores, esse método implica uma baixa eficiência (tanto de tempo, como de espaço) e envolve problemas quando aplicado a grandes bancos de dados, já que o modelo tem que ser ajustado toda vez que os dados mudarem. Já os sistemas de aprendizagem *on-line*, segundo Géron (2019), apresentam um funcionamento distinto. Esses são capazes de serem ajustados de maneira incremental. Os dados são adicionados a eles de forma sequencial, sejam individualmente ou em pequenos grupos. Por isso, Hoi *et al*. (2018) afirmam que a aprendizagem *on-line* é mais eficiente para as aplicações contemporâneas, nas quais a quantidade de dados é grande, e seu fluxo, muitas vezes, é contínuo.

A terceira e última classificação oferecida por Géron (2019) refere-se à maneira como os sistemas se generalizam. Após o ajuste, é necessário que o sistema tenha um bom resultado frente a problemas com os quais ele nunca tenha se deparado antes. Essa ação frente a novos problemas chama-se generalização. Nesta classificação, também há duas categorias: aprendizado baseado em exemplos (*instance-based*) e aprendizado baseado em modelos (*model-based*). No primeiro caso, os sistemas agem baseados em comparação com os exemplos já conhecidos. Já na segunda classificação, deve-se construir um modelo para os exemplos apresentados. Assim que construído, tal modelo é usado para se realizarem predições.

2.3 RELAÇÕES COM A ECONOMETRIA

Em seu estudo, Pinto (2011) aborda o conceito de econometria, além de discutir sua aplicação em problemas. Esse autor que a econometria é o resultado da incorporação do conhecimento econômico, matemático e estatístico. É usada amplamente para testar teorias econômicas, auxiliar *policy makers* em tomadas de decisões e entender relações entre variáveis. A sua proposta é estimar parâmetros consistentes entre variáveis dependentes e independentes. Dessa forma, busca estabelecer padrões econômicos.

Varian (2014) debate as diferenças entre os métodos econométricos e os algoritmos de *machine learning* quando aplicados em problemas econômicos. Enquanto um dos principais objetivos da econometria é a inferência causal, os sistemas de *machine learning* enfatizam a inferência preditiva. Assim, enquanto o ML utiliza análise preditiva e mineração de dados para prever uma variável em função de todas as outras, a econometria emprega métodos estatísticos para a modelagem causal de relações econômicas. Dessa forma, portanto, as disciplinas podem ser vistas como complementares, e não como substitutas.

Mullaimathan e Spiess (2017) aprofundam a discussão, enfatizando a relação entre o *machine learning* e a econometria. Inicialmente, aludem que o ML pode ser visto como complementar à econometria, uma vez que os objetivos de suas utilizações diferem. Enquanto a última se preocupa, sobretudo, com a estimação dos parâmetros, o ML visa à projeção de determinada variável. Esses sistemas buscam por estruturas gerais e são avaliados conforme sua capacidade de encontrar tais estruturas.

Breiman (2001) especifica que há dois objetivos principais durante a análise de determinado grupo de dados: predição e informação. Quando o principal objetivo é o primeiro (predição), o pesquisador busca respostas para situações futuras, em que novas variáveis serão adicionadas ao problema. Esse é o objetivo dos sistemas de *machine learning*, No segundo caso (informação), procuram-se extrair informações sobre a natureza das variáveis. Esse, entretanto, é o foco das técnicas empregadas na econometria (como o MQO). Assim, um método (*machine learning*) busca a acurácia de seus resultados, enquanto o outro (econometria) preza pela interpretabilidade do problema observado.

Segundo Mullaimathan e Spiess (2017), além de incrementar os problemas de previsão, os algoritmos de *machine learning* também são capazes de lidarem com novos tipos

de dados, como fotos, vozes e textos. Além disso, sua capacidade de detectar padrões acontece mesmo perante estruturas mais complexas. Assim, o método é capaz de modelar problemas complexos em formas funcionais, sem sobreajustes, e que apresentem boa performance quando aplicadas fora da amostra. Os autores definem, também, os algoritmos em termos matemáticos: eles consideram uma função de perda (L ($\hat{y}, y$)) como entrada e procuram por uma nova função ($\hat{f}$) que tenha a menor perda esperada (E(x,y) [L($\hat{f}$(x),y)] em um novo ponto de dados da mesma distribuição. Assim, Vasconcelos (2017) explicita que o objetivo desse processo é maximizar o ajustamento do modelo para cada subgrupo da base de dados.

Em seu trabalho, Mullaimathan e Spiess (2017) discutem e comparam a performance de algoritmos de ML com o método dos mínimos quadrados ordinários (MQO), tradicionalmente, empregado na econometria. Esses autores abordam um problema de projeção de preços de imóveis norte-americanos. Para isso, selecionam uma amostra de 10.000 imóveis, com 150 características para cada um (como, por exemplo, número de banheiros, quartos, localização, tamanho, etc.). Ao utilizarem diferentes métodos de previsão (MQO, árvore de decisão, LASSO, *random forest* e *ensemble)*, contemplando tanto a econometria, como o ML, os autores avaliam os resultados para amostras *out-of-sample* e *in-sample*. Em ambas, os resultados de previsão mais precisos foram os encontrados mediante o *random forest*, um algoritmo de *machine learning*, baseado em árvores de decisão. Comparativamente, a diferença maior em relação aos outros métodos ocorreu nas amostras *out-of-sample*. Nessas, o R^2 (coeficiente de determinação) encontrado foi 85,1%, enquanto o MQO apresentou R^2 igual a 47,3%. Assim, no exemplo observado, o modelo de árvore de decisão foi capaz de explicar, praticamente, 85% do resultado encontrado.

Mullaimathan e Spiess (2017) também alertam para algumas precauções que devem ser tomadas ao se utilizar esse tipo de algoritmo: os estimadores gerados, muitas vezes, não são consistentes (ou seja, são viesados). Os autores alertam que o ML não foi desenvolvido com o mesmo propósito da econometria. Assim, há um *trade-off* entre redução da incerteza dos estimadores e viés nos estimadores (VASCONCELOS, 2017). Desse modo, portanto, Mullaimathan e Spiess (2017) sugerem que a pesquisa realizada mediante ML deve buscar funções relevantes à variável dependente ($\hat{y}$), e não, aos estimadores.

Vasconcelos (2017) afirma que o uso de *machine learning* vem se tornando cada vez mais popular entre economistas. Ele entende que três características podem explicar essa

popularização. Primeiro, ela possui um alto grau de eficiência quando a previsão desejada é *out-of-sample*. Entretanto, esse privilegiado grau de eficiência está relacionado à presença de estimadores viesados. Segundo, o método permite um alto grau de automatização de problemas complexos. E, por último, o ML trata diretamente dos problemas de *big data*. Esse último ponto conecta-se com o segundo. O *big data* pode ser definido como "grandes bases de dados com um número elevado de preditores relativos às observações disponíveis" (VASCONCELOS, 2017).

2.4 ÁRVORES DE DECISÃO

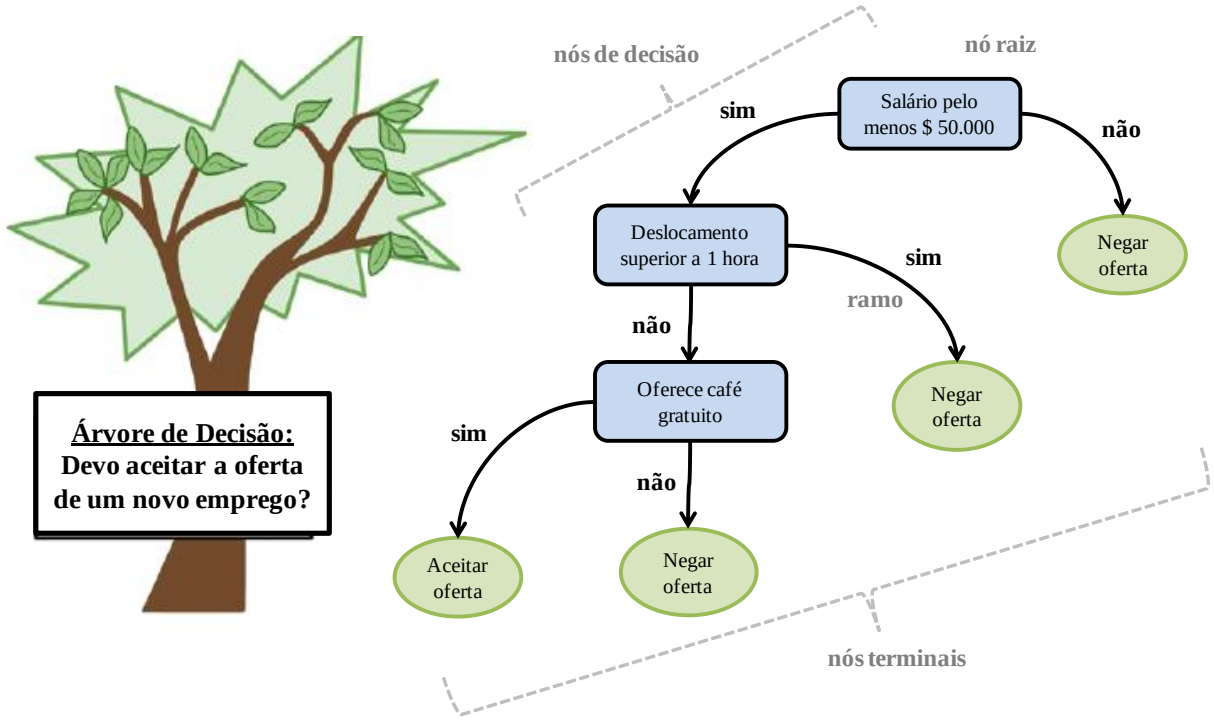
Hoje, a gama de algoritmos de *machine learning* é imensa. Dessa forma, para uma aplicação adequada dos modelos de previsões, é importante também compreender quais os melhores algoritmos para a resolução de determinados problemas. Fernández-Delgado *et al*. (2014) investigam os 179 principais tipos de algoritmos, oriundos de 17 famílias. Para isso, eles utilizam 121 conjuntos de dados, a fim de atingirem conclusões precisas e sem a influência de correlações entre os dados. Os autores afirmam que os pesquisadores, em sua maior parte, geralmente, aplicam aqueles modelos com os quais estão acostumados a lidar. Mas isso, não necessariamente, leva aos melhores resultados, visto que uma gama de métodos desconhecidos por eles (ou simplesmente ignorados) deixam de ser testados.

Fernández-Delgado *et al*. (2014) realizaram um amplo estudo sobre a performance de um grande número de indicadores. Segundo o teorema desenvolvido por Wolpert (1996), o melhor tipo de algoritmo não será o mesmo para todos os conjuntos de dados. Ciente disso, o estudo de Fernández-Delgado *et al*. (2014) propôs um ranking, onde classificou a performance de todos os indicadores em termos probabilísticos. O melhor resultado foi atingido com a família de *random forest*, mais precisamente pelo *parallel random forest*. Esse classificador atingiu uma acurácia máxima de 94,1% (ou seja, a partir do modelo construído com base nos dados *in-sample*, classificou corretamente 94,1% dos dados *out-of-sample*). Ainda, o algoritmo teve acurácia superior a 90% em 102 dos 124 conjuntos de dados. Logo em seguida, o segundo modelo com melhor resultado foi o SVM com núcleos gaussianos e polinomiais. Assim, segundo os próprios autores, tal pesquisa significa grande avanço no estudo comparativo de classificadores, pois não se restringe à análise de apenas uma família de classificadores, nem a um banco de dados limitado.

Entre os sistemas de *machine learning*, Quinlan (1986) enfatiza árvores de decisões como um método menos complexo que os demais, mas, ao mesmo tempo, capaz de resolver problemas de alta complexidade. Lantz (2015), por sua vez, afirma que árvores de decisão são eficazes para resolver problemas de predição e simples de serem compreendidas, já que podem ser demonstradas visualmente, sem a necessidade de um conhecimento estatístico prévio. Ainda, permite o emprego de diferentes tipos de dados. Esses algoritmos utilizam a estrutura de uma árvore para modelar as relações entre as variáveis e os possíveis resultados (ou decisões). O problema inicia no nó-raiz e, a partir daí, ramos levam a diferentes nós de decisão, até serem atingidos os resultados finais (chamados, originalmente, de *leaf nodes*). Tal abordagem é conhecida como "dividir e conquistar" (*divide and conquer*), pois os dados vão sendo divididos em subconjuntos cada vez menores. Cada subconjunto é associado a uma classe. Esse processo termina quando uma destas três situações é satisfeita: 1) todos os resultados dos nós pertencem à mesma classe; 2) não há mais características que possam ser diferenciadas entre os resultados atingidos pelos nós; ou 3) a árvore atingiu o limite pré-estabelecido. Dessa forma, o problema pode ser facilmente interpretado. Por isso, talvez, seja essa a técnica de *machine learning* mais utilizada.

A fim de exemplificar o funcionamento de uma árvore de decisão, Lantz (2015) apresenta, graficamente, um modelo genérico. Nesse seu exemplo, a árvore projeta se uma oferta de emprego deve ser aceita ou negada. A Figura 1, abaixo, apresenta tal árvore, distinguindo os tipos de nós.

Figura 1 – Exemplo de uma árvore de decisão



Fonte: elaborada pelo autor (2019).

No exemplo da Figura 1, a leitura da árvore é iniciada no nó raiz, localizado na parte superior. Após isso, escolhe-se qual dos ramos é o apropriado para, assim, atingir um novo nó de decisão ou, então, um dos nós terminais. Nesse exemplo, há apenas duas opções para os ramos: sim ou não. Entretanto, os ramos podem apresentar mais opções. O processo é finalizado quando um dos nós terminais (indicados em verde, na Figura 1) é atingido e, com isso, obtém-se a resposta para a questão proposta.

Segundo Da Silva (2005), as partições de determinada árvore de decisão são definidas de acordo com o ganho de informação que será gerado. Nos modelos destinados a processos classificatórios (como árvores de decisão), uma das ferramentas mais tradicionais para medir o ganho de informação é a entropia. Lantz (2015) menciona que, em linhas gerais, a entropia mede a aleatoriedade (ou desordem) de um conjunto de dados. Um conjunto com entropia alta possui uma grande aleatoriedade e, assim, não apresenta um padrão comportamental evidente. Em termos matemáticos, a entropia de certo conjunto S pode ser expressa pela seguinte equação:

$$\text{Entropia } (S) = \sum_{i=1}^{C} - p_i \; log_2 \; (p_i).$$

Nessa equação, *c* é o número de classes e p_i representa a proporção de valores pertencentes à classe *i*. Assim, a fim de buscar padrões para explicar o problema, o sistema sempre procurará por atributos que diminuam sua entropia, ou seja, que proporcionem um maior ganho de informação. O ganho de informação gerado ao incluir um novo subconjunto ao modelo pode ser calculado da seguinte forma:

$$\text{Ganho de Informação } (F) = \text{Entropia } (S_1) - \text{Entropia } (S_2).$$

Nesse caso, *F* é um atributo, e o lado direito da equação representa a diferença entre a entropia antes e depois do particionamento do conjunto de dados. Assim, quanto maior o ganho de informação, mais homogêneos são os subgrupos criados pela árvore de decisão.

Breiman (1998) alerta, também, para o problema do *overfitting (*sobreajuste). Segundo Da Silva (2005), o *overfitting* "é o ajuste demasiado dos dados de treinamento". Uma árvore de decisão pode crescer indefinidamente, criando subgrupos cada vez menores, a fim de aumentar a entropia total (LANTZ, 2015). Isso levaria ao problema do sobreajuste, que é perfeitamente ajustado ao conjunto de dados observado; entretanto, é ineficaz quando aplicado sobre a amostra-teste. Breiman (1998) recomenda a realização da poda da árvore para evitar a ocorrência desse problema e, assim, melhorar a capacidade de projeção *out-of-sample* do modelo. Basicamente, a poda limita o crescimento da árvore até um determinado nó e pode ser empregada de duas maneiras: restringindo o crescimento da árvore antes de sua construção, ou eliminando os nós que possuem pouco impacto na classificação do problema, após a sua construção completa (DA SILVA, 2005).

Os problemas no mundo real podem se tornar extremamente complexos, envolvendo um grande número de variáveis e, consequentemente, uma árvore de decisão com muitos nós. Para resolver esses problemas, o algoritmo de *machine learning* mais utilizado é o C5.0, desenvolvido pelo cientista da computação J. Ross Quinlan (LANTZ, 2015). Lantz (2015) afirma que esse algoritmo apresenta desempenho muito similar a outros muito mais complexos (como redes neurais ou *support vector machines* (SVM)). Esse autor elenca pontos positivos e negativos desse tipo de abordagem. Entre os fatores favoráveis, cita que o algoritmo possui um ótimo desempenho na maioria dos problemas, é um processo automático capaz de lidar com diversos tipos de variáveis, exclui variáveis desnecessárias, pode ser usado tanto em bases de dados pequenas, como em grandes, seus resultados podem ser facilmente interpretados, além de ser mais eficiente que muitos outros modelos complexos. Por outro

lado, menciona como ponto negativo que árvores de decisão com muitas ramificações, geralmente, são tendenciosas: elas são fáceis de sobreajustar o modelo, mas pequenas mudanças nos dados utilizados para realizar o ajuste podem trazer grandes mudanças no resultado final.

Árvores de decisão podem ser utilizadas para resolver as mais diversas situações. Lantz (2015) cita três exemplos em que o uso é bem sucedido: modelos de notas de crédito, estudo do perfil de consumidores e diagnóstico de condições médicas. No exemplo inicial, usa-se para definir se determinada solicitação de crédito será atendida ou não. No segundo caso, o classificador determinará o grau de satisfação de cada cliente em relação ao consumo de um produto ou serviço. Por fim, haveria a previsão do diagnóstico médico apenas com base em exames, sintomas e/ou a velocidade do avanço de determinada doença. Géron (2019), por sua vez, refere a aplicação de forma mais genérica. Afirma que árvores de decisão apresentam resultados satisfatórios tanto para tarefas de classificação, como para tarefas de predição. Em seu estudo, Nyman e Ormerod (2017) utilizam o algoritmo para projetar o PIB dos Estados Unidos e do Reino Unido. Já a tarefa de classificação pode ser encontrada no ensaio de Öcal *et al*. (2015), onde aplicam árvores de decisão a fim de classificarem o desempenho financeiro de determinadas empresas.

Quinlan (1986) discute sobre a família de sistemas de aprendizados conhecida como árvores de decisão. Esse autor afirma que o *machine learning* é objeto de pesquisa desde os anos 50, quando a inteligência artificial passou a ser reconhecida como disciplina. As árvores de decisão, por sua vez, são classificadas dentro de uma única família, devido ao mesmo sistema de representação adotado. Quinlan (1986) assegura que a abordagem empregada em árvores de decisão é relativamente simples, deixando de lado redes semânticas e representações de primeira ordem. Dessa forma, as metodologias de aprendizado utilizadas em árvores de decisão são consideradas menos complexas do que as de sistemas que se expressam por meio de linguagens mais poderosas. Mesmo assim, ainda são capazes de atingir excelentes resultados, resolvendo problemas práticos e significativos.

Segundo Carbonell, Michalski e Mitchell (1983), os sistemas de aprendizado de máquina podem ser classificados de acordo com alguma destas três características: a estratégia de aprendizado adotada, a representação do conhecimento adquirido pelo sistema e o domínio de aplicação do sistema. Árvores de decisão possuem um elo em comum em todos esses aspectos. Quinlan (1986) diz, resumidamente, que esses sistemas desenvolvem árvores

de decisão a fim de classificarem objetos. Eis alguns exemplos de tais classificações: indicação da possibilidade de chuva mediante observações atmosféricas, diagnóstico médico pela análise dos sintomas, etc. Esses sistemas procuram por padrões na amostra previamente dada e a reexaminam durante várias etapas do aprendizado. A amostra na qual as regras para classificação (árvore de decisão) são desenvolvidas pode ser obtida por meio de duas formas. A primeira pode ser conhecida pelo acesso a um banco de dados já existente, composto por um histórico de observações (como o banco de dados de um hospital, com informações completas sobre os pacientes). Em geral, esse tipo de dados fornece informações confiáveis. Entretanto, como eles não foram organizados, podem ser redundantes ou omissos em relação a algumas observações importantes. A outra forma de obtenção dos dados é mediante a coleta por um especialista, em que haja o preparo dos dados para uma correta aplicação à regra de classificação. Mas é interessante compreender que pode haver diversas árvores corretas para o mesmo problema. Ainda, a essência da indução é construir árvores que não classifiquem corretamente apenas os objetos da amostra, mas também os fora dela. Para isso, a árvore deve capturar uma relação significativa entre as variáveis do problema.

2.5 ESTUDOS APLICADOS

Visto isso, convém analisarem-se as principais pesquisas aplicadas sobre a utilização de algoritmos de ML em projeções econômicas, bem como seus métodos e resultados. Nyman e Ormerod (2017) discutem as falhas da comunidade econômica em não prever a recessão de 2008/09. Os autores assinalam que a discussão sobre projeções tem espaço no debate acadêmico há, pelo menos, 50 anos. Em sua análise, realizam previsões com dados anteriores a esse período, utilizando o *random forest* e o mínimos quadrados ordinários (MQO). Suas conclusões apontam para uma limitação em relação ao método empregado tradicionalmente pela econometria (MQO) quando estes são empregados em tarefas de projeção. Há 20 anos, Fildes e Stekler (2000) concluíram que nenhum método apresentou incremento de precisão significativo no período analisado. Em seu estudo, os autores verificam que a pesquisa econômica mostra um crescimento significativo desde 1960. Novas teorias e metodologias foram implantadas, principalmente, na macroeconomia. Apesar disso, esses autores denunciam que a comunidade científica concedeu pouca atenção para o desenvolvimento de novas técnicas de previsão, comprometendo, assim, a qualidade dos estudos práticos. Por fim,

os autores também constatam que as previsões econômicas não apenas falharam em prever recessões, como também subestimaram períodos de expansão econômica.

Em sua aplicação, Nyman e Ormerod (2017) realizam previsões de curto-prazo para o crescimento do PIB, utilizando dados trimestrais do Reino Unido e dos Estados Unidos. Para isso, empregam dois métodos a fim de compará-los: *random forest* e mínimos quadrados ordinários. O trabalho utilizou previsões para seis períodos à frente. Em linhas gerais, o *random forest* apresentou previsões altamente satisfatórias (ou seja, em conformidade com o que, de fato, houve) com um R^2 de 0,290. Em contrapartida, o R^2 apresentado pelo MQO foi 0,042. Dessa forma, a acurácia de previsão do *random forest* foi superior. Visto em termos gráficos, percebe-se que as projeções não captaram completamente a variação do PIB, mas foram suficientes para prever sua queda e uma futura recessão. Ainda, as previsões não indicaram nenhum momento de crise quando, de fato, isso não aconteceu. Os resultados foram muito próximos, tanto para os dados dos Estados Unidos, como para aqueles do Reino Unido.

Outra aplicação importante foi proposta por Öcal *et al*. (2015). Buscando prever o sucesso ou insucesso financeiro futuro de determinada companhia, os autores desenvolveram uma árvore de decisão a partir do algoritmo C5.0. Para isso, foram utilizados, inicialmente, 35 dados financeiros (tanto quantitativos, como qualitativos) de 206 empresas manufatureiras listadas na Bolsa de Valores de Istambul. Tais elementos compreendem o período de 2007 a 2013. Segundo os autores, diversos tipos de algoritmos são citados na literatura vigente como possíveis para resolver casos aplicados. Entretanto, árvores de decisão, além de apresentarem excelente performance, também são fáceis de serem interpretadas (QUINLAN, 1986). Visto isso, mostram-se preferíveis para o problema proposto. No caso sugerido, Öcal *et al*. (2015) determinam que uma empresa não obteve sucesso quando houve a ocorrência de alguma das seguintes situações: retirada do mercado devido a dificuldades financeiras, suspensão da negociação das ações devido a dificuldades financeiras, troca de mercado devido a dificuldades financeiras, obrigação da realização de declarações mensais devido a dificuldades financeiras ou aviso para tomar medidas cautelares devido à perda de capital ou pedido judicial. A árvore de decisão obtida leva em conta as seguintes variáveis: lucro sobre vendas líquidas antes dos impostos, taxa de alavancagem, rotatividade do capital de giro, estrutura patrimonial, liquidez corrente, fluxo de caixa sobre passivo total e, finalmente, EBITDA sobre ativos totais. Dessa forma, o modelo sugere que o conjunto dessas variáveis é

capaz de projetar o sucesso ou insucesso financeiro de determinada companhia. Por fim, o modelo encontrado por meio do C5.0 apresentou acurácia de 85,13% para a amostra total do estudo. Tal nível de precisão foi considerado satisfatório pelos autores.

Dessa forma, é importante a proposição da aplicação de um sistema de *machine learning* para o caso brasileiro. Tais sistemas mostram-se, para muitos casos, satisfatoriamente assertivos para tarefas preditivas, além de possibilitarem uma nova relação com os dados. Ademais, a literatura econômica apresenta pouca evolução na discussão sobre previsões (FILDES; STEKLER, 2002). Como consequência, discussões interdisciplinares são comprometidas, prejudicando o avanço, também, de outras áreas da economia.

3 METODOLOGIA

Este trabalho verifica a eficiência de um algoritmo de árvore de decisão para a previsão do sucesso ou insucesso financeiro de instituições financeiras com sede no Brasil. Visto que a literatura já apresenta aplicação bem sucedida para um problema semelhante, optou-se por aplicá-lo ao cenário local, onde pesquisas neste campo ainda são escassas. Em virtude do argumento de acadêmicos sobre os benefícios e limitações dos diversos algoritmos, decidiu-se pela utilização do C5.0[2], dadas sua simplicidade para apresentação, sua precisão e sua adequação para problemas de classificação binária, como o proposto neste trabalho. Além disso, chama atenção o fato de a discussão econômica, tradicionalmente, produzir projeções muito pouco assertivas (MULLAINATHAN; SPIESS, 2017), o que redundou no questionamento sobre os métodos tradicionais.

A investigação partiu de casos particulares: indicadores financeiros (quantitativos) de 193 instituições financeiras brasileiras, classificadas pelo Banco Central do Brasil como bancos múltiplos ou bancos comerciais. Tais empresas foram dividas entre aquelas que obtiveram sucesso e aquelas que não o conseguiram. Como *proxy* de insucesso, usou-se a inserção da companhia ao regime de liquidação extrajudicial. Assim, aquelas que não ingressaram ao regime de liquidação extrajudicial foram classificadas como instituições com sucesso. Após a seleção dessas empresas, analisaram-se todas as informações disponibilizadas pelo Bacen a respeito delas. A autarquia disponibiliza dados relativos aos seus balanços patrimoniais e demonstrações de resultados do exercício, abrangendo o período que se estende desde 1994 até 2019. Por meio desses dados, pôde-se selecionar 25 indicadores financeiros para compor as variáveis do algoritmo. Por fim, efetuou-se um ajuste dos dados extremos, a fim de eliminar os efeitos negativos que produzem. Esse ajuste foi realizado mediante o processo de Winsorização e atingiu, aproximadamente, 3% da amostra.

Assim, após o ajuste das variáveis, pôde-se iniciar a construção da árvore de decisão, por intermédio do uso do algoritmo C5.0. Para isso, 4.850 variáveis estiveram disponíveis. O programa contendo o algoritmo foi escrito e processado dentro do ambiente computacional R, versão 3.6.1, para Windows. O algoritmo exige que um percentual da amostra seja utilizado como teste para o modelo ajustado. Assim, optou-se por selecionar 20% das variáveis

[2] Algoritmo de *machine learning* baseado em árvore de decisão. Desenvolvido pelo cientista da computação J. Ross Quinlan.

disponíveis para essa função. Dessa forma, 3.875 variáveis foram utilizadas no ajuste do modelo, enquanto 975 compuseram a amostra de teste. Com o programa concluído, gerou-se a árvore de decisão.

Visto que a árvore apresentada é válida apenas para a ordenação e o tamanho da amostra proposta, reaplicou-se o algoritmo diversas vezes, com novas configurações e tamanhos de amostras-teste. Mediante as informações obtidas dessas novas aplicações, mostraram-se faixas de precisão do modelo, além de uma nova árvore de decisão. Para a construção desta faixa de precisão, reaplicou-se o algoritmo 75 vezes, em cinco tamanhos distintos de amostras-teste. Os resultados foram expostos em forma de tabela, contendo os valores máximo, mínimo, médio e a variância para cada tamanho de amostra-teste. Já a nova árvore de decisão foi gerada utilizando-se 50% dos dados para compor a amostra-teste.

A partir dos modelos encontrados, procurou-se produzir afirmações gerais sobre as empresas do setor estudado. Tais afirmações foram a respeito dos principais fatores que levam uma instituição financeira à falência, bem como dos pontos de atenção que devem ser levados em consideração por tais empresas. Dessa forma, como casos particulares foram o ponto de partida para a realização de generalizações, o método empregado foi o indutivo.

4 COLETA DE DADOS

Para a realização da projeção do sucesso ou insucesso financeiro de determinadas empresas, é importante, inicialmente, que todos os dados necessários sejam coletados, analisados e, devidamente, justificados. Isso permitirá que o algoritmo apresente um desempenho satisfatório. A pesquisa desenvolvida por Öcal *et al*. (2015) definiu que uma empresa não obteve sucesso quando houve sua retirada do mercado, suspensão da negociação das ações, troca de mercado ou obrigação de divulgação mensal de determinadas informações. Um critério semelhante pode ser utilizado para casos nacionais.

4.1 SEGMENTAÇÃO DO ESTUDO

Dado que há uma grande quantidade de setores econômicos com diferentes características e peculiaridades, entende-se que é necessário segmentar a pesquisa em um determinado ramo, para evitar qualquer tipo de viés nos dados, além de propiciar uma discussão mais adequada sobre os resultados encontrados. Além disso, é imprescindível que haja acesso aos dados financeiros do grupo de empresas selecionadas. Assim, será possível aplicar o algoritmo de árvore de decisão, bem como diferenciá-las entre companhias que tiveram êxito no período analisado e outras que não o obtiveram.

Segundo dados da B3, o segmento brasileiro que possui o maior número de empresas de capital aberto é bancos. Tal segmento é compreendido dentro do subsetor “intermediários financeiros”, que, por sua vez, apresenta-se no setor “financeiro”. Ele comporta, atualmente, 26 companhias listadas. Além disso, o setor financeiro possui uma grande importância na atual estrutura econômica, já que participa, ativamente, dos demais setores (ALBERTIN, 1999). Isso fica evidenciado ao se verificar que, atualmente, esse é o setor nacional com o maior volume de compras interempresariais (EUROMONITOR, 2019). Visto isso, esse grupo mostra-se de grande relevância para compor a amostra deste estudo.

Entretanto, por mais que não haja uma indicação exata mostrando o tamanho mínimo adequado do banco de dados utilizado para a aplicação do algoritmo, a quantidade de empresas de capital aberto classificadas como “bancos” parece ser insuficiente, podendo comprometer a qualidade dos resultados encontrados. Todas as demais pesquisas citadas

lidam com quantidades maiores de elementos. Öcal *et al*. (2015), por exemplo, lidam com uma amostra de 206 companhias. Dessa forma, é importante que dados de um grupo maior de empresas estejam disponíveis. Assim, é fundamental analisar a disponibilidade, também, dos dados de empresas não listadas na bolsa.

O Banco Central do Brasil divulga dados financeiros de todas as instituições financeiras do país, tanto das em funcionamento, como daquelas que tiveram suas atividades encerradas por motivos falimentares. Dessa forma, é possível adotar um critério semelhante ao de Öcal *et al*. (2015) para diferenciá-las em dois grupos: as empresas que obtiveram sucesso e as que não o alcançaram. Com isso, serão consideradas companhias de sucesso aquelas que estão atualmente em funcionamento, enquanto o grupo contrário será composto pelas instituições que estiveram em regime de liquidação extrajudicial dentro do período estipulado.

Atualmente, as instituições financeiras são segmentadas em diversas categorias. O Banco Central do Brasil as fragmenta da seguinte forma: bancos múltiplos, bancos múltiplos cooperativos, bancos comerciais, bancos de câmbio, bancos comerciais estrangeiros, caixas econômicas federais, cooperativas de crédito, administradoras de consórcios, bancos de investimento, bancos de desenvolvimento, sociedades corretoras de TVM e câmbio, sociedades distribuidoras de TVM, sociedades de crédito direto, sociedades de crédito, financiamento e investimento, sociedades de crédito imobiliário APE, sociedades de crédito ao microempreendedor, sociedades de arrendamento mercantil, sociedade de empréstimo entre pessoas, sociedades de investimento, agências de fomento, companhias hipotecárias e, por fim, instituições de pagamento. Frente à grande variedade de instituições financeiras, é importante fragmentar os dados com o intuito de inserir no modelo apenas instituições que possuam estruturas e funcionamentos semelhantes. Além disso, é essencial que o grupo selecionado conte com um número adequado de empresas, tanto em funcionamento, como em regime falimentar. Com isso, procura-se evitar que haja algum tipo de viés no modelo estudado.

Visto isso, optou-se por selecionar bancos múltiplos e bancos comerciais para compor a amostra. Tal conjunto atende às condições propostas acima. De acordo com a Resolução CMN 2.099, de 1994,

> os bancos múltiplos são instituições financeiras privadas ou públicas que realizam as operações ativas, passivas e acessórias das diversas instituições financeiras, por

> intermédio das seguintes carteiras: comercial, de investimento e/ou de desenvolvimento, de crédito imobiliário, de arrendamento mercantil e de crédito, financiamento e investimento.

Além disso, ‘o banco múltiplo deve ser constituído com, no mínimo, duas carteiras, sendo uma delas, obrigatoriamente, comercial ou de investimento, e ser organizado sob a forma de sociedade anônima’. Já os bancos comerciais, segundo a mesma resolução do CMN (Conselho Monetário Nacional),

> [...] são instituições financeiras privadas ou públicas que têm como objetivo principal proporcionar suprimento de recursos necessários para financiar, a curto e a médio prazos, o comércio, a indústria, as empresas prestadoras de serviços, as pessoas físicas e terceiros em geral. A captação de depósitos à vista, livremente movimentáveis, é atividade típica do banco comercial, o qual pode também captar depósitos a prazo.

Segundo relação do Banco Central do Brasil, de agosto de 2019, há, atualmente, 130 bancos múltiplos em funcionamento no país e 14 bancos comerciais, totalizando 144 corporações. Dessas, a grande maioria (96 empresas) tem sua sede no estado de São Paulo. Além disso, há, ainda, 41 bancos múltiplos e oito bancos comerciais que estiveram em regime de liquidação extrajudicial desde 1994 e têm seus indicadores financeiros divulgados pela mesma fonte, somando 49 companhias. Dessa forma, os 144 bancos em funcionamento serão considerados como empresas de sucesso, enquanto aqueles que foram postos em liquidação extrajudicial serão categorizados como empresas de insucesso. A relação completa das 193 instituições, bem como informações complementares, podem ser visualizadas no Apêndice A.

4.2 INFORMAÇÕES DISPONÍVEIS

Posto isso, é importante, a partir deste momento, analisar os indicadores financeiros disponíveis, suas características e decidir quais serão utilizados na aplicação. O Banco Central do Brasil, por intermédio de seu banco de dados, divulga as diversas notificações de todas as instituições financeiras do país. Tais informações são apuradas pelos próprios bancos, enviadas ao Bacen, que as compila e divulga em seu portal. Essas divulgações englobam o período compreendido desde dezembro de 1994 até o último trimestre deste ano. Desde seu início até o final de 1999, as exposições ocorreram semestralmente. A partir de 2000,

entretanto, passaram a ser trimestrais. Os dados divulgados sobre as instituições financeiras são referentes, basicamente, aos seus balanços patrimoniais. Até 1999, são disponibilizados os seguintes relatórios com informações sobre cada companhia: ativo, passivo, resultado da intermediação financeira, resultado líquido, depósito e, por fim, resumo. Já para os demais períodos, os relatórios disponíveis são: ativo, passivo, demonstração do resultado e resumo. Dessa forma, poder-se-á utilizar as informações referentes aos relatórios mais recentes (ativo, passivo, demonstração do resultado e resumo), considerando-se que estão disponíveis em todo o período citado. Rufino *et al*. (2014) apontam que todos os bancos comerciais e múltiplos devem embasar seus relatórios no Plano Contábil das Instituições do Sistema Financeiro Nacional (Cosif), expedido pelo Banco Central do Brasil. Tal plano foi criado e 1987 e apresenta os critérios e procedimentos contábeis a serem observados pelas instituições financeiras, bem como a estrutura de contas e modelos de documentos (BANCO CENTRAL DO BRASIL, 2019). Assim, isso garante a uniformidade dos dados de todas as instituições, independente da robustez da companhia e do período analisado.

O balanço patrimonial de determinada companhia apresenta informações a respeito de sua posição econômica e financeira em determinado momento (ASSAF NETO, 2015). Além disso, a sistematização das informações por meio do balanço constitui um modo conveniente de organizar e resumir tudo o que a empresa possui, o que deve e a diferença entre esses valores. Aquilo que ela detém é definido como ativo, enquanto suas obrigações compõem o passivo (ROSS *et al*., 2013). Já a demonstração do resultado do exercício (DRE) fornece informações sobre os lucros e prejuízos auferidos pela empresa em determinado exercício social, sendo esse resultado consequência de receitas, custos e despesas incorridas pela empresa (ASSAF NETO, 2015).

Visto isso, é importante, agora, analisar com maiores detalhes todas as informações disponibilizadas sobre as instituições financeiras. Como já dito, essas estão decompostas em quatro relatórios: ativo, passivo, demonstração do resultado e resumo. A Tabela 1, a seguir, apresenta todas as informações disponíveis em cada relatório. Considera-se importante salientar que o último relatório da tabela (“geral”) cita as informações que são comuns a todos os quatro relatórios. Na totalidade, há 85 informações disponíveis para cada instituição financeira. Porém, algumas dessas variáveis correspondem a somatórios das demais. Para facilitar a compreensão dessas relações, mostram-se as variáveis acompanhadas de caracteres postos entre parênteses. Assim sendo, segue a tabela contendo os dados à disposição:

Tabela 1 – Dados disponíveis das instituições financeiras

(Continua)

RELATÓRIO	DADOS	DETALHES
Ativo	Disponibilidades (*a*)	
	Aplicações Interfinanceiras de Liquidez (*b*)	
	Títulos e Valores Mobiliários e Instrumentos Financeiros Derivativos (*c*)	
	Operações de Crédito	Operações de Crédito (*d1*)
		Provisão sobre Operações de Crédito (*d2*)
		Operações de Crédito Líquidas de Provisão (*d*)
	Arrendamento Mercantil	Arrendamento Mercantil a Receber (*e1*)
		Imobilizado de Arrendamento (*e2*)
		Credores por Antecipação de Valor Residual (*e3*)
		Provisão sobre Arrendamento Mercantil (*e4*)
		Arrendamento Mercantil Líquido de Provisão (*e*)
	Outros Créditos - Líquidos de Provisão (*f*)	
	Outros Ativos Realizáveis (*g*)	
	Permanente Ajustado (*h*)	
	Ativo Total Ajustado (*i*) = (*a*) + (*b*) + (*c*) + (*d*) + (*e*) + (*f*) + (*g*) + (*h*)	
	Credores por Antecipação de Valor Residual	
	Ativo Total (k) = (*i*) - (*j*)	
Passivo	Depósito Total (*a*)	Depósitos à Vista (*a1*)
		Depósitos de Poupança (*a2*)
		Depósitos Interfinanceiros (*a3*)
		Depósitos a Prazo (*a4*)
		Outros Depósitos (*a5*)
		Depósito Total (*a*)
	Obrigações Por Operações Compromissadas (*b*)	
	Recursos de Aceites e Emissão de Títulos (*c*)	Letras de Crédito Imobiliário (*c1*)
		Letras de Crédito do Agronegócio (*c2*)
		Letras Financeiras (*c3*)
		Obrigações por Títulos e Valores Mobiliários no Exterior (*c4*)
		Outros Recursos de Aceites e Emissão de Títulos (*c5*)
		Recursos de Aceites e Emissão de Títulos (*c*)
	Obrigações por Empréstimos e Repasses (*d*)	
	Captações (*e*) = (*a*) + (*b*) + (*c*) + (*d*)	

(Continuação)

RELATÓRIO	DADOS	DETALHES
Passivo	Instrumentos Derivativos (*f*)	
	Outras Obrigações (*g*)	
	Passivo Circulante e Exigível a Longo Prazo (h) = (*e*) + (*f*) + (*g*)	
	Resultados de Exercícios Futuros (*i*)	
	Patrimônio Líquido (*j*)	
	Passivo Total (*k*) = (*h*) + (*i*) + (*j*)	
Demonstração de Resultado	Receitas de Intermediação Financeira (*a*)	Rendas de Operações de Crédito (*a1*)
		Rendas de Operações de Arrendamento Mercantil (*a2*)
		Rendas de Operações com TVM (*a3*)
		Rendas de Operações com Instrumentos Financeiros Derivativos (*a4*)
		Resultado de Operações de Câmbio (*a5*)
		Rendas de Aplicações Compulsórias (*a6*)
		Receitas de Intermediação Financeira (*a*) = (*a1*) + (*a2*) + (*a3*) + (*a4*) + (*a5*) + (*a6*)
	Despesas de Intermediação Financeira (*b*)	Despesas de Captação (*b1*)
		Despesas de Obrigações por Empréstimos e Repasses (*b2*)
		Despesas de Operações de Arrendamento Mercantil (*b3*)
		Resultado de Operações de Câmbio (*b4*)
		Resultado de Provisão para Créditos de Difícil Liquidação (*b5*)
		Despesas de Intermediação Financeira (*b*) = (*b1*) + (*b2*) + (*b3*) + (*b4*) + (*b5*)
	Resultado de Intermediação Financeira (*c*) = (*a*) + (*b*)	
	Outras Receitas/Despesas Operacionais (*d*)	Rendas de Prestação de Serviços (*d1*)
		Rendas de Tarifas Bancárias (*d2*)
		Despesas de Pessoal (*d3*)
		Despesas de Administrativas (*d4*)
		Despesas Tributárias (*d5*)
		Resultado de Participações (*d6*)
		Outras Receitas Operacionais (*d7*)
		Outras Despesas Operacionais (*d8*)
		Outras Receitas/Despesas Operacionais (*d*) = (*d1*) + (*d2*) + (*d3*) + (*d4*) + (*d5*) + (*d6*) + (*d7*) + (*d8*)
	Resultado Operacional (*e*) = (*c*) + (*d*)	

(Continuação)

RELATÓRIO	DADOS	DETALHES
Demonstração de Resultado	Resultado Não Operacional (*f*)	
	Resultado antes da Tributação, Lucro e Participação (*g*) = (*e*) + (*f*)	
	Imposto de Renda e Contribuição Social (*h*)	
	Participação nos Lucros (*i*)	
	Lucro Líquido (*j*) = (*g*) + (*h*) + (*i*)	
	Juros sobre Capital Social das Cooperativas (*k*)	
Resumo	Ativo Total	
	Carteira de Crédito Classificada	
	Passivo Circulante e Exigível a Longo Prazo e Resultado de Exercícios Futuros	
	Captações	
	Patrimônio Líquido	
	Lucro Líquido	
	Número de Agências	
	Número de Postos de Atendimento	
Geral	Conglomerado	
	Tipo de Consolidado Bancário	
	Tipo de Controle	
	Tipo de Instituição	
	Cidade	
	Data-base do relatório	

Fonte: elaborada pelo autor com base em Banco Central do Brasil (2019).

Diante disso, é fundamental, agora, rever todos os dados disponíveis e propor uma avaliação adequada dessas informações. Inicialmente, entende-se que seja um equívoco inserir todos os dados de forma absoluta, uma vez que os tamanhos das instituições selecionadas diferem expressivamente. Isso é evidenciado ao se compararem os dados das maiores e das menores instituições: em dezembro de 2018, por exemplo, o maior patrimônio líquido da amostra utilizada correspondeu a mais de sete mil vezes o tamanho do menor (BANCO CENTRAL DO BRASIL, 2019). Assim, para lidar com problemas relacionados ao porte das companhias, Ross *et al*. (2013) aconselham o uso de indicadores financeiros. Basicamente, um indicador é um ou mais dados divididos por outro(s). Dessa forma, o problema do tamanho da empresa é eliminado, pois o tamanho é suprimido no processo de divisão (ROSS *et al*., 2013). À vista disso, é necessário estabelecer quais indicadores financeiros são importantes para compor a amostra.

4.3 PROPOSIÇÃO DE INDICADORES FINANCEIROS

Atualmente, não há qualquer conjunto de indicadores que sejam universalmente aceitos pela academia para a avaliação financeira de uma instituição (CAPELLETO; CORRAR, 2008). Além do mais, os indicadores referentes a bancos possuem características distintas das de outros setores, devido às especificidades do ambiente no qual estão inseridos (RUFINO *et al*., 2014). Assim sendo, é pertinente analisarem-se as principais produções específicas sobre o tema, a fim de se selecionarem indicadores que possam ser utilizados para o caso proposto. Assaf Neto (2015) foca seu estudo em instituições financeiras, examinando as principais características de seus balanços patrimoniais. Além do mais, elenca os principais indicadores financeiros e os classifica em três grandes grupos: solvência e liquidez, capital e risco e, por último, rentabilidade e lucratividade. Na construção de um modelo para a previsão de risco bancário, Rocha (1999) elenca uma série de indicadores considerados como importantes para o problema. Nesse caso, a autora os extrai de boletins de uma consultoria (Austin Asis) especializada na análise de risco bancário, sem classificá-los em grupos específicos.

Dessa forma, há uma série de indicadores propostos para analisar problemas específicos de empresas bancárias. Entretanto, é imprescindível que todos os dados necessários para a construção dessas relações estejam à disposição. A Tabela 2, a seguir, apresenta todos os indicadores selecionados, bem como suas fontes e as relações correspondentes.

Tabela 2 – Indicadores selecionados

FONTE	ÍNDICE	RELAÇÃO
Rocha (1999)	Alavancagem	Captação Total / Patrimônio Líquido
Rocha (1999)	Capital de Giro	(Patrimônio Líquido - Ativo Permanente) / Patrimônio Líquido
Rocha (1999)	Custo Administrativo	Despesa Administrativa / Captação Total
Assaf Neto (2015)	Custo Médio de Captação	Despesas Financeiras de Captação de Mercado / Depósitos a Prazo
Rocha (1999)	Custo Pessoal	Despesa Pessoal / Captação Total
Rocha (1999)	Custo Total	Despesa Total / Captação Total
Assaf Neto (2015)	Eficiência Operacional	Despesas Operacionais / Receitas de Intermediação Financeira
Assaf Neto (2015), Rocha (1999)	Encaixe	Disponibilidades / Depósitos a Vista
Assaf Neto (2015)	Imobilização do Capital Próprio	Ativo Permanente / Patrimônio Líquido
Rocha (1999)	Imobilização Própria	(Ativo Permanente - Imobilizado de Arrendamento) / Patrimônio Líquido
Assaf Neto (2015)	Independência Financeira	Patrimônio Líquido / Ativo Total
Assaf Neto (2015)	Índice Empréstimos / Depósitos	Operações de Crédito / Depósitos
Assaf Neto (2015)	Juros Passivos	Despesa de Intermediação Financeira / Passivo Total
Assaf Neto (2015)	Liquidez Imediata	(Disponibilidades + Aplicações Interfinanceiras de Liquidez) / Depósitos a Vista
Assaf Neto (2015)	Lucratividade dos Ativos	Receitas de Intermediação Financeira / Ativo Total
Rocha (1999)	Margem Bruta	Resultado Bruto da Intermediação Financeira / Receita da Intermediação Financeira
Assaf Neto (2015)	Margem Financeira	Resultado Bruto da Intermediação Financeira / Ativo Total
Assaf Neto (2015), Rocha (1999)	Margem Líquida	Lucro Líquido / Receita de Intermediação Financeira
Rocha (1999)	Margem Operacional	Resultado Operacional / Receita da Intermediação Financeira
Assaf Neto (2015)	Participação dos Empréstimos	Operações de Crédito / Ativo total
Assaf Neto (2015)	Relação Capital / Depositantes	Patrimônio Líquido / Depósitos
Assaf Neto (2015), Rocha (1999)	Rentabilidade do PL	Lucro Líquido / Patrimônio Líquido
Assaf Neto (2015)	Retorno Médio das Operações de Crédito	Receitas Financeiras de Operações de Crédito /Operações de Crédito
Assaf Neto (2015)	Retorno sobre o Investimento Total	Lucro Líquido / Ativo Total

Fonte: elaborada pelo autor com base em Assaf Neto (2015) e Rocha (1999).

Assim, esses 24 indicadores financeiros serão utilizados como dados de entrada do algoritmo. Ao se verificarem todas as variáveis indicadas nas relações dos índices, é possível constatar-se que apenas uma não consta na Tabela 1: ativo permanente. Entretanto, seu valor pode ser obtido por meio da soma das variáveis “permanente ajustado” e “imobilizado de arrendamento” (BANCO CENTRAL DO BRASIL, 2019). Desse modo, todos os dados necessários para a construção dos índices estão disponíveis. É possível perceber-se, então, que

os índices selecionados abrangem amplamente a análise de uma instituição financeira. E, tal abrangência tem o intuito de permitir um bom desempenho da árvore de decisão.

5 APLICAÇÃO DO ALGORITMO E DISCUSSÃO SOBRE OS RESULTADOS

Este capítulo apresenta todo o processo envolvido na construção de um modelo explicativo para a falência bancária por meio de um algoritmo, bem como as discussões em torno de seus resultados. Após o relato sobre o tratamento dos dados, será exibida a árvore de decisão principal, além de seus detalhes. Após isso, o algoritmo será reaplicado diversas vezes, para que possa ser aprofundada a discussão sobre sua precisão e sobre as diferentes árvores de decisão que podem ser geradas para explicar o mesmo problema.

5.1 TRATAMENTO DOS DADOS

Em seu estudo, Öcal *et al*. (2015) realizam uma seleção prévia dos dados antes de efetuarem a aplicação do algoritmo C5.0. Essa seleção constitui-se na retirada de dados extremos do modelo, ou *outliers*. Além disso, Lantz (2015) menciona que essa preparação dos dados é extremamente importante para a qualidade do resultado final. Assim sendo, tal procedimento também será realizado na amostra desta pesquisa, a fim de que a árvore de decisão apresente um desempenho satisfatório.

Outliers podem ser definidos como "observações que possuem uma combinação única de características significativamente distintas das demais observações" (HAIR JR. *et al*., 2014). Ch'ng (2016) afirma que é um erro simplesmente ignorá-los, visto que tal atitude pode provocar grandes distorções no resultado final. Por isso, o autor indica que a preparação dos dados é vital para o processo. Contudo, é importante ressaltar que nenhuma técnica estatística é capaz de afirmar com total precisão o que deve ser feito com esses valores extremos (JOHNSON, 1998). Tukey (1973) sugere o uso de diagramas de caixa (*box-and-whisker plots*) como um método eficiente e simples de identificar tais dados. Essa ferramenta utiliza os valores do primeiro e do terceiro quartil (*Q1* e *Q2*, respectivamente) de determinado conjunto de dados, bem como suas diferenças, para construir um intervalo. Todos os valores que estiverem fora do intervalo $[(Q1 - 1{,}5(Q1 - Q3)), (Q3 + 1{,}5(Q1 - Q3))]$ são considerados como possíveis *outliers* (CH'NG, 2016).

Ch'ng (2016) defende a ideia de que os dados extremos devem ser acomodados ou eliminados da amostra. Afirma, também, que a eliminação ou a acomodação desses valores é

uma decisão exclusiva do pesquisador. Importa considerar, no entanto, que ambos os procedimentos têm seus pontos falhos. Osborne (2002) sugere que os *outliers* não sejam removidos da amostra, pois uma grande quantidade de informações pode ser perdida. Dessa forma, o autor sugere que todos os dados, mesmo os extremos, sejam acomodados. Newton e Rudestam (1999), por sua vez, afirmam que a acomodação dos valores extremos gera uma transformação da amostra, podendo alterar as relações originais das variáveis do modelo. Para as duas situações, Ch'ng (2016) assegura que os métodos mais populares para o gerenciamento dos dados são o *Winsorize means* e o *Trimmed means*. Enquanto o primeiro lida com a acomodação dos dados, o segundo os elimina.

Visto que a pesquisa proposta neste estudo trabalha com uma grande quantidade de variáveis independentes para cada variável dependente, a exclusão de empresas que comportam dados extremos parece ser problemática, pois acarretaria uma perda significativa de dados considerados como adequados. Por isso, optou-se por acomodar todos os dados extremos na amostra, em vez de excluí-los. Ainda, a análise da amostra foi realizada dividindo-a em dois grupos: um contendo apenas as empresas classificadas como "sucesso", e outro com apenas as definidas como "insucesso". Para isso, seguiram-se as indicações de Tukey (1973) e Ch'ng (2016): primeiramente, identificaram-se todos os valores que estão fora do intervalo $[(Q1 - 1{,}5(Q1 - Q3)), (Q3 + 1{,}5(Q1 - Q3))]$; após, todos esses dados fora do intervalo foram acomodados nas amostras, por meio do método *Winsorize*. Esse procedimento consiste na remoção de determinado percentil superior e/ou inferior da amostra e na sua substituição pelo valor mais alto e/ou mais baixo remanescente (DIXON, 1960). O valor de tal percentil é determinado previamente pelo pesquisador. Além disso, a aplicação desse processo pode ser realizada dentro do ambiente computacional R.

Dessa forma, a fim de gerenciar os dados extremos, a amostra foi divida em dois grupos. Considerando as duas amostras ("sucesso" e "insucesso"), apenas o indicador "participação em empréstimos" das instituições financeiras de "sucesso" apresentou todos os dados dentro dos limites propostos pelo *boxplot*. Assim, somente esse conjunto de dados deixou de passar por algum tratamento. Todos os demais grupos foram submetidos ao processo *Winsorize*. A fim de se evitar uma grande distorção em relação à estrutura original dos dados, optou-se pela utilização dos percentis 1% e 99%. As modificações atingiram 132 dados do conjunto, representando, aproximadamente, 3% do total. O código do programa utilizado para tal processo pode ser visto no Apêndice C.

A amostra final contém 193 empresas, das quais 75% são consideradas empresas de sucesso. Ao todo, há 25 indicadores. Além das 24 relações já apresentadas na Tabela 2, também foi incluído o indicador "sucesso", que será utilizado como variável dependente. Nesse caso, foi atribuído o valor 1 a empresas de sucesso, enquanto as de insucesso receberam o valor 0. Ao todo, 4.825 variáveis foram utilizadas para a construção da árvore de decisão. É importante ressaltar que, deste total, 229 são dados inexistentes. A grande maioria dos componentes desse grupo não possuem depósitos a prazo ou depósitos a vista, sendo formada, sobretudo, por bancos focados no financiamento de produtos específicos, como carros ou caminhões. Com isso, os indicadores que utilizam algum desses dados, como custo médio de captação, encaixe ou liquidez imediata, tornam-se inexistentes. Entretanto, como o algoritmo utilizado é capaz de lidar com dados inexistentes (LANTZ, 2015), optou-se por deixá-los na amostra.

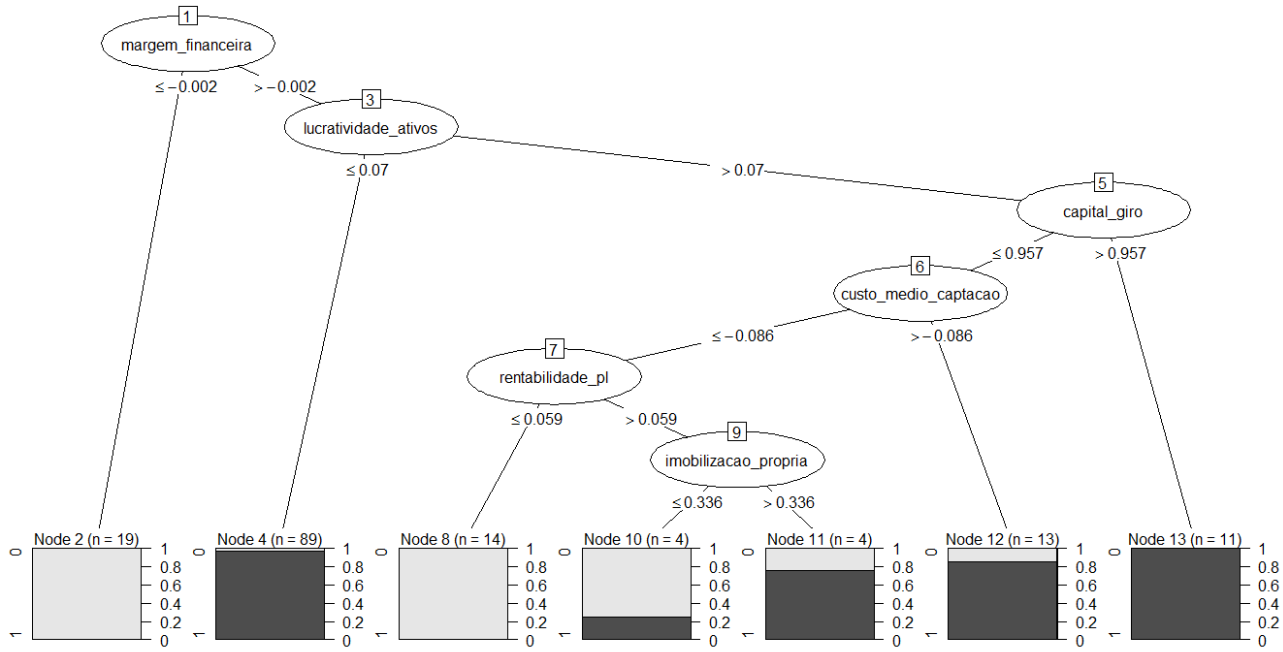
5.2 APLICAÇÃO DO ALGORITMO

A partir dos dados reunidos, foi possível realizar a construção da árvore de decisão. Para isso, utilizou-se o algoritmo C5.0. Como já visto anteriormente, esse algoritmo apresenta vários benefícios que justificam a sua escolha diante dos demais: relativa facilidade de execução (QUINLAN, 1986), de interpretação e alta performance frente a problemas complexos (LANTZ, 2015). Além do mais, convém ressaltar que árvores de decisão são adequadas para lidar com problemas onde a variável de decisão é binária (VASCONCELOS, 2017), como no caso do problema proposto. O programa contendo tal algoritmo foi escrito e processado no *software* R, versão 3.6.1, para Windows. Instruções para a elaboração do programa, bem como algumas nuances do algoritmo, podem ser encontradas em manuais disponibilizados pelo próprio *software*, como o C5.0 Classification Models, ou, então, nas obras de Géron (2019) ou Lantz (2015). O programa completo desenvolvido para a aplicação desse problema pode ser visualizado no Apêndice B.

Lantz (2015) indica que a amostra deve ser divida em duas partes: uma com o objetivo de ajustar o modelo, e outra a fim de avaliar o modelo construído. Nesse caso, 20% dos dados da amostra foram utilizados para compor o segundo grupo. Além disso, de acordo com as categorias de *machine learning* analisadas por Géron (2019), pode-se afirmar que o aprendizado do algoritmo é supervisionado (pois os dados usados no ajuste já contêm as

soluções esperadas) em lotes (visto que, nesse caso, não há como os dados serem inseridos de maneira incremental) e, por fim, baseado em exemplos (já que o sistema age em comparação com os exemplos já conhecidos). Visto isso, as figuras abaixo apresentam, de forma visual, a árvore de decisão gerada pelo programa. A primeira imagem exibe a árvore de maneira gráfica, tendo sido gerada por intermédio da função *plot*, no *software* R. Já a segunda a apresenta na forma como na qual foi gerada pelo algoritmo C5.0.

Figura 2 - Árvore de decisão, expressa de forma gráfica



Fonte: elaborada pelo autor (2019).

Figura 3 – Árvore de decisão, gerada pelo algoritmo C5.0

```
margem_financeira <= -0.002031436: 0 (19)
margem_financeira > -0.002031436:
:...lucratividade_ativos <= 0.06966712: 1 (89/3)
    lucratividade_ativos > 0.06966712:
    :...capital_giro > 0.9567282: 1 (11)
        capital_giro <= 0.9567282:
        :...custo_medio_captacao > -0.0860748: 1 (13.4/2)
            custo_medio_captacao <= -0.0860748:
            :...rentabilidade_pl <= 0.05905775: 0 (14)
                rentabilidade_pl > 0.05905775:
                :...imobilizacao_propria <= 0.3356186: 0 (4.6/0.6)
                    imobilizacao_propria > 0.3356186: 1 (3)
```

Fonte: elaborada pelo autor (2019).

Assim, é possível perceber que seis indicadores são utilizados para explicar o sucesso ou insucesso das instituições financeiras. Além disso, a árvore apresenta, ao todo, 13 nós. Desses nós, seis são de decisão, enquanto sete são terminais. Vale ressaltar, também, o significado dos nós terminais presentes na Figura 2: o retângulo claro representa "insucesso", enquanto o escuro significa "sucesso". Ademais, a leitura da Figura 2 é iniciada no canto superior esquerdo da imagem e encerra quando o nó terminal é atingido. Na Figura 3, por sua vez, os nós terminais são indicados pelos números 0 e 1, que representam, respectivamente, "insucesso" e "sucesso". Dessa forma, a leitura inicia na parte superior e é finalizada quando um desses valores é encontrado. Portanto, segundo o modelo gerado, uma instituição financeira que apresenta uma margem financeira menor ou igual a -0,002, por exemplo, será classificada como "insucesso". Assim como essa, todas as demais alternativas de classificação podem ser visualizadas nas Figuras 2 e 3.

Quando aplicado na amostra-teste escolhida, o modelo apresentou um nível de assertividade de 89,7%, ou seja, dos 39 casos, classificou corretamente a variável "sucesso" em 35. Foram detectados quatro erros, dos quais três correspondem ao caso em que a empresa havia apresentado sucesso, enquanto o outro corresponde a insucesso. A Figura 4 apresenta essas informações compiladas numa tabela. Além disso, a figura mostra, também, os percentuais correspondentes a cada situação em relação a toda a amostra.

Figura 4 - Precisão de acerto do algoritmo C5.0, quando aplicado sobre a amostra-teste correspondente a 20% do total

	Sucesso Projetado		
Sucesso Real	0	1	Total
0	6	1	7
	0.154	0.026	
1	3	29	32
	0.077	0.744	
Total	9	30	39

Fonte: elaborada pelo autor (2019).

Desse modo, é possível constatar que 74,4% dos dados correspondem a variáveis classificadas como "sucesso" e que foram projetadas corretamente pelo modelo, enquanto 15,4% correspondem a variáveis determinadas como "insucesso" e que também foram classificadas corretamente. Somando-se ambos os valores, encontra-se a precisão total do

modelo (89,7%). Quando comparado aos demais trabalhos sobre o tema, esse nível de assertividade é considerado satisfatório. Dessa forma, não há necessidade de o modelo passar por um processo de melhoria. A árvore de decisão apresentada por Öcal *et al*. (2015) foi capaz de explicar 85,13% do problema proposto. Já Mullaimathan e Spiess (2017), ao compararem diversos métodos de previsão, encontraram seu melhor sistema preditivo numa árvore de decisão com acurácia de 85,1%.

É importante observar, também, que os seis indicadores utilizados pelo modelo para explicar o sucesso ou o insucesso são utilizados de maneira distinta pela árvore. Como é possível perceber na Figura 2, todos os problemas iniciam por "margem financeira". Isso torna essa relação a mais utilizada e, consequentemente, a mais relevante para a explicação do problema proposto. A Tabela 3 apresenta a relação de todos os indicadores utilizados na árvore de decisão, em ordem de seus percentuais de utilização:

Tabela 3 – Indicadores utilizados na árvore de decisão

INDICADOR	UTILIZAÇÃO
Margem Financeira	100,00%
Lucratividade dos Ativos	87,66%
Capital de Giro	29,87%
Custo Médio de Captação	22,08%
Rentabilidade do PL	14,29%
Imobilização Própria	5,19%

Fonte: elaborada pelo autor (2019).

Percebe-se que o indicador margem financeira foi utilizado para explicar 100% da amostra, enquanto a imobilização própria só esteve presente na explicação de 5,19% da amostra ajustada. Dada a importância desses indicadores para o modelo construído, torna-se necessário detalhar cada um, indicando suas características e pontos de atenção.

Quatro desses indicadores são classificados por Assaf Neto (2015) dentro da categoria "rentabilidade e lucratividade": margem financeira, lucratividade dos ativos, custo médio de captação e rentabilidade do patrimônio líquido (PL). O autor afirma que, diferente de outros setores, as empresas financeiras costumam apresentar um retorno sobre o investimento baixo. Dessa forma, o valor do primeiro nó, por exemplo, parece plausível, já que considera um valor de margem financeira relativamente baixo. Capital de giro, por sua vez, é definido por

Brom e Balian (2007) como “os recursos a serem desembolsados antes dos recebimentos das vendas dos produtos ou serviços de uma empresa e que são necessários para mantê-la funcionando”. Ainda, afirmam que o indicador é de extrema importância, pois reflete o que ocorre na organização, como sua capacidade de pagar em dia suas obrigações. Por fim, imobilização própria é definida por Rocha (1999) como a relação entre a diferença do ativo permanente e do imobilizado do arrendamento sobre o patrimônio líquido, conforme é possível visualizar na Tabela 2. Assaf Neto (2015) especifica que quanto menor o nível de uso de capital próprio, maior será o risco operacional do banco. Entretanto, o autor alerta que o indicador não leva em consideração a qualidade dos ativos e, por isso, não deve ser analisado de maneira isolada.

5.3 NOVAS APLICAÇÕES DO ALGORITMO

É importante observar que a árvore de decisão apresentada nas Figuras 2 e 3, bem como sua precisão, é válida apenas para a ordenação dos dados proposta e para o tamanho da amostra-teste estabelecido. Dessa forma, sempre que houver uma alteração de qualquer um desses fatores, será gerada uma árvore de decisão inédita e, consequentemente, com uma estrutura distinta. Assim, optou-se por realizar novas aplicações do algoritmo com diferentes ordenações e tamanhos da amostra-teste, a fim de enriquecer a discussão e verificar outros resultados possíveis para o problema.

Inicialmente, o algoritmo foi executado outras vezes, apenas com o intuito de se verificar o nível de precisão do modelo. Para isso, novas ordenações e tamanhos da amostra-teste foram propostos. Isso permite uma observação mais assertiva sobre a qualidade do algoritmo para prever o caso proposto. Sendo assim, reaplicou-se o programa 75 vezes, considerando-se cinco faixas de tamanho da amostra destinada ao ajuste do modelo e 15 ordenações distintas para cada uma dessas faixas. A Tabela 4 apresenta os resultados sintetizados, contendo o percentual máximo de precisão obtido em cada faixa de tamanho da amostra, o valor mínimo, o médio e a variância dos dados:

Tabela 4 – Precisões do algoritmo para o caso proposto

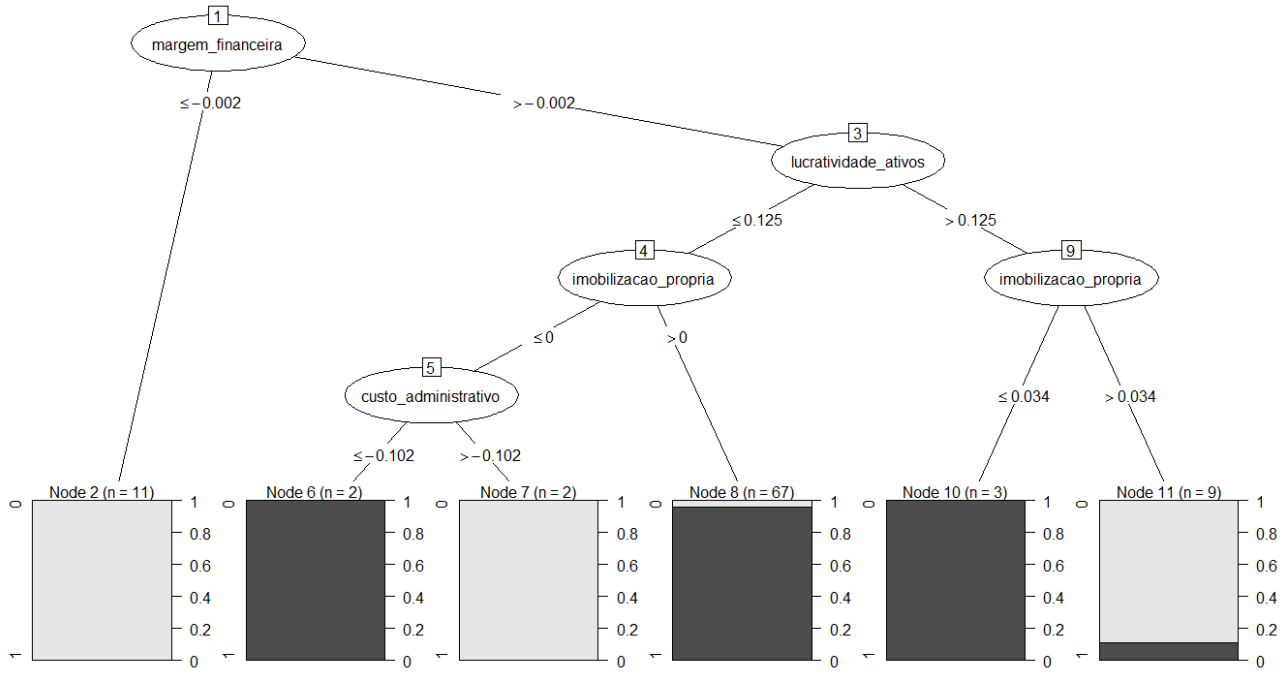
PERCENTUAL DA AMOSTRA DESTINADA AO AJUSTE DO MODELO	MÁXIMO	MÍNIMO	MÉDIO	VARIÂNCIA
70%	93,1%	72,4%	85,5%	0,22%
75%	93,8%	77,1%	86,7%	0,26%
80%	94,9%	76,9%	86,7%	0,28%
85%	100,0%	75,9%	86,4%	0,43%
90%	100,0%	78,9%	89,1%	0,45%

Fonte: elaborada pelo autor (2019).

Constata-se que a precisão média do C5.0 para o caso proposto situa-se entre 85,5% e 89,1%. A amostra contendo 90% dos dados destinados ao ajuste do modelo apresentou os maiores valores médios. Entretanto, a variância dos resultados obtidos para essa faixa de tamanho é significativamente maior que a das três primeiras. E essa variância mais elevada pode comprometer a qualidade dos resultados.

Além da faixa de precisão, considera-se importante apresentar em detalhes uma nova árvore de decisão, a fim de expor suas diferenças em relação ao modelo já apresentado. Nessa nova aplicação, alterou-se significativamente o tamanho da amostra-teste, para que possíveis diferenças fossem verificadas. Assim, a amostra-teste para esse caso correspondeu a 50% do tamanho da amostra (ou, mais precisamente, a 95 empresas). As Figuras 5 e 6 mostram essa nova árvore. Essas figuras estão dispostas da mesma forma que as da sessão anterior: a primeira exibe a árvore de forma gráfica, enquanto a segunda a expõe da exata forma pela qual foi gerada pelo algoritmo.

Figura 5 - Nova árvore de decisão, expressa de forma gráfica



Fonte: elaborada pelo autor (2019).

Figura 6 - Nova árvore de decisão, gerada pelo algoritmo C5.0

```
margem_financeira <= -0.002031436: 0 (11)
margem_financeira > -0.002031436:
:...lucratividade_ativos > 0.1254252:
    :...imobilizacao_propria <= 0.03399373: 1 (3)
    :   imobilizacao_propria > 0.03399373: 0 (9/1)
    lucratividade_ativos <= 0.1254252:
    :...imobilizacao_propria > 0.0003774184: 1 (66/2)
        imobilizacao_propria <= 0.0003774184:
        :...custo_administrativo <= -0.1024727: 1 (2)
            custo_administrativo > -0.1024727: 0 (3)
```

Fonte: elaborada pelo autor (2019).

As figuras mostram que esta nova árvore gerada apresentou uma estrutura distinta em relação ao primeiro modelo. Inicialmente, esta é menor, contendo 11 nós. Desses, seis são terminais. Além disso, apenas quatro indicadores foram utilizados para explicar o sucesso ou insucesso financeiro de uma instituição financeira. Desses indicadores, três apareceram, também, na primeira árvore: margem financeira, lucratividade dos ativos e imobilização própria.

O novo modelo apresentou uma precisão de 81,8%, estando próximo das médias exibidas na Tabela 4. Das 99 empresas presentes na amostra-teste, 81 foram classificadas

corretamente. Nos casos em que houve erro, 10 correspondiam a companhias de “insucesso” e que foram classificadas como “sucesso”, enquanto oito correspondiam ao caso contrário. A figura 7, abaixo, apresenta essas informações de forma sintética:

Figura 7 - Precisão de acerto do algoritmo C5.0 quando aplicado sobre a amostra-teste correspondente a 50% do total

	Sucesso Projetado		
Sucesso Real	0	1	Total
0	15	10	25
	0.152	0.101	
1	8	66	74
	0.081	0.667	
Total	23	76	99

Fonte: elaborado pelo autor (2019).

Assim, nota-se que, por mais que o nível de precisão tenha sido adequado, esse modelo apresentou um desequilíbrio: 55% dos erros de previsão correspondem aos casos em que a empresa deveria ter sido classificada como “insucesso”. Entretanto, esse grupo de empresas correspondia a apenas 25% da amostra total.

Da mesma forma que na primeira árvore apresentada, os indicadores possuem grau de importância distinto para explicar o problema proposto. A Tabela 5 expõe todos os indicadores em ordem de utilização:

Tabela 5 – Indicadores utilizados na árvore de decisão

INDICADOR	UTILIZAÇÃO
Margem Financeira	100,00%
Lucratividade dos Ativos	88,30%
Imobilização Própria	88,30%
Custo Administrativo	5,32%

Fonte: elaborada pelo autor (2019).

A relação margem financeira é utilizada na explicação de todas as empresas, enquanto custo administrativo é utilizado na elucidação de apenas 5,32% das companhias. É relevante

observar-se que os três indicadores que também aparecem na primeira aplicação (margem financeira, lucratividade dos ativos e imobilização própria) são os mais utilizados neste caso.

Logo, é possível interpretar esse três indicadores presentes em ambas as árvores de decisão (margem financeira, lucratividade dos ativos e imobilização própria) como essenciais para a análise ou a administração de uma instituição financeira, visto que são determinantes para a continuidade de tais empresas. Desses indicadores, margem financeira e lucratividade dos ativos são classificados por Assaf Neto (2015) dentro da categoria "rentabilidade e lucratividade". Já o indicador imobilização própria, segundo o mesmo autor, está relacionado com a qualidade dos ativos da empresa e, consequentemente, com seu risco operacional.

6 CONCLUSÃO

Este estudo investigou diferentes sistemas de *machine learning*, enfatizando suas possíveis aplicações na análise econômica, bem como suas diferenças em relação aos métodos, tradicionalmente, empregados nessa análise. Além disso, optou-se por abordar um problema de maneira prática, criando uma árvore de decisão, a fim de projetar a possível falência de uma instituição financeira. A aplicação assentou-se no cenário brasileiro, onde estudos práticos semelhantes ainda são escassos. Dessa forma, pôde-se observar tanto se o resultado encontrado esteve alinhado com os apontamentos da academia, bem como os desafios no processo de recolhimento de dados e na construção do programa para a aplicação do algoritmo.

A árvore de decisão apresentada nas Figuras 2 e 3 foi capaz de classificar com precisão 89,7% das empresas contidas na amostra-teste. Visto que esse percentual de precisão só pode ser atribuído à ordenação dos dados proposta e ao tamanho da amostra-teste aplicado, executou-se o programa diversas vezes, com ordenações e tamanhos diferentes, para que a sua precisão pudesse ser observada com maior acurácia. A Tabela 4 exibiu as precisões máximas, mínimas, médias e suas variâncias para cinco tamanhos de amostra-teste. Enquanto o modelo proposto apresentou precisão média entre 85,5% e 89,1% (variando de acordo com o tamanho da amostra-teste), as árvores de decisão apresentadas por Öcal *et al*. (2015) e Mullaimathan e Spiess (2017) foram capazes de prever corretamente 85,13% e 85,10% do problema, respectivamente. Assim, ao se compararem os resultados deste estudo com os demais, considera-se satisfatório o nível de precisão do modelo apresentado. Consequentemente, a árvore de decisão, bem como o recolhimento e preparo dos dados, também foram exitosos.

Além disso, uma nova árvore de decisão foi apresentada, com um tamanho de amostra-teste significativamente distinto em relação à da primeira aplicação. Com isso, ficou evidenciado que os modelos gerados pelo algoritmo não são únicos. Qualquer reordenação dos dados pode resultar em uma árvore distinta. Entretanto, pôde-se perceber algumas semelhanças entre as duas árvores apresentadas: três indicadores – margem financeira, lucratividade dos ativos e imobilização própria – são utilizados nos dois modelos para explicar a solvência de uma instituição financeira. Desses indicadores, margem financeira e lucratividade dos ativos podem ser considerados como os principais, visto que o primeiro é utilizado para explicar 100% dos dados nas duas árvores, enquanto o segundo possui

utilização superior a 80% em ambos os modelos. Assaf Neto (2015) aponta que esses dois indicadores são relacionados à rentabilidade e lucratividade das empresas. Dessa forma, evidencia-se a importância que esse segmento merece por parte do setor, ainda que se entenda que os valores devam ser analisados por região, e não, de maneira exata.

O estudo, entretanto, apresentou algumas limitações. O algoritmo abordado suporta apenas dados bidimensionais. Dessa forma, impossibilitou-se uma abordagem envolvendo séries temporais de todos os indicadores. Acredita-se que uma aplicação envolvendo a evolução dos indicadores em um determinado período de tempo poderia apresentar resultados ainda mais enriquecedores. Destaca-se que todos os dados recolhidos são disponibilizados pelo Banco Central do Brasil em sua página e abrangem o período de 1994 a 2019. Esses dados, porém, são limitados e não englobam todas as informações sobre as instituições financeiras. Com isso, muitos indicadores propostos pela literatura não puderam ser utilizados, pois as informações necessárias para suas construções não estavam disponíveis nessa fonte. Outra limitação é a impossibilidade de o programa desenvolvido aprender de maneira incremental. Isso pode limitar seu uso em casos que envolvam bancos de dados maiores e que necessitem de atualizações com uma frequência maior. Por fim, é de se salientar que o estudo não considerou mudanças macroeconômicas e setoriais ocorridas durante o período analisado.

Visto que a aplicação do C5.0 não exige do pesquisador um conhecimento avançado sobre ciência computacional, outros estudos envolvendo classificação podem ser desenvolvidos pelas mais diversas áreas acadêmicas. Entende-se, também, que diversos problemas econômicos podem ser abordados por árvores de decisão, principalmente para casos brasileiros. As abordagens no país ainda são escassas, permitindo que uma vasta gama de problemas seja explorada. Sugere-se, então, que outros algoritmos também sejam explorados, aproveitando-se os benefícios de suas particularidades.

REFERÊNCIAS

ALBERTIN, Alberto Luiz. Comércio Eletrônico: um estudo no setor bancário. **Revista de Administração Contemporânea**, v. 3, n. 1, p. 47-70, abr. 1999.

ASSAF NETO, Alexandre. **Estrutura e análise de balanços**: um enfoque econômico-financeiro. 11. ed. São Paulo: Atlas, 2015.

BANCO CENTRAL DO BRASIL. **Consulta de instituições sob regime especial**. Disponível em: <https://www.bcb.gov.br/acessoinformacao/legado?url=https:%2F%2Fwww4.bcb.gov.br%2Flid%2Fliquidacao%2Fd1946%2Fconsulta_form.asp%3Fidpai%3Dregesp>. Acesso em: 07 out. 2019.

BANCO CENTRAL DO BRASIL. **Dados selecionados**, c2019. Disponível em: <https://www3.bcb.gov.br/ifdata/index.html>. Acesso em: 09 out. 2019.

BANCO CENTRAL DO BRASIL. **Plano Contábil das Instituições do Sistema Financeiro Nacional (COSIF)**. Disponível em: <https://www.bcb.gov.br/acessoinformacao/legado?url=https:%2F%2Fwww.bcb.gov.br%2Fhtms%2Fcosif%2Fdefault.asp>. Acesso em: 07 out. 2019.

BANCO CENTRAL DO BRASIL. **Relação de Instituições em Funcionamento no País (transferência de arquivos)**, c2019. Disponível em: <https://www.bcb.gov.br/estabilidadefinanceira/relacao_instituicoes_funcionamento>. Acesso em: 02 out. 2019.

BANCO CENTRAL DO BRASIL. **Resolução 2.099**, de 17 de agosto de 1994. (??? Completar...)

BRASIL BOLSA BALCÃO. **Consultas**. Disponível em: <http://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/renda-variavel/acoes/consultas/classificacao-setorial>. Acesso em: 01 out. 2019.

BRASIL BOLSA BALCÃO. **Empresas com listagem cancelada no mercado de bolsa**. Disponível em: <http://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/renda-variavel/acoes/consultas/empresas-com-listagem-cancelada-no-mercado-de-bolsa>. Acesso em: 05 out. 2019.

BREIMAN, Leo. Arcing Classifiers. **The Annals of Statistics**, v. 26, n. 3, p. 801-824, June, 1998.

BREIMAN, Leo. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). **Statistical Science**, v. 16, n. 3, p. 199-231, 2001.

BROM, Luiz Guilherme; BALIAN, José Eduardo Amato. **Análise de investimentos e capital de giro**. 2. ed. São Paulo: Saraiva, 2007.

CAPELLETO, Lucio Rodrigues; CORRAR, Luiz João. Índices de risco sistêmico para o setor bancário. **Revista Contabilidade & Finanças**, v. 19, n. 47, p. 6-18, ago. 2008.

CARBONELL, Jaime; MICHALSKI, Ryszard; MITCHELL, Tom. **Machine learning**: an artificial intelligence approach. Palo Alto: Springer, 1983.

CH'NG, Chee Keong. **Winsorize tree algorithm for handling outliers in classification problem**. 2015. Tese (Pós-doutorado) – Universiti Utara Malaysia, Kedah Darul Aman, 2016.

DA SILVA, Luiza Maria Oliveira. **Uma Aplicação de Árvores de Decisão, Redes Neurais e KNN para a Identificação de Modelos ARMA Não-Sazonais e Sazonais**. 2005. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2005.

DIXON, W. J. Simplified Estimation from Censored Normal Samples. **The Annals of Mathematical Statistics**, v. 31, n. 2, p. 385-391, June, 1960.

EUROMONITOR INTERNATIONAL. **Finance and Insurance in Brazil**. Março, 2019. Disponível em: <portal.euromonitor.com>. Acesso em: 09 out. 2019.

FERNÁNDEZ-DELGADO, Manuel; CERNADAS, Eva; BARRO, Senén; AMORIM, Dinani. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? **Journal of Machine Learning Research**, v. 15, n. 1, p. 3133-3181, Oct. 2014.

FILDES, Robert; STEKLER, Herman. The state of macroeconomic forecasting. **Journal of Macroeconomics**, v. 24, n. 4, p. 435-468, Autumn, 2002.

GÉRON, Aurélien. **Hands-on Machine Learning with Scikit-Learn, Keras & TernsorFlow**. 2. ed. Sebastopol: O'Reilly Media, 2019.

HAIR Jr., Joseph F.; BLACK, William C.; BABIN, Barry J.; ANDERSON, Rolph E. **Multivariate Data Analysis**. 7. ed. London: Pearson Education, 2014.

HOI, Steven C. H.; SAHOO, Doyen; LU, Jing; ZHAO, Peilin. Online Learning: a comprehensive survey. **SMU Technical Report**, v. 1, n. 1, Oct. 2018.

JOHSON, Dallas E. **Applied multivariate method for data analysis**. Pacific Grove: Duxbury Press, 1998.

JORDAN, M. I.; MITCHELL, T. M.. Machine learning: trends, perspectives and prospects. **Science**, v. 349, n. 6245, p. 255-60, July 2015.

LANTZ, Brett. **Machine Learning with R**. 2. ed. Birmingham: Pack Publishing, 2015.

MITCHELL, Tom M. **Machine Learning**. 1. ed. New York: McGraw-Hil, 1997.

MULLAINATHAN, Sendhil; SPIESS, Jann. Machine Learning: An Applied Econometric Approach. **Journal of Econometric Perspectives**, v. 31, n. 2, p. 87-106, Spring 2017.

NEWTON, Era R.; RUDESTAM, Kjell Erik. **Your statistical consultant:** answers to your data analysis questions. Newbury Park: Sage, 1999.

NYMAN, Rickard; ORMEROD, Paul. **Predicting Economic Recessions Using Machine Learnings Algorithms**. New York: Cornell University, 2017.

ÖCAL, Nurcan; ERCAN, Metin Kamil; KADIOĞLU, Eyüp. Predicting Financial Failuere Using Decision Tree Algorithms: an empirical test on the manufacturing industry at Borsa. Istambul. **Canadian Center of Science and Education**, v. 7, n. 7, p. 189-206, June 2015.

OSBORNE, Jason W. Notes on the Use of Data. **Transformation, Practical Assessment, Research & Evaluation**, v. 8, n. 6, p. 1-7, May 2002.

PINTO, Hugo. The role of econometrics in economic science: an essay about the monopolization of economic methodoly by econometric methods. **Journal of Socio-Economics**, v. 40, n. 4, p. 436- 443, Apr. 2011.

QUINLAN, John. Induction of Decision Trees. **Machine Learning**, v.1, n.1, p. 81-106, 1986.

ROCHA, Fabiana. Previsão de falência bancária: um modelo de risco proporcional. **Pesquisa e Planejamento Econômico**, Rio de Janeiro, v. 29, n. 1, p. 137-152, abr. 1999.

ROSS, Stephen A.; WESTERFIELD, Randolph W.; JORDAN, Bradford D.; LAMB, Roberto. **Fundamentos de Administração Financeira**. 9. ed. Porto Alegre: AMGH, 2013.

RUFINO, Maria Audenôra; MAZER, Lílian Perobon; MACHADO, Márcia Reis; CAVALCANTE, Paulo Roberto Nóbrega. Sustentabilidade e *Performance* dos Indicadores de Rentabilidade e Lucratividade: um estudo comparativo entre os bancos integrantes e não integrantes do ISE da BMF&Bovespa. **Revista Ambiente Contábil**, v. 6, n. 1, p. 1-18, jun. 2014.

SOKOLOV-MLADENOVIC, Svetlana; MILOVANCEVIC, Milos; MLADENOVIC, Igor; ALIZAMIR, Meysam. Economic growth forecasting by artificial neural network with extreme learning machine based on trade, import and export parameters. **Computers in Human Behavior**, vol. 65, n. 1, p. 43-45, Dec. 2016.

THE COMPREHENSIVE R ARCHIVE NETWORK. **C5.0 Classification Models**. Disponível em: <https://cran.rstudio.com/web/packages/C50/vignettes/C5.0.html#classification-trees>. Acesso em: 14 out. 2019.

TUKEY, John W. **Exploratory Data Analysis**. Boston: Addison-Wesley Publishing Company, 1977.

VARIAN, Hal R. Big Data: New Tricks for Econometrics. **The Journal of Economic Perspectives**, v. 28, n. 2, p. 3-27, Spring, 2014.

VASCONCELOS, Bruno Freitas Boynad de. **Poder preditivo de métodos de Machine Learning com processo de seleção de variáveis**: uma aplicação às projeções de produto de países. 2017. Tese (Doutorado em Economia) – Universidade de Brasília, Brasília, 2017.

WOLPERT, David. Stacked Generalization. **Neural Networks**, v.5, n.1, p. 241-259, 1992.

APÊNDICE A – INSTITUIÇÕES FINANCEIRAS UTILIZADAS NO ALGORITMO, INCLUINDO SEUS SEGMENTOS E CLASSIFICAÇÕES

Tabela 6 – Instituições financeiras utilizadas na aplicação do algoritmo

(Continua)

NOME DA INSTITUIÇÃO	SEGMENTO	CLASSIFICAÇÃO
BANCO A.J. RENNER S.A.	Banco Múltiplo	Sucesso
BANCO ABC BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO ABN AMRO S.A.	Banco Múltiplo	Sucesso
BANCO AGIBANK S.A.	Banco Comercial	Sucesso
BANCO AGRIMISA S.A.	Banco Múltiplo	Insucesso
BANCO ALFA S.A.	Banco Comercial	Sucesso
BANCO ANDBANK (BRASIL) S.A.	Banco Múltiplo	Sucesso
BANCO APLICAP S.A.	Banco Comercial	Insucesso
BANCO ARAUCARIA S.A.	Banco Múltiplo	Insucesso
BANCO ARBI S.A.	Banco Comercial	Sucesso
BANCO AZTECA DO BRASIL S.A.	Banco Múltiplo	Insucesso
BANCO B3 S.A.	Banco Comercial	Sucesso
BANCO BAMERINDUS DO BRASIL S.A.	Banco Múltiplo	Insucesso
BANCO BANDEPE S.A.	Banco Múltiplo	Sucesso
BANCO BARI DE INVESTIMENTOS E FINANCIAMENTOS S.A.	Banco Múltiplo	Sucesso
BANCO BMD S.A.	Banco Comercial	Insucesso
BANCO BMG S.A.	Banco Múltiplo	Sucesso
BANCO BNP PARIBAS BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO BOCOM BBM S.A.	Banco Múltiplo	Sucesso
BANCO BRADESCARD S.A.	Banco Múltiplo	Sucesso
BANCO BRADESCO BBI S.A.	Banco Múltiplo	Sucesso
BANCO BRADESCO BERJ S.A.	Banco Múltiplo	Sucesso
BANCO BRADESCO CARTÕES S.A.	Banco Múltiplo	Sucesso
BANCO BRADESCO FINANCIAMENTOS S.A.	Banco Múltiplo	Sucesso
BANCO BRADESCO S.A.	Banco Múltiplo	Sucesso
BANCO BRASILEIRO COMERCIAL S.A	Banco Múltiplo	Insucesso
BANCO BRJ S.A.	Banco Múltiplo	Insucesso
BANCO BS2 S.A.	Banco Múltiplo	Sucesso
BANCO BTG PACTUAL S.A.	Banco Múltiplo	Sucesso
BANCO BVA S.A.	Banco Múltiplo	Insucesso
BANCO C6 S.A.	Banco Múltiplo	Sucesso
BANCO CAIXA GERAL - BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO CAPITAL S.A.	Banco Múltiplo	Sucesso
BANCO CARGILL S.A.	Banco Múltiplo	Sucesso
BANCO CATERPILLAR S.A.	Banco Múltiplo	Sucesso
BANCO CEDULA S.A.	Banco Múltiplo	Sucesso
BANCO CETELEM S.A.	Banco Múltiplo	Sucesso
BANCO CIFRA S.A.	Banco Múltiplo	Sucesso
BANCO CITIBANK S.A.	Banco Múltiplo	Sucesso
BANCO CLASSICO S.A.	Banco Múltiplo	Sucesso
BANCO CNH INDUSTRIAL CAPITAL S.A.	Banco Múltiplo	Sucesso
BANCO COMERCIAL BANCESA S.A.	Banco Múltiplo	Insucesso
BANCO CRÉDIT AGRICOLE BRASIL S.A.	Banco Múltiplo	Sucesso

(Continuação)

NOME DA INSTITUIÇÃO	SEGMENTO	CLASSIFICAÇÃO
BANCO CREDIT SUISSE (BRASIL) S.A.	Banco Múltiplo	Sucesso
BANCO CREFISA S.A.	Banco Comercial	Sucesso
BANCO CREFISUL S.A.	Banco Múltiplo	Insucesso
BANCO CRUZEIRO DO SUL S.A	Banco Múltiplo	Insucesso
BANCO CSF S.A.	Banco Múltiplo	Sucesso
BANCO DA AMAZONIA S.A.	Banco Comercial	Sucesso
BANCO DA CHINA BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO DAYCOVAL S.A.	Banco Múltiplo	Sucesso
BANCO DE LAGE LANDEN BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO DIGIO S.A.	Banco Múltiplo	Sucesso
BANCO DO BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO DO ESTADO DE SERGIPE S.A.	Banco Múltiplo	Sucesso
BANCO DO ESTADO DO PARÁ S.A.	Banco Múltiplo	Sucesso
BANCO DO ESTADO DO PIAUI S.A.	Banco Comercial	Insucesso
BANCO DO ESTADO DO RIO GRANDE DO SUL S.A.	Banco Múltiplo	Sucesso
BANCO DO NORDESTE DO BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO DO PROGRESSO S.A.	Banco Múltiplo	Insucesso
BANCO DRACMA S.A.	Banco Múltiplo	Insucesso
BANCO ECONÔMICO S.A.	Banco Múltiplo	Insucesso
BANCO EMPRESARIAL S.A.	Banco Múltiplo	Insucesso
BANCO FATOR S.A.	Banco Múltiplo	Sucesso
BANCO FIBRA S.A.	Banco Múltiplo	Sucesso
BANCO FICSA S.A.	Banco Múltiplo	Sucesso
BANCO FIDIS S/A	Banco Múltiplo	Sucesso
BANCO FINAXIS S.A.	Banco Comercial	Sucesso
BANCO FORD S.A.	Banco Múltiplo	Sucesso
BANCO GMAC S.A.	Banco Múltiplo	Sucesso
BANCO GUANABARA S.A.	Banco Múltiplo	Sucesso
BANCO HEXABANCO S.A.	Banco Múltiplo	Insucesso
BANCO HONDA S.A.	Banco Múltiplo	Sucesso
BANCO HYUNDAI CAPITAL BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO IBM S.A.	Banco Múltiplo	Sucesso
BANCO INBURSA S.A.	Banco Múltiplo	Sucesso
BANCO INDUSTRIAL DO BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO INDUSVAL S.A.	Banco Múltiplo	Sucesso
BANCO INTER S.A.	Banco Múltiplo	Sucesso
BANCO INTERFINANCE S.A.	Banco Múltiplo	Insucesso
BANCO INTERIOR DE SÃO PAULO S.A.	Banco Múltiplo	Insucesso
BANCO INTERPART S.A.	Banco Múltiplo	Insucesso
BANCO INTERUNION S.A.	Banco Múltiplo	Insucesso
BANCO INVESTCORP S.A.	Banco Comercial	Insucesso
BANCO INVESTCRED UNIBANCO S.A.	Banco Comercial	Sucesso
BANCO ITAÚ BBA S.A.	Banco Múltiplo	Sucesso
BANCO ITAÚ CONSIGNADO S.A.	Banco Múltiplo	Sucesso
BANCO ITAÚ VEÍCULOS S.A.	Banco Múltiplo	Sucesso
BANCO ITAUBANK S.A.	Banco Múltiplo	Sucesso
BANCO ITAUCARD S.A.	Banco Múltiplo	Sucesso
BANCO ITAULEASING S.A.	Banco Múltiplo	Sucesso
BANCO J. SAFRA S.A.	Banco Múltiplo	Sucesso
BANCO J.P. MORGAN S.A.	Banco Múltiplo	Sucesso
BANCO JOHN DEERE S.A.	Banco Múltiplo	Sucesso
BANCO KDB DO BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO KEB HANA DO BRASIL S.A.	Banco Comercial	Sucesso

(Continuação)

NOME DA INSTITUIÇÃO	SEGMENTO	CLASSIFICAÇÃO
BANCO KOMATSU DO BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO LAVRA S.A.	Banco Múltiplo	Insucesso
BANCO LOSANGO S.A. - BANCO MÚLTIPLO	Banco Múltiplo	Sucesso
BANCO LUSO BRASILEIRO S.A.	Banco Múltiplo	Sucesso
BANCO MAIS S.A.	Banco Múltiplo	Insucesso
BANCO MARTINELLI S.A.	Banco Múltiplo	Insucesso
BANCO MÁXIMA S.A.	Banco Comercial	Sucesso
BANCO MAXINVEST S.A.	Banco Múltiplo	Sucesso
BANCO MERCANTIL DO BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO MERCANTIL S.A.	Banco Múltiplo	Insucesso
BANCO MERCEDES-BENZ DO BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO MIZUHO DO BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO MODAL S.A.	Banco Múltiplo	Sucesso
BANCO MONEO S.A.	Banco Múltiplo	Sucesso
BANCO MORADA S.A	Banco Múltiplo	Insucesso
BANCO MORGAN STANLEY S.A.	Banco Múltiplo	Sucesso
BANCO MUFG BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO NACIONAL S.A.	Banco Múltiplo	Insucesso
BANCO NEON S.A.	Banco Comercial	Insucesso
BANCO OLÉ BONSUCESSO CONSIGNADO S.A.	Banco Múltiplo	Sucesso
BANCO OPEN S.A.	Banco Múltiplo	Insucesso
BANCO ORIGINAL DO AGRONEGÓCIO S.A.	Banco Múltiplo	Sucesso
BANCO ORIGINAL S.A.	Banco Múltiplo	Sucesso
BANCO OURINVEST S.A.	Banco Múltiplo	Sucesso
BANCO PACCAR S.A.	Banco Múltiplo	Sucesso
BANCO PAN S.A.	Banco Múltiplo	Sucesso
BANCO PAULISTA S.A.	Banco Múltiplo	Sucesso
BANCO PINE S.A.	Banco Múltiplo	Sucesso
BANCO PONTUAL S.A.	Banco Múltiplo	Insucesso
BANCO PORTO SEGURO S.A.	Banco Múltiplo	Insucesso
BANCO PROSPER S.A.	Banco Múltiplo	Insucesso
BANCO PSA FINANCE BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO RABOBANK INTERNATIONAL BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO RANDON S.A.	Banco Múltiplo	Sucesso
BANCO RCI BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO RENDIMENTO S.A.	Banco Comercial	Sucesso
BANCO RIBEIRAO PRETO S.A.	Banco Múltiplo	Sucesso
BANCO RODOBENS S.A.	Banco Múltiplo	Sucesso
BANCO ROSA S.A.	Banco Múltiplo	Insucesso
BANCO RURAL S.A.	Banco Múltiplo	Insucesso
BANCO SAFRA S.A.	Banco Múltiplo	Sucesso
BANCO SANTANDER (BRASIL) S.A.	Banco Múltiplo	Sucesso
BANCO SANTOS NEVES S.A.	Banco Múltiplo	Insucesso
BANCO SANTOS S.A.	Banco Múltiplo	Insucesso
BANCO SÃO JORGE S.A.	Banco Comercial	Insucesso
BANCO SEMEAR S.A.	Banco Múltiplo	Sucesso
BANCO SIMPLES S.A.	Banco Múltiplo	Insucesso
BANCO SISTEMA S.A.	Banco Múltiplo	Sucesso
BANCO SMARTBANK S.A.	Banco Múltiplo	Sucesso
BANCO SOCIETE GENERALE BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO SOFISA S.A.	Banco Múltiplo	Sucesso
BANCO SUMITOMO MITSUI BRASILEIRO S.A.	Banco Múltiplo	Sucesso
BANCO TOPÁZIO S.A.	Banco Múltiplo	Sucesso

(Continuação)

NOME DA INSTITUIÇÃO	SEGMENTO	CLASSIFICAÇÃO
BANCO TOYOTA DO BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO TRIANGULO S.A.	Banco Múltiplo	Sucesso
BANCO TRICURY S.A.	Banco Múltiplo	Sucesso
BANCO UNIVERSAL S.A.	Banco Comercial	Insucesso
BANCO VEGA S.A.	Banco Múltiplo	Insucesso
BANCO VETOR S.A.	Banco Múltiplo	Insucesso
BANCO VIPAL S.A.	Banco Múltiplo	Sucesso
BANCO VOLKSWAGEN S.A.	Banco Múltiplo	Sucesso
BANCO VOLVO BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO VOTORANTIM S.A.	Banco Múltiplo	Sucesso
BANCO VR S.A.	Banco Múltiplo	Sucesso
BANCO WESTERN UNION DO BRASIL S.A.	Banco Comercial	Sucesso
BANCO WOORI BANK DO BRASIL S.A.	Banco Múltiplo	Sucesso
BANCO YAMAHA MOTOR DO BRASIL S.A.	Banco Múltiplo	Sucesso
BANCOSEGURO S.A.	Banco Múltiplo	Sucesso
BANESTES S.A. BANCO DO ESTADO DO ESPIRITO SANTO	Banco Múltiplo	Sucesso
BANFORT - BANCO FORTALEZA S.A.	Banco Múltiplo	Insucesso
BANK OF AMERICA MERRILL LYNCH BANCO MÚLTIPLO S.A.	Banco Múltiplo	Sucesso
BCV - BANCO DE CRÉDITO E VAREJO S.A.	Banco Múltiplo	Sucesso
BFC BANCO S.A.	Banco Múltiplo	Insucesso
BFI - BANCO DE FINANCIAMENTO INTERNACIONAL S.A.	Banco Múltiplo	Insucesso
BIG S.A. - BANCO IRMAOS GUIMARAES	Banco Múltiplo	Insucesso
BNY MELLON BANCO S.A.	Banco Comercial	Sucesso
BRB - BANCO DE BRASILIA S.A.	Banco Múltiplo	Sucesso
CHINA CONSTRUCTION BANK (BRASIL) BANCO MÚLTIPLO S/A	Banco Múltiplo	Sucesso
COMMERZBANK BRASIL S.A. - BANCO MÚLTIPLO	Banco Múltiplo	Sucesso
DAYCOVAL LEASING - BANCO MÚLTIPLO S.A.	Banco Múltiplo	Sucesso
DEUTSCHE BANK S.A. - BANCO ALEMAO	Banco Múltiplo	Sucesso
GOLDMAN SACHS DO BRASIL BANCO MULTIPLO S.A.	Banco Múltiplo	Sucesso
HIPERCARD BANCO MÚLTIPLO S.A.	Banco Múltiplo	Sucesso
ICBC DO BRASIL BANCO MÚLTIPLO S.A.	Banco Múltiplo	Sucesso
INTESA SANPAOLO BRASIL S.A. - BANCO MÚLTIPLO	Banco Múltiplo	Sucesso
ITAÚ UNIBANCO HOLDING S.A.	Banco Múltiplo	Sucesso
ITAÚ UNIBANCO S.A.	Banco Múltiplo	Sucesso
KIRTON BANK S.A. - BANCO MÚLTIPLO	Banco Múltiplo	Sucesso
MILBANCO S.A.	Banco Múltiplo	Insucesso
NOVO BANCO CONTINENTAL S.A. - BANCO MÚLTIPLO	Banco Múltiplo	Sucesso
OMNI BANCO S.A.	Banco Múltiplo	Sucesso
PARAIBAN - BANCO DO ESTADO DA PARAIBA S.A.	Banco Comercial	Insucesso
PARANÁ BANCO S.A.	Banco Múltiplo	Sucesso
PLURAL S.A. BANCO MÚLTIPLO	Banco Múltiplo	Sucesso
SCANIA BANCO S.A.	Banco Múltiplo	Sucesso
SCOTIABANK BRASIL S.A. BANCO MÚLTIPLO	Banco Múltiplo	Sucesso
STATE STREET BRASIL S.A. – BANCO COMERCIAL	Banco Comercial	Sucesso

Fonte: elaborado pelo autor com base em Banco Central do Brasil (2019).

APÊNDICE B – DESCRIÇÃO DO PROGRAMA UTLIZADO NA CONSTRUÇÃO DA ÁVORE DE DECISÃO

```
# carregar os dados e dar nome à variável dependente

solvencia <- read.csv(file.choose())

str(solvencia)


# create random training and test datasets

set.seed(123)

train_sample <- sample(193, 154)

str(train_sample)

set.seed(12345)

solvencia_rand <- solvencia[order(runif(193)), ]

solvencia_train <- solvencia_rand[1:154, ]

solvencia_test <- solvencia_rand[155:193, ]

prop.table(table(solvencia_train$sucesso))

prop.table(table(solvencia_test$sucesso))


# build the classifier

require(C50)

solvencia_train$sucesso <- factor(solvencia_train$sucesso)

solvencia_model <- C5.0(solvencia_train[-1], solvencia_train$sucesso)

solvencia_model

summary(solvencia_model)
```

```
# evaluate model performance
solvencia_pred <- predict(solvencia_model, solvencia_test)
require(gmodels)
CrossTable(solvencia_test$sucesso, solvencia_pred,
           prop.chisq=FALSE, prop.c=FALSE, prop.r=FALSE,
           dnn=c('actual sucesso', 'predicted sucesso'))


## using boosting through trials parameter
solvencia_boost10 <- C5.0(solvencia_train[-1], solvencia_train$sucesso,
                          trials = 10)


#summary(solvencia_boost10)
solvencia_boost_pred10 <- predict(solvencia_boost10, solvencia_test)
CrossTable(solvencia_test$sucesso, solvencia_boost_pred10,
           prop.chisq=FALSE, prop.c=FALSE, prop.r=FALSE,
           dnn=c('actual sucesso', 'predicted sucesso'))


#apresentar árvore graficamente
plot(solvencia_model)
```

APÊNDICE C – DESCRIÇÃO DO PROGRAMA UTLIZADO NO PROCESSO DE WINSORIZAÇÃO DOS DADOS

```
#Install and Load Packages

list.of.packages <- c("DescTools", "AER", "sandwich", "lmtest", "car", "dplyr", "stargazer", "ggplot2", "foreign",

                      "openintro","OIdata", "gdata", "doBy","ivpack", "psych","plm", "readxl")

new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"])]

if(length(new.packages)) install.packages(new.packages)

lapply(list.of.packages, require, character.only = TRUE)


##Load database (apenas as empresas de “sucesso”)

database <- read_excel(file.choose())


#Winsorize (destinado apenas às empresas de “sucesso”)

#obs. 1: percentis definidos de acordo com os limites do boxplot

#obs. 2: o programa deve ser reaplicado para as empresas de “insucesso”

attach(database)

database$alavancagem_NW<-Winsorize(database$alavancagem,na.rm=T, probs = c(0.00, 0.99))

database$capital_giro_NW<-Winsorize(database$capital_giro,na.rm=T, probs = c(0.01, 1.00))

database$custo_administrativo_NW<-Winsorize(database$custo_administrativo,na.rm=T, probs = c(0.01, 1.00))

database$custo_medio_captacao_NW<-Winsorize(database$custo_medio_captacao,na.rm=T, probs = c(0.01, 1.00))

database$custo_pessoal_NW<-Winsorize(database$custo_pessoal,na.rm=T, probs = c(0.01, 1.00))

database$custo_total_NW<-Winsorize(database$custo_total,na.rm=T, probs = c(0.01, 1.00))

database$eficiencia_operacional_NW<-Winsorize(database$eficiencia_operacional,na.rm=T, probs = c(0.01, 0.99))
```

```
database$encaixe_NW<-Winsorize(database$encaixe,na.rm=T, probs = c(0.00, 0.99))

database$imobilizacao_capital_proprio_NW<-Winsorize(database$imobilizacao_capital_proprio,na.rm=T, probs = c(0.00, 0.99))

database$imobilizacao_propria_NW<-Winsorize(database$imobilizacao_propria,na.rm=T, probs = c(0.00, 0.99))

database$independencia_financeira_NW<-Winsorize(database$independencia_financeira,na.rm=T, probs = c(0.00, 0.99))

database$indice_emprestimos_sobre_depositos_NW<-Winsorize(database$indice_emprestimos_sobre_depositos,na.rm=T, probs = c(0.00, 0.99))

database$juros_passivos_NW<-Winsorize(database$juros_passivos,na.rm=T, probs = c(0.01, 1.00))

database$liquidez_imediata_NW<-Winsorize(database$liquidez_imediata,na.rm=T, probs = c(0.00, 0.99))

database$lucratividade_ativos_NW<-Winsorize(database$lucratividade_ativos,na.rm=T, probs = c(0.00, 0.99))

database$margem_bruta_NW<-Winsorize(database$margem_bruta,na.rm=T, probs = c(0.01, 0.99))

database$margem_financeira_NW<-Winsorize(database$margem_financeira,na.rm=T, probs = c(0.01, 0.99))

database$margem_liquida_NW<-Winsorize(database$margem_liquida,na.rm=T, probs = c(0.01, 0.99))

database$margem_operacional_NW<-Winsorize(database$margem_operacional,na.rm=T, probs = c(0.01, 0.99))

database$participacao_emprestimos_NW<-Winsorize(database$participacao_emprestimos,na.rm=T, probs = c(0.00, 1.00))

database$relacao_capital_sobre_depositantes_NW<-Winsorize(database$relacao_capital_sobre_depositantes,na.rm=T, probs = c(0.00, 0.99))

database$rentabilidade_pl_NW<-Winsorize(database$rentabilidade_pl,na.rm=T, probs = c(0.01, 0.99))

database$retorno_medio_operacoes_credito_NW<-Winsorize(database$retorno_medio_operacoes_credito,na.rm=T, probs = c(0.00, 0.99))

database$retorno_sobre_investimento_total_NW<-Winsorize(database$retorno_sobre_investimento_total,na.rm=T, probs = c(0.01, 0.99))


database$capital_giro<-Winsorize(database$capital_giro_NW,na.rm=T, probs = c(0.00, 1.00))

database$alavancagem<-Winsorize(database$alavancagem_NW,na.rm=T, probs = c(0.00, 0.99))
```

```
database$capital_giro<-Winsorize(database$capital_giro_NW,na.rm=T, probs = c(0.01, 1.00))

database$custo_administrativo<-Winsorize(database$custo_administrativo_NW,na.rm=T, probs = c(0.01, 1.00))

database$custo_medio_captacao<-Winsorize(database$custo_medio_captacao_NW,na.rm=T, probs = c(0.01, 1.00))

database$custo_pessoal<-Winsorize(database$custo_pessoal_NW,na.rm=T, probs = c(0.01, 1.00))

database$custo_total<-Winsorize(database$custo_total_NW,na.rm=T, probs = c(0.01, 1.00))

database$eficiencia_operacional<-Winsorize(database$eficiencia_operacional_NW,na.rm=T, probs = c(0.01, 0.99))

database$encaixe<-Winsorize(database$encaixe_NW,na.rm=T, probs = c(0.00, 0.99))

database$imobilizacao_capital_proprio<-Winsorize(database$imobilizacao_capital_proprio_NW,na.rm=T, probs = c(0.00, 0.99))

database$imobilizacao_propria<-Winsorize(database$imobilizacao_propria_NW,na.rm=T, probs = c(0.00, 0.99))

database$independencia_financeira<-Winsorize(database$independencia_financeira_NW,na.rm=T, probs = c(0.00, 0.99))

database$indice_emprestimos_sobre_depositos<-
Winsorize(database$indice_emprestimos_sobre_depositos_NW,na.rm=T, probs = c(0.00, 0.99))

database$juros_passivos<-Winsorize(database$juros_passivos_NW,na.rm=T, probs = c(0.01, 1.00))

database$liquidez_imediata<-Winsorize(database$liquidez_imediata_NW,na.rm=T, probs = c(0.00, 0.99))

database$lucratividade_ativos<-Winsorize(database$lucratividade_ativos_NW,na.rm=T, probs = c(0.00, 0.99))

database$margem_bruta<-Winsorize(database$margem_bruta_NW,na.rm=T, probs = c(0.01, 0.99))

database$margem_financeira<-Winsorize(database$margem_financeira_NW,na.rm=T, probs = c(0.01, 0.99))

database$margem_liquida<-Winsorize(database$margem_liquida_NW,na.rm=T, probs = c(0.01, 0.99))

database$margem_operacional<-Winsorize(database$margem_operacional_NW,na.rm=T, probs = c(0.01, 0.99))

database$participacao_emprestimos<-Winsorize(database$participacao_emprestimos_NW,na.rm=T, probs = c(0.00, 1.00))

database$relacao_capital_sobre_depositantes<-
Winsorize(database$relacao_capital_sobre_depositantes_NW,na.rm=T, probs = c(0.00, 0.99))
```

```
database$rentabilidade_pl<-Winsorize(database$rentabilidade_pl_NW,na.rm=T, probs = c(0.01, 0.99))

database$retorno_medio_operacoes_credito<-
Winsorize(database$retorno_medio_operacoes_credito_NW,na.rm=T, probs = c(0.00, 0.99))

database$retorno_sobre_investimento_total<-
Winsorize(database$retorno_sobre_investimento_total_NW,na.rm=T, probs = c(0.01, 0.99))



detach(database)



#exportar a tabela nova em excel

library(openxlsx)

write.xlsx(database, "C:/....xlsx")
```