UNIVERSIDADE FERDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO
CURSO DE MESTRADO ACADÊMICO EM ADMINISTRAÇÃO

PEDRO IGOR DE SOUSA DAMASCENO

RISCO DE INSOLVÊNCIA E SENTIMENTO TEXTUAL BANCÁRIO: UMA ANÁLISE DOS BANCOS DE CAPITAL ABERTO NO BRASIL

João Pessoa
2021

PEDRO IGOR DE SOUSA DAMASCENO

RISCO DE INSOLVÊNCIA E SENTIMENTO TEXTUAL BANCÁRIO: UMA ANÁLISE DOS BANCOS DE CAPITAL ABERTO NO BRASIL

Dissertação apresentada como requisito parcial para obtenção do título de mestre em Administração no Programa de Pós-Graduação em Administração da Universidade Federal da Paraíba - UFPB.

Área de concentração: Administração e Sociedade

Linha de Pesquisa: Finanças e Métodos Quantitativos

Ênfase: Finanças e Métodos Quantitativos

Orientador**:** Prof. Dr. Cássio da Nóbrega Besarria

João Pessoa

2021

Catalogação na publicação
Seção de Catalogação e Classificação

D155r Damasceno, Pedro Igor de Sousa.
Risco de insolvência e sentimento textual bancário: uma análise dos bancos de capital aberto no Brasil / Pedro Igor de Sousa Damasceno. - João Pessoa, 2021.
54 f. : il.

Orientação: Cássio da Nóbrega Besarria.
Dissertação (Mestrado) - UFPB/CCSA.

1. Risco de insolvência. 2. Bancos de capital aberto. 3. Sentimento textual. 4. Aprendizagem de máquina. I. Besarria, Cássio da Nóbrega. II. Título.

UFPB/BC CDU 336.7(043)

Elaborado por Gracilene Barbosa Figueiredo - CRB-15/794

PEDRO IGOR DE SOUSA DAMASCENO

RISCO DE INSOLVÊNCIA E SENTIMENTO TEXTUAL BANCÁRIO: UMA ANÁLISE DOS BANCOS DE CAPITAL ABERTO NO BRASIL

Dissertação apresentada ao Programa de Pós-Graduação em Administração do Centro de Ciências Sociais Aplicadas da Universidade Federal da Paraíba, Linha de pesquisa: Finanças e Métodos Quantitativos, em cumprimento ao requisito parcial para obtenção do Título de mestre em Administração.

Área de Concentração: Administração e Sociedade.

Aprovada em: 24/02/2021.

BANCA EXAMINADORA

Cássio da N. Besarria

Prof.(a) Dr.(a) Cássio da Nóbrega Besarria
Orientador - PPGA/UFPB

Orleans Silva Martins

Prof.(a) Dr.(a) Orleans Silva Martins
Titular Interno – PPGA/UFPB

Prof.(a) Dr.(a) José Alves Dantas
Titular Externo – PPGCont/UNB

Pedro Igor de Sousa Damasceno

Pedro Igor de Sousa Damasceno
Mestrando (a)

AGRADECIMENTOS

Agradeço a Deus, por ter me permitido seguir esse caminho, guiando-me e abençoando ao longo dessa jornada.

À minha família, por todo o apoio incondicional que recebi e pelas constantes demonstrações de amizade.

Ao meu orientador, professor Dr. Cássio da Nóbrega Besarria, pela maneira como conduziu a orientação, agradeço profundamente pelos incentivos, pelos ensinamentos acadêmicos e de vida, pela paciência e por ter confiado na minha capacidade para desenvolver a pesquisa.

À professora Dra. Maria Daniella Oliveira Pereira da Silva, pelos ensinamentos acadêmicos, por todas as contribuições prestadas para a minha formação profissional e para a execução desta pesquisa.

Ao Programa de Pós-Graduação em Administração da Universidade Federal da Paraíba (PPGA-UFPB) e aos professores do programa, pela formação acadêmica de excelência que me foi propiciada, contribuindo para o meu crescimento profissional e pessoal.

Aos professores Dr. Orleans Silva Martins e Dr. José Alves Dantas, por terem aceitado fazer parte da minha banca, pelo tempo dedicado e pelas contribuições para esta pesquisa.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo financiamento desta pesquisa.

Aos meus amigos da turma 44, por todos os momentos que tivemos e pelo aprendizado que tive com vocês ao longo desses dois anos, em especial agradeço à Arlan Macêdo e Camila Fernandes, pelas discussões enriquecedoras.

RESUMO

Este estudo teve como objetivo analisar se o sentimento textual explica o maior risco de insolvência dos bancos de capital aberto no Brasil, com uma amostra composta por 17 empresas e 450 observações referentes ao período do quarto trimestre de 2012, até o quarto trimestre de 2019. A estratégia empírica adotada é dividida em três partes: a primeira consiste na utilização do algoritmo não supervisionado *k-means* para classificar os bancos de acordo com o seu risco de insolvência, sendo elaborada uma nova medida de probabilidade de falência durante esse processo. Nessa etapa foi observado que 66 observações haviam sido classificadas como de alto risco e 384 de baixo risco, sendo assim uma métrica mais rigorosa que o *Z-score*, quanto a classificação de bancos com alto risco de insolvência. Em seguida, foram empregados os métodos supervisionados de aprendizagem de máquina *naive bayes* e *random forest*, e o modelo *logit*, para identificar qual dessas técnicas estatísticas é mais robusta para a predição da variável construída na etapa anterior. A partir da matriz de confusão e do critério de acurácia foi possível identificar que o modelo logístico apresentou o maior poder preditivo. Por fim, foi realizada a terceira etapa para avaliar se o sentimento textual, a variação percentual real do Produto Interno Bruto (PIB), a capitalização, lucratividade, liquidez e o tamanho dessas firmas explicam o risco de uma insolvência bancária. Os resultados demonstram que bancos com maior probabilidade de falência apresentam um sentimento textual mais otimista.

Palavras-chave: Risco de insolvência. Bancos de capital aberto. Sentimento textual. Aprendizagem de máquina.

ABSTRACT

This study aimed to analyze whether the textual sentiment explains the greater risk of insolvency of publicly held banks in Brazil, with a sample composed of 17 companies and 450 observations referring to the period from the fourth quarter of 2012, until the fourth quarter of 2019. The empirical strategy adopted is divided in three parts: the first consists of using the unsupervised algorithm k-means to classify banks according to their risk of insolvency, a new measure of bankruptcy probability was elaborated during this process. At this stage, it was observed that 66 observations were classified as high risk, and 384 as low risk, thus being a more rigorous metric than the Z-score, regarding the classification of banks with high risk of insolvency. Then, supervised machine learning methods naive bayes and random forest and the logit model were used to identify which of these statistical techniques is more robust for the prediction of the variable constructed in the previous step. From the confusion matrix and the accuracy criterion it was possible to identify that the logistic model presented the greatest predictive power. Finally, a third step was taken to assess whether textual sentiment, the real percentage change in Gross Domestic Product (GDP), capitalization, profitability, liquidity, and the size of these firms explain the risk of bank insolvency. The results show that banks with a higher probability of bankruptcy have a more optimistic textual feeling.

Keywords: Insolvency risk. Publicly traded banks. Textual feeling. Machine learning.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O colapso de uma companhia bancária resulta, sobretudo, da má gestão dos seus executivos e afeta tanto os detentores de depósitos e funcionários daquela corporação, como toda a economia de um país, cuja necessidade de socorrer bancos em situações de falência pode ocasionar em custos onerosos que impactam ainda mais o desenvolvimento econômico daquela nação (ARAÚJO, 2019; DEL GAUDIO *et al*., 2019; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

Neste contexto, comportamentos de insolvência bancária passaram a ganhar cada vez mais destaque na literatura considerando a elevada repercussão financeira adversa que é gerada, uma vez que tanto os investidores como os donos de depósitos tendem a perder a confiança nessas instituições inadimplentes, o que pode contaminar os demais bancos presentes no mercado e no longo prazo resultar em uma crise bancária, essa última denota em consequências ainda mais severas que vão desde a paralisação da oferta de crédito para empresas e famílias até a fuga de capitais daquele país (BARBOSA, 2017).

Nesse sentido, tecnicamente a insolvência de um banco ocorre quando as perdas incorridas por essa instituição não podem ser cobertas pelos seus recursos próprios (BOYD; GRAHAM, 1986; LEPETIT; STROBEL, 2013). Diante dessa possibilidade a literatura tem desenvolvido medidas de risco de insolvência, entre essas se destacam o sistema *CAMELS* e o *Z-score*, em que o último indica a distância em que o banco se encontra de um comportamento de insolvência (LEPETIT; STROBEL, 2013; VIEIRA; GIRÃO, 2016; SUSS; TREITEL, 2019; VIEIRA *et al.*, 2020; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

Embora pesquisas anteriores tenham se destinado à previsão de falências bancárias (BARBOSA 2017; ROSA; GARTNER, 2018; CLIMENT; MOMPARLER; CARMONA, 2019), o presente estudo está associado à classificação dos bancos de acordo com o seu risco de insolvência, sendo consideradas para tanto as companhias bancárias de capital aberto no Brasil. Dessa maneira, são empregadas as variáveis preditoras indicadas pela literatura, com o intuito de elaborar uma métrica de risco de insolvência capaz de categorizar essas instituições.

Nesse sentido, este estudo se diferencia dos demais por, em um primeiro momento, aplicar a técnica de aprendizagem de máquina não supervisionada, utilizando o algoritmo *k-means* para reconhecer os padrões na disposição dos dados. Essa análise será empregada para realizar à classificação das empresas bancárias, em agrupamentos associados ao maior ou menor risco de falência (BARBOSA 2017; FERNANDES; CHIAVEGATTO FILHO, 2019).

O resultado da categorização obtida pelo *k-means*, foi comparado com a classificação de risco de insolvência dos bancos indicada pelo *Z-score*. Ao verificar o resultado pelos diferentes métodos, foi possível identificar que o algoritmo agrupou de maneira condizente ao *Z-score*, 50,82%, e 98,78% das observações, respectivamente, para os grupos de maior e menor probabilidade de insolvência.

Cabe ressaltar que a vantagem da utilização do método *k-means*, é que conforme descrito por Viswanathan, Srinivasan e Hariharan (2020), a aplicação desse algoritmo de classificação permite assegurar a heterogeneidade entre os agrupamentos, além de garantir a homogeneidade dentro dos *clusters* (grupos) estabelecidos pela máquina.

A partir da medida de risco de insolvência bancária obtida por meio do *k-means*, foram aplicados em um segundo momento os algoritmos supervisionados de aprendizagem de máquina *naive bayes* e *random forest*, e o modelo clássico de regressão logística *logit*, com o objetivo de identificar qual dos modelos de classificação é o mais consistente para à predição da variável construída.

A escolha dos métodos de aprendizagem de máquina e do modelo logístico, é ampara por pesquisas que tratam sobre a previsão de falências em bancos, pois essa abordagem conceitual é destinada a comparação de modelos econométricos, com o intuito de identificar aqueles que demonstram os melhores resultados preditivos (LIBERMAN; BARBOSA; PIRES, 2018; CLIMENT; MOMPARLER; CARMONA, 2019; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

Os resultados dessa etapa indicam que o modelo *naive bayes* apresentou uma acurácia de previsão de 81,5%, já o modelo *logit* obteve uma acurácia de 84,4%. Quanto ao *random forest*, esse modelo demonstrou problemas de *over-fitting*, fazendo com que os resultados de acurácia desse algoritmo supervisionado fossem desconsiderados. Assim, o modelo logístico foi estabelecido como o mais adequado para a previsão dessa nova métrica de risco de insolvência, elaborada por meio do *k-means*.

Para a definição das variáveis empregadas nos modelos econométricos, foram utilizadas as mesmas medidas reportadas pela literatura bancária. Chiaramonte *et al.* (2016), Vieira e Girão (2016), Rosa e Gartner (2018), Ferreira, Zanini e Alves (2019), e Andrade (2020) defendem a utilização de informações contábeis e dados macroeconômicos para explicarem comportamentos relacionados ao risco de insolvência em bancos. Adicionalmente, uma segunda abordagem destacada por Gupta, Simaan e Zaki (2018), Del Gaudio *et al.* (2019) e Gandhi, Loughran e McDonald (2019) considera que, além do uso de dados de caráter

quantitativo, também devem ser aplicadas informações qualitativas, uma vez que essas tendem a refletir o ambiente de atuação das empresas.

A literatura que trata sobre o risco de insolvência em bancos, destaca o impacto que o tamanho da instituição pode apresentar sobre esse comportamento, pois, firmas maiores podem aproveitar o fato de serem companhias *too-big-to-fail*, e adotarem um modelo de negócios mais arriscado (VIEIRA; GIRÃO, 2016; FERREIRA; ZANINI; ALVES, 2019; ANDRADE, 2020). Em contrapartida, bancos de grande porte tendem a reduzir a sua probabilidade de falência, por meio da elevada quantidade de serviços disponibilizados aos seus clientes, fazendo com que a empresa atue de maneira diversificada e apresente ganhos menos voláteis (DEL GAUDIO *et al.*, 2019; VIEIRA *et al.*, 2020).

A lucratividade, por sua vez, é associada de forma inversa ao risco de falência, assim bancos com lucros menores tendem a ser insolventes e consequentemente apresentam ganhos mais voláteis (VIEIRA; GIRÃO, 2016; LIBERMAN; BARBOSA; PIRES, 2018; ROSA; GARTNER, 2018; VIEIRA *et al.*, 2020). A capitalização também é evidenciada pois, bancos com maiores riscos de insolvência apresentam níveis inferiores de capitalização (VIEIRA; GIRÃO, 2016; FERREIRA; ZANINI; ALVES, 2019; VIEIRA *et al.*, 2020).

Nesse sentido, firmas bancárias com um menor nível liquidez são associadas a um incremento na sua probabilidade de falência (VIEIRA; GIRÃO, 2016; BARBOSA 2017; FERREIRA; ZANINI; ALVES, 2019). Por fim, o Produto Interno Bruto (PIB) também é destacado na literatura, podendo reduzir o risco de falência bancária em momentos de crescimento econômico, ou aumentando a probabilidade de insolvência em situações de crises financeiras, nos países (CHIARAMONTE *et al.*, 2016; VIEIRA; GIRÃO, 2016; ROSA; GARTNER, 2018; FERREIRA; ZANINI; ALVES, 2019).

Dessa forma, estudos que tratam sobre o risco da insolvência bancária têm objetivado identificar novas variáveis capazes de explicar esses comportamentos. Nesse sentido Gupta, Simaan e Zaki (2018) indicam que o sentimento textual é capaz de predizer a falência dessas companhias, considerando ainda que o tom textual otimista tem um maior poder preditivo para essas firmas, identificando que os bancos insolventes apresentam sentimentos mais positivos do que os seus pares não falidos.

Corroborando essa perspectiva Gandhi, Loughran e McDonald (2019) também identificaram que o sentimento textual explica a saída dos bancos das bolsas de valores norte-americanas. Contudo, os autores consideram que o sentimento pessimista apresenta maior capacidade de previsão, ao identificarem uma relação entre o aumento desse tipo de tom textual e fechamentos de capital de companhias bancárias. Adicionalmente, Del Gaudio *et al.* (2019)

endossam que o sentimento textual consegue explicar o risco de insolvência, indicando uma relação entre o tom pessimista e o aumento na probabilidade de falência dos bancos europeus.

Considerando essa abordagem conceitual o presente estudo elaborou uma variável de sentimento textual para os bancos de capital aberto listados na bolsa brasileira, fazendo uso dos relatórios trimestrais (ITR) e das Demonstrações Financeiras Padronizadas (DFP) dessas empresas, disponibilizadas no endereço eletrônico da Comissão de Valores Mobiliários (CVM).

A escolha desses relatórios financeiros para a estimação do sentimento textual, considerou a proposta indicada por Gupta, Simaan e Zaki (2018), e Jiang *et al*. (2019) ao destacarem que o uso de mais de um tipo de documento é indicado para essa prática ao permitir a criação de um indicador robusto de sentimento. Além disso, relatórios financeiros descrevem a conjuntura dos bancos em relação aos seus riscos, sua lucratividade e solvência, bem como apresentam perspectivas sobre o comportamento futuro do negócio.

Em seguida, foi aplicada a técnica denominada de *bag of words* para mensurar o sentimento textual presente nos relatórios destacados, empregando o dicionário e o algoritmo de leitura desenvolvidos por Silva e Machado (2019), o qual permite reconhecer os termos presentes nos documentos que estão associados ao sentimento textual positivo ou negativo e mediante contagem de frequência dessas palavras, determinar o tom pessimista ou otimista daquele relatório analisado (LOUGHRAN; MCDONALD, 2011; KEARNEY; LIU, 2014).

Dessa forma, considerando os estudos anteriores que empregaram o sentimento textual e que fizeram uso da metodologia de *bag of words* (GUPTA; SIMAAN; ZAKI, 2018; DEL GAUDIO *et al.*, 2019; GANDHI; LOUGHRAN; MCDONALD, 2019), esta pesquisa objetiva responder à seguinte problemática: **O sentimento textual pode explicar um comportamento de maior risco de insolvência por parte dos bancos listados no mercado acionário brasileiro?**

1.1 Objetivos da pesquisa

1.1.1 Objetivo geral

Analisar se o sentimento textual explica o maior risco de insolvência dos bancos de capital aberto no Brasil, no período do quarto trimestre de 2012, até o quarto trimestre de 2019.

1.2.2 Objetivos específicos

- Elaborar uma medida de risco de insolvência para os bancos de capital aberto no Brasil, utilizando o algoritmo *k-means* (VISWANATHAN; SRINIVASAN; HARIHARAN, 2020);
- Identificar entre os modelos utilizados no estudo qual apresenta a maior capacidade preditiva, para a métrica de risco de falência desenvolvida;
- Construir uma variável de sentimento textual para os bancos presentes no mercado de capitais brasileiro (SILVA; MACHADO, 2019).

1.2 Justificativa

A importância de estudar o risco de insolvência dos bancos é relacionada a diversos argumentos na literatura, pois esse setor é observado de forma peculiar pelas entidades reguladoras devido a sua tarefa de manter a estabilidade do sistema financeiro, cujos bancos acabam por se concretizar como a base desses sistemas, sobretudo em países emergentes (FERREIRA; ZANINI; ALVES, 2019; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

A necessidade desse acompanhamento especial por parte dos bancos centrais reside no fato de que episódios resultantes em falências bancárias podem ter inúmeras consequências, sejam econômicas ou sociais, e devido às relações estabelecidas entre os bancos comerciais, essas empresas podem ocasionar crises bancárias que denotam em diversos problemas para as nações (BARBOSA, 2017; LIBERMAN; BARBOSA; PIRES, 2018).

As crises bancárias ocorrem quando um banco em situação de vulnerabilidade passa a afetar uma outra companhia bancária, que no longo prazo podem contaminar um grupo de empresas desse setor e em larga escala ocasionar em uma crise financeira. Essa última, por sua vez, desafia a estabilidade econômica de um país, resultando em uma desaceleração da economia e em uma redução na capacidade de liquidez das corporações, o que gera desempregos e aumenta o comportamento de inadimplência da população (BARBOSA, 2017).

Além disso, a falência de um banco também resulta em problemas sociais, uma vez que pessoas físicas e jurídicas confiaram os seus depósitos a essas instituições (ALVES, 2009; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

Nessa perspectiva, informações qualitativas como o sentimento textual podem aprimorar o monitoramento dos bancos por parte dos agentes supervisores, tornando-se um

elemento de fundamental importância devido à complexidade associada a essa tarefa de fiscalização, além de possibilitar uma atuação preventiva desses órgãos, reduzindo o custo de socorro a essas companhias financeiras e permitindo uma intervenção em tempo hábil para casos de falência iminente (GANDHI; LOUGHRAN; MCDONALD, 2019; SUSS; TREITEL, 2019).

De maneira complementar, Del Gaudio *et al.* (2019) e Gandhi, Loughran e McDonald (2019) endossam que ao longo das últimas décadas os reguladores aumentaram os requisitos de divulgação de informações para os bancos em todo o mundo, principalmente no que se refere ao risco das suas operações; no entanto, segundo os autores essas companhias financeiras permanecem sendo instituições inerentemente com pouca transparência e elevada assimetria.

Em meio a esse contexto, a motivação e relevância desta pesquisa consiste no fato de que o sentimento textual além de auxiliar as entidades reguladoras, também pode ajudar os investidores e os donos de depósitos na alocação dos seus recursos. De forma específica, para a academia a presente pesquisa se soma ao corpo de estudos que tratam sobre o risco de insolvência dos bancos, contribuindo com uma nova medida para mensurar esses comportamentos.

Adicionalmente, os resultados deste trabalho também auxiliam na identificação de modelos preditivos associados à probabilidade de insolvência bancária, além de fornecer evidências empíricas de que informações qualitativas podem explicar o risco bancário associado a uma possível falência, até então não observado em estudos anteriores no Brasil, sendo assim explanadas as contribuições para a academia, sociedade e para a prática das entidades reguladoras do setor bancário.

2 REVISÃO DA LITERATURA

Este capítulo objetiva apresentar uma revisão da literatura que trata sobre os principais conceitos aplicados neste trabalho, abordando as medidas contábeis de risco de falência bancária e sua relação com os modelos de aprendizagem de máquina, empregados para a classificação dos bancos, de acordo com a sua probabilidade de insolvência. Posteriormente, é destacado o impacto que informações qualitativas apresentam na explicação de comportamentos de maior risco por parte das instituições bancárias, bem como um breve resumo dos estudos que trataram sobre essa temática.

2.1 Previsão de falências

Estudos que tratam sobre a predição de falências em bancos são divididos em duas linhas de pesquisa: a primeira é destinada à comparação de diferentes modelos econométricos com o intuito de identificar aqueles que demonstram maior robustez para a previsão de insolvências bancárias; a segunda concepção conceitual objetiva descobrir novas variáveis que consigam explicar esses comportamentos de interesse (BARBOSA, 2017; LIBERMAN; BARBOSA; PIRES, 2018).

Considerando essa primeira linha de pesquisa, trabalhos voltados para esta abordagem fazem uso de modelos estatísticos clássicos como a análise discriminante e principalmente os modelos logísticos; também são empregados métodos de aprendizagem de máquina supervisionados, os quais aplicam inteligência artificial (LIBERMAN; BARBOSA; PIRES, 2018; ROSA; GARTNER, 2018; SUSS; TREITEL, 2019).

A utilização de técnicas de aprendizagem de máquina nessa literatura se ampara no fato de que esses modelos possuem soluções de elevada eficiência na estimação de relações complexas entre as variáveis (BARBOSA, 2017). A literatura que trata de modelos preditivos para falência dos bancos tem empregado diferentes métodos supervisionados, como o *support vector machine* (*SVM*), *random forest*, *boosting*, o classificador de k vizinhos mais próximos (*KNN*), entre outros (BARBOSA, 2017; GUPTA; SIMAAN; ZAKI, 2018; CLIMENT; MOMPARLER; CARMONA, 2019; SUSS; TREITEL, 2019; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

Nesse sentido, o uso de técnicas de aprendizagem de máquina não supervisionada em estudos sobre bancos é bastante limitado. Viswanathan, Srinivasan e Hariharan (2020) ao identificarem inconsistências entre a classificação de *rating* dessas empresas bancárias e à

ocorrência de falência em bancos indianos, utilizaram a técnica de aprendizagem não supervisionada *k-means* para categorizar essas firmas.

Para tanto, foram observados 44 bancos indianos públicos e privados, entre o período de 2005 a 2017, aplicando 9 indicadores contábeis apontados pela literatura bancária para realizar os agrupamentos em *clusters*, resultando em grupos de alta, média e baixa saúde financeira, sendo definidos após a comparação dos agrupamentos com os dados contábeis dessas firmas.

Dessa forma, se sobressaíram principalmente as medidas relacionadas ao crescimento de ativos inadimplentes, adequação de capital e o retorno sobre os ativos (ROA), para a determinação desses grupos de acordo com a situação financeira de cada instituição bancária.

Em seguida, foram utilizadas as técnicas de análise discriminante, *classification of regression tree* (CART) e *random forest*, para identificar qual desses métodos é o mais robusto para a predição da classificação criada pelo *k-means*. O modelo *random forest* apresentou uma acurácia de 95,93%, já a análise discriminante obteve um resultado 95,36% e o modelo CART, por sua vez, com 89,5% de acurácia preditiva. Diante desses resultados, apesar do algoritmo *random forest* ter obtido uma vantagem marginal, os autores consideraram a análise discriminante como o modelo mais parcimonioso, sendo elencado como o método mais consistente aos seus dados.

Contudo, a utilização de algoritmos destinados a problemas de classificação também não se limita à literatura bancária, sendo aplicados em estudos das demais áreas de finanças, em especial se destacando os trabalhos de Silva, Ribeiro e Matias (2016), que analisam a capacidade de previsão de modelos supervisionados em relação a comportamentos de *default* de crédito, e Silva, Rêgo e Frascaroli (2019), que empregaram a análise de *clusters* e o *random forest* para a predição de mudanças na categorização de *ratings* soberanos.

A literatura que trata sobre a predição de falências empresariais foi estabelecida, sobretudo, pelos trabalhos desenvolvidos por Beaver (1966) e Altman (1968), considerando a utilização da análise discriminante para identificar comportamentos capazes de distinguir firmas solventes e falidas, partindo do pressuposto de que companhias insolventes apresentavam medidas contábeis diferentes daquelas que não estavam em situação de vulnerabilidade financeira, cuja corrente conceitual, posteriormente, também passou a ser empregada para o setor bancário (LIBERMAN; BARBOSA; PIRES, 2018; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

Ressalta-se que a medida elaborada por Altman (1968) para identificar companhias que apresentam maior probabilidade de falência não se aplica para as empresas financeiras,

sobretudo ao setor bancário, conforme argumentam Chiaramonte *et al.* (2016). Em meio a esse contexto, é importante destacar que o *Z-score* de Altman (1968) faz uso do método de análise discriminante, em detrimento da medida de risco de insolvência bancária homônima, desenvolvida principalmente a partir das pesquisas de Boyd e Graham (1986), e Boyd, Nicolò e Jalal (2006), que empregam um modelo de cálculo distinto.

Nessa perspectiva, ressalta-se ainda que o *Z-score* elaborado para o setor bancário se caracteriza como um indicador de risco de inadimplência, que considera as informações contábeis das firmas para "apontar" à distância que aquele empreendimento bancário se encontra de uma possível insolvência (BOYD; GRAHAM, 1986; BOYD; NICOLÒ; JALAL, 2006; LEPETIT; STROBEL, 2013; CHIARAMONTE *et al.*, 2016; VIEIRA; GIRÃO, 2016; VIEIRA *et al.*, 2020).

Alguns estudos utilizam o *Z-score* para analisar variáveis capazes de explicar o aumento na probabilidade de falência dos bancos, ou seja, identificando se o modelo de negócios e, por conseguinte, os seus indicadores contábeis, que refletem essas decisões de gestão, explicam o incremento no risco de insolvência bancária, sem que necessariamente a companhia tenha entrado em processo de falência, embora em alguns casos, comportamentos de maior risco de inadimplência são associados a dificuldades financeiras por parte dos bancos (VIEIRA; GIRÃO, 2016; FERREIRA; ZANINI; ALVES, 2019; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

No entanto, o *Z-score* apresenta algumas limitações, uma vez que a sua confiabilidade é diretamente relacionada com a qualidade dos dados contábeis reportadas pelos bancos, pois, caso essas firmas suavizem suas informações financeiras o *Z-score* pode resultar em pontuações de risco que não refletem a realidade da companhia (CHIARAMONTE *et al.*, 2016).

Chiaramonte, Croci e Poli (2015) destacam que entre as vantagens do uso do *Z-score* está a sua aplicação por meio de um cômputo relativamente simples, porém, que apresenta resultados consistentes e semelhantes às covariáveis aplicadas pelo sistema *CAMELS;* este, por sua vez, consiste em um segundo indicador de risco de insolvência bancária e que também utiliza dados contábeis para tanto.

Em detrimento do *Z-score* que resulta em uma pontuação associada à possibilidade de insolvência naquele banco, o *CAMELS* faz uso de um conjunto de variáveis contábeis relacionadas à adequação de capital, qualidade dos ativos, qualidade gerencial, sensibilidade ao risco do mercado, rentabilidade e liquidez; embora tenha sido desenvolvido pelas instituições reguladoras norte-americanas, é utilizado por diversas entidades de supervisão em todo o

mundo, para identificar empresas financeiras mais arriscadas (ROSA; GARTNER, 2018; ARAÚJO, 2019; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

A tarefa de mensurar o risco bancário de insolvência também é associada na literatura a variáveis macroeconômicas; dentre essas ganha destaque o PIB, pois diante de um crescimento econômico ocorre uma tendência de expansão nas atividades produtivas, o que se associa diretamente com a intermediação bancária. Porém, em períodos de redução do PIB existe uma propensão para o efeito contrário, uma vez que a demanda de crédito se reduz e a capacidade das empresas em honrar os seus empréstimos bancários também diminui, deixando os bancos mais expostos a riscos associados à sua inadimplência (CHIARAMONTE *et al.*, 2016; VIEIRA; GIRÃO, 2016; ROSA; GARTNER, 2018; FERREIRA; ZANINI; ALVES, 2019; ANDRADE, 2020).

Nesse sentido, além da utilização de dados de caráter quantitativo, Gupta, Simaan e Zaki (2018), Del Gaudio *et al.* (2019) e Gandhi, Loughran e McDonald (2019) defendem a aplicação de informações qualitativas de âmbito público, em específico aplicando o sentimento textual presente nos relatórios financeiros dessas firmas, para explicar comportamentos de insolvência bancária de maneira conjunta com as demais variáveis quantitativas indicadas pela literatura. Considerando as abordagens destacadas, em seguida, é explanado sobre a relação entre o risco de insolvência dos bancos e o sentimento textual.

2.2 Sentimento textual e insolvência bancária

As pesquisas que aplicam sentimento textual se intensificaram na medida em que descobertas nesse campo evidenciaram que palavras e frases em discursos demonstram características de articulação com o sentimento dos seus autores (LIU; ZHANG, 2012). Assim, a análise do sentimento textual pode ser tida como uma área de pesquisa incipiente e que considera informações de cunho qualitativo para aplicações em modelos quantitativos (DEL GAUDIO *et al*, 2019; SILVA; BESARRIA; SILVA, 2019).

As informações textuais empregadas em tais pesquisas apresentam certas variações de acordo com o problema que se propõem a estudar; no entanto, existe uma maior predominância em três fontes de dados, sendo estes os relatórios financeiros de caráter público, textos em redes sociais na *internet* e artigos elaborados pela imprensa especializada em finanças (KEARNEY; LIU, 2014).

Também é importante ressaltar que os primeiros trabalhos dessa temática foram desenvolvidos objetivando o entendimento da relação entre o sentimento textual e o

comportamento dos mercados de capitais internacionais, o que fez com que esse campo posteriormente se tornasse uma área de pesquisa em ascensão no Brasil (KEARNEY; LIU, 2014; GALDI; GONÇALVES, 2018).

Outro fator de destaque é que ao longo do tempo diversas técnicas de análise de sentimento foram elaboradas, pois as narrativas textuais inicialmente eram examinadas manualmente, sendo limitadas a um pequeno número de observações. Desde então, as pesquisas mais recentes fazem uso de métodos computacionais, dentre estes a técnica de *bag of words*, que permite a análise de uma grande quantidade de documentos por meio de algoritmos automatizados (PAGLIARUSSI; AGUIAR; GALDI, 2016; MACHADO; SILVA, 2017).

De maneira complementar, Liu e Zhang (2012) e Galdi e Gonçalves (2018) endossam que a análise de sentimento por meio de técnicas automatizadas, como a de *bag of words*, também possibilita reduzir o efeito de possíveis vieses pessoais, ao considerarem que existe uma tendência cognitiva, de que informações que estão alinhadas com as opiniões individuais são mais valorizadas na observação de dados textuais, fator esse que poderia produzir resultados menos consistentes quando se faz uso de alguma técnica manual.

Mediante a aplicação da análise textual é possível identificar o tom contido na comunicação verbal, podendo representar um posicionamento voltado a uma postura pessimista ou otimista daquele discurso (SILVA; MACHADO, 2019). Todavia, para a realização de estudos que aplicam a técnica de *bag of words*, é essencial atentar para a aplicação do dicionário (lista de palavras), o qual é empregado pelo algoritmo para a identificação computacional dos termos, associados ao sentimento contido no texto.

Entre os dicionários utilizados para a análise de sentimento textual, inicialmente foi aplicado pela literatura o *Harvard General Inquirer* (*GI/Harvard*) desenvolvido para a área de psicologia, e em um segundo momento empregado em estudos de finanças. Posteriormente, Loughran e McDonald (2011) apresentaram críticas quanto à utilização da lista de palavras estabelecidas pela universidade de *Havard*, identificando inconsistências e até mesmo omissões quando empregadas em textos de caráter financeiro, uma vez que foi estabelecida para identificar o comportamento social.

Em virtude dessas incongruências destacadas, Loughran e McDonald (2011) elaboraram um dicionário próprio, voltado para a observação de documentos textuais financeiros, os quais obtiveram resultados mais consistentes em relação ao *GI/Havard*. No Brasil, Pagliarussi, Aguiar e Galdi (2016), seguindo a mesma metodologia desenvolvida por Loughran e McDonald (2011), construíram uma lista de palavras próprias, fazendo uso dos relatórios das companhias não financeiras listadas no país. Ressalta-se ainda que na literatura foram elaborados outros

dicionários em âmbito nacional, a exemplo das listas de palavras desenvolvidas por Machado e Silva (2017), Silva e Machado (2019), e Silva, Besarria e Silva (2019), os quais, mesmo apresentando certas particularidades quanto à sua construção, consideram o mesmo método estabelecido por Loughran e McDonald (2011).

No que se refere ao uso de relatórios financeiros para a análise textual, Jiang *et al.* (2019) defendem a utilização de mais de um tipo de relatório, com o intuito de obter um indicador robusto de sentimento textual. Em meio a essa perspectiva, Gupta, Simaan e Zaki (2018) endossam que para a estimação do tom textual de companhias bancárias é importante que os relatórios apresentem uma descrição sobre a conjuntura de atuação dos bancos, principalmente, em relação à sua solvência e lucratividade.

Assim, considerando o fato de que os indicadores financeiros e econômicos podem não revelar todas as informações sobre o risco de insolvência bancária, pesquisas recentes passaram a empregar as informações qualitativas, derivadas dos relatórios financeiros dessas firmas, para examinar comportamentos de falência em bancos (GUPTA; SIMAAN; ZAKI, 2018; DEL GAUDIO *et al.*, 2019; GANDHI, LOUGHRAN; MCDONALD, 2019).

Dessa forma, seguindo a mesma abordagem sugerida por Gupta, Simaan e Zaki (2018), e Jiang *et al*. (2019), o presente estudo construiu uma variável de sentimento textual que considera a utilização de relatórios trimestrais (ITR) e das Demonstrações Financeiras Padronizadas (DFP), os quais consistem em uma medida robusta de sentimento ao se fazer uso de mais de um tipo de documento textual, requerido por agentes reguladores e condizente com o contexto de atuação das empresas bancárias listadas no mercado acionário brasileiro.

Estudos que tratam sobre a relação entre sentimento textual e insolvências bancárias podem ser considerados como uma linha de pesquisa incipiente, sendo observada uma quantidade restrita de trabalhos que tratam sobre essa temática, destacando-se as pesquisas de Gupta, Simaan e Zaki (2018), Gandhi, Loughran e McDonald (2019) e Del Gaudio *et al.* (2019).

O primeiro trabalho que examinou a relação entre sentimento textual e falências em bancos foi desenvolvido por Gupta, Simaan e Zaki (2018) ao observarem *bank holding companies* (*BHC*), de capital aberto nos Estados Unidos, utilizando a técnica de aprendizagem de máquina supervisionada *support vector machine* (*SVM*), para identificar se o sentimento pessimista, otimista ou a combinação de ambos os tons textuais demonstravam melhor capacidade preditiva para as firmas insolventes presentes na sua amostra.

Para tanto, os autores fizeram uso de uma elevada quantidade de relatórios 10-K, entre o período de 2005 até 2012, dividindo a sua base de dados em subconjuntos de treinamento e teste, e utilizando o dicionário de Loughran e McDonald (2011) para identificar os termos

associados ao sentimento positivo ou negativo. No entanto, é importante ressaltar que os autores aplicaram ponderações nos pesos das palavras, para a estimação textual, de maneira distinta da proposta por Loughran e McDonald (2011), em que durante a etapa de treinamento dos dados as ponderações desses termos foram calculadas.

Assim, após realizar diversas estimações com e sem o controle das atividades de fusões e aquisições, ocorridas devido a problemas financeiros nessas companhias, bem como controlando as firmas bancárias que receberam ajuda governamental durante o período da crise financeira de 2008, os autores identificaram que o sentimento positivo apresenta um maior poder preditivo para bancos falidos ou em situação de dificuldade financeira, em detrimento do tom pessimista e a utilização conjunta de ambos os tipos de sentimentos textuais.

Dessa forma, Gupta, Simaan e Zaki (2018) argumentam que os bancos falidos apresentam tons textuais mais positivos em relação as firmas bancárias solventes. Indicando por meio dos seus achados que os bancos inadimplentes podem ter utilizado o sentimento otimista com o intuito de repassar sinais positivos para os seus acionistas e investidores. Tais gestores bancários poderiam ainda estar confiando nas suas decisões empresariais para retirar essas companhias de situações futuras de insolvência, sendo demasiadamente otimistas acerca das perspectivas de desempenho do seu negócio.

Considerando essa linha de pesquisa, Gandhi, Loughran e McDonald (2019) observaram o sentimento textual presente nos relatórios 10-K, de bancos listados nas bolsas de valores dos Estados Unidos, partindo do pressuposto de que gestores de firmas em dificuldade financeira tendem a fazer um maior uso de palavras negativas em seus relatórios; assim, foi estudado pelos autores apenas o tom pessimista desses documentos.

Gandhi, Loughran e McDonald (2019) empregaram o modelo estatístico do tipo *logit* para associar comportamentos futuros de fechamento de capital por parte desses bancos, com o aumento do sentimento textual negativo nos seus relatórios financeiros, aplicando para tanto o dicionário de Loughran e McDonald (2011). Os autores indicam que as companhias bancárias que apresentam subsequentes problemas financeiros, após a sua saída de um dos mercados acionários norte-americanos, aumentavam a porcentagem de termos associados ao sentimento textual pessimista em seus relatórios 10-K, antes do fechamento de capital.

Posteriormente, Del Gaudio *et al.* (2019) fizeram uso do *Z-score* para observar a associação entre o risco de insolvência e o sentimento textual de bancos europeus entre o período de 2012 até 2017, por meio dos relatórios anuais dessas companhias bancárias. Assim, foi empregado o dicionário de Loughran e McDonald (2011) para identificar o tom contido nesses documentos.

Embora os autores não especifiquem qual modelo de regressão foi aplicado em seu estudo, mediante o seu resultado foi possível identificar que os bancos com menores probabilidades de se tornarem inadimplentes, demonstraram sentimentos textuais positivos nos seus relatórios. De forma complementar um teste de robustez foi realizado em seu estudo, e identificou que o sentimento textual impacta positivamente no retorno sobre os ativos (*ROA*) e na capitalização desses bancos, porém apresenta uma relação negativa com o desvio padrão móvel do ROA, demonstrando que tons otimistas estão associados a um menor risco de problemas financeiros para aquelas companhias.

As evidências empíricas apresentadas demonstram que o sentimento textual apresenta uma relação com o risco de insolvência dos bancos. Dessa maneira foram identificados resultados divergentes pela literatura, que indicam uma relação positiva ou negativa dessa variável qualitativa com a probabilidade de um comportamento futuro de inadimplência, nesse setor. Desse modo, o presente estudo investigou a relação entre o sentimento textual e uma nova medida de risco de falência bancária, desenvolvida mediante técnicas de aprendizagem de máquina não supervisionada.

3 METODOLOGIA

Como estratégia empírica neste estudo são propostas três etapas: em um primeiro momento foi aplicado o algoritmo *k-means* para realizar os agrupamentos (*clusters*) dos bancos presentes na amostra, elaborando uma medida de risco de insolvência que passou a ser considerada como a variável resposta desta pesquisa, após a sua comparação com os resultados obtidos pelo *Z-score*. Para o desenvolvimento dessa variável de risco foram utilizados os componentes do *Z-score* e as demais variáveis explicativas do estudo descritas nas seções 3.4 e 3.5, exceto a variável PIB. Os dados contábeis utilizados durante essa etapa foram coletados na Economática®.

Posteriormente, foram empregadas técnicas estatísticas de aprendizagem de máquina supervisionada, especificamente os algoritmos de classificação *naive bayes* e *random forest*, bem como o modelo logístico (*logit)*, para prever a classificação das empresas bancárias, com o intuito de identificar qual modelo preditivo é o mais robusto para a variável de risco de insolvência construída.

A última etapa objetiva analisar a forma como as variáveis independentes utilizadas nesta pesquisa, sobretudo o sentimento textual, afetam o risco de insolvência bancária. Essa discussão segue a mesma proposta encontrada nos trabalhos de Gupta, Simaan e Zaki (2018), Del Gaudio *et al.* (2019) e Gandhi, Loughran e McDonald (2019).

3.1 Modelos de aprendizagem de máquina não supervisionados

Considerando os objetivos deste trabalho em construir uma medida de risco de insolvência para os bancos de capital aberto no Brasil, foram utilizados métodos de aprendizagem de máquina (*machine learning*), os quais se configuram como uma área da inteligência artificial que aplica técnicas computacionais, fazendo uso dos dados disponíveis para aprimorar o seu desempenho em atividades associadas à classificação, sendo empregada de forma ampla em pesquisas de finanças, entre as quais destinadas à predição de falências empresariais e ao comportamento futuro dos preços das ações (SENA; SILVA; ARRIAL, 2010).

A principal característica que diferencia as técnicas supervisionadas e não supervisionadas de aprendizagem de máquina consiste na determinação prévia dos valores da variável a ser predita, sendo fornecido para a máquina os números assumidos para essa variável, no primeiro caso. Já nos modelos não supervisionados não é disponibilizado ao algoritmo os

valores da variável resposta, fazendo com que o mesmo identifique os padrões de comportamento dos dados e por meio desses realize a classificação das observações (FERNANDES; CHIAVEGATTO FILHO, 2019; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

Corroborando essa perspectiva Barbosa (2017) destaca ainda que na aprendizagem não supervisionada apenas os dados brutos são fornecidos à máquina, fazendo com que sobretudo os algoritmos que elaboram *clusters*, como o *k-means*, passem a agrupar os dados em conjuntos naturais de acordo com o seu comportamento, em detrimento das técnicas supervisionadas, em que são disponibilizados os atributos das variáveis preditoras e os rótulos, os quais consistem nos valores-alvos que devem ser estimados pelo algoritmo.

3.1.1 *K-Means* e agrupamentos em *clusters*

O agrupamento em *cluster* por meio da técnica *k-means* consiste em um método de aprendizado de máquina não supervisionado e que permite classificar uma base de dados por meio de agrupamentos que minimizam o erro quadrado. De maneira específica o *k-means* faz uso de um número pré-determinado de *clusters* (k) que são considerados pelo algoritmo para que sejam calculadas as distâncias de cada observação da amostra em relação aos *clusters* estabelecidos (VARELLA; QUADRELLI, 2017).

Em seguida, são estimados os centroides, que compreendem o valor médio dos objetos presentes em cada agrupamento ou *cluster*, repetindo esse processo em forma de *loop* até que não ocorra uma variação significativa na distância das observações presentes na base de dados, em relação aos centroides dos k *clusters* estabelecidos (COUTO JÚNIOR; GALDI, 2012; VARELLA; QUADRELLI, 2017).

Cabe ressaltar que na análise desse estudo, o *Z-score* por ser um indicador amplamente utilizado na literatura, foi empregado como métrica de comparação com os resultados obtidos pelo método de agrupamento. Assim, para a realização do agrupamento dos *clusters* foi considerado o desvio padrão móvel do retorno sobre os ativos ($\sigma ROA_{i,t}$), empregado para a execução do algoritmo *k-means*.

Os resultados iniciais demonstram que a classificação para o grupo de menor risco é basicamente a mesma, seja pelo *Z-score* ou por meio dos *clusters*. A Tabela 1 mostra que o método de agrupamento de *clusters* classificou 98,78% das observações de menor risco rotuladas pelo *Z-score*. A maior diferença acontece quando tratamos do grupo de elevado risco. Como pode ser visto na Tabela 1, o *Z-score* categorizou 122 observações no grupo de maior

risco. No entanto, pelo método de *clusters*, houve *matching* com apenas 62 dessas observações, representando 50,82%. Ressalta-se ainda, que ambos os percentuais foram obtidos após a comparação das mesmas observações, pelas duas medidas de classificação empregadas. Esse resultado sugere que o método de classificação pelo agrupamento é mais rigoroso, quanto a categorização do grupo de maior risco, do que o *Z-score*.

Tabela 1 – Comparação do agrupamento dos *clusters* com a classificação do *Z-score*

CLASSIFICAÇÃO	TOTAL DE OBSERVAÇÕES COM O *Z-SCORE*	TOTAL DE OBSERVAÇÕES COM O σROA	OBSERVAÇÕES COMUNS QUANTO AO RISCO	PERCENTUAL DE SIMILARIDADE
Grupo de maior risco (1)	122	66	62	50,82%
Grupo de menor risco (0)	328	384	324	98,78%
Total de observações	450	450	386	-

Fonte: Dados da pesquisa.

Dessa forma, ao comparar os dois métodos empregados, optou-se pela utilização da variável $\sigma ROA_{i,t}$, executada pelo algoritmo *k-means*, considerando a homogeneidade dentro dos *clusters*, propiciada pela aplicação desse método computacional, que também permite assegurar a heterogeneidade entre estes grupos (VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

Nesse sentido, o desvio padrão móvel do retorno sobre os ativos ($\sigma ROA_{i,t}$), consiste em um indicador da volatilidade dos lucros bancários (FERREIRA; ZANINI; ALVES, 2019; VIEIRA *et al.*, 2020). Conforme argumentam Chiaramonte *et al.* (2016), uma segunda limitação inerente ao *Z-score* se refere à sua sensibilidade ao σROA, uma vez que bancos com valores elevados para essa medida tendem a apresentar pontuações menores no *Z-score*, indicando maiores probabilidades para um comportamento futuro de insolvência.

Em seguida, foi realizada a segunda etapa da estratégia empírica deste estudo, destinada à identificação do método mais robusto para a predição dessa variável de risco de insolvência.

3.2 Métodos de aprendizagem de máquina supervisionados

Os modelos supervisionados de aprendizagem de máquina ganham destaque pela sua aplicabilidade, pois, dependendo do objetivo da pesquisa, esses métodos podem ser utilizados em regressões, cuja variável resposta apresenta valores contínuos, bem como para modelos de classificação, em que a variável predita é categórica, fazendo com que o algoritmo estime em qual classe uma nova observação na base de dados deve fazer parte, considerando para essa definição os atributos das variáveis preditoras e as categorias (ou rótulos) previamente estabelecidas para a variável resposta (SILVA; RIBEIRO; MATIAS, 2016; FERNANDES; CHIAVEGATTO FILHO, 2019).

Dessa forma, fez-se uso de uma divisão na base de dados em subamostras de treinamento e teste (*cross-validation*), cuja primeira é empregada para o ajuste do modelo e a última destinada ao exame do seu desempenho (validação), sendo essa etapa essencial para a execução de modelos supervisionados (BARBOSA, 2017; ROSA; GARTNER, 2018; FERNANDES; CHIAVEGATTO FILHO, 2019).

Em virtude desse fracionamento, 70% dos dados da amostra final foram considerados para construção da subamostra de treinamento e 30% aplicados para a amostra teste. Ressalta-se ainda que esses percentuais também foram utilizados por Rosa e Gartner (2018) e Viswanathan, Srinivasan e Hariharan (2020) ao realizarem essa etapa nos seus estudos. A escolha das observações que fazem parte das subamostras foi realizada de maneira aleatória pelo algoritmo e mantendo as proporções para a variável categórica nos subconjuntos de treino e teste, condizentes com os percentuais observados na base de dados completa.

3.2.1 Modelo de classificação *naive bayes*

Dentre os modelos supervisionados de aprendizagem de máquina, as redes bayesianas consistem em um dos métodos mais utilizados e que se configuram como uma classe de modelos estatísticos que apresentam resultados significativos para lidar com eventos de elevada incerteza, sendo aplicado o teorema de Bayes para identificar comportamentos relacionados à dependência probabilística condicional (SILVA; RIBEIRO; MATIAS, 2016; SOMBRA *et al.*, 2020).

Especificamente esse modelo utiliza um grupo de variáveis aleatórias (atributos) retratadas em um grafo estatístico por meio de nós e arcos, os quais são definidos em função de uma relação de precedência (condicional), ou seja, refletem a probabilidade de ocorrência de

um evento de interesse em estudo, em virtude da ocorrência de um outro evento tido como condicional, obtendo dessa forma a tabela de probabilidade condicional. Por meio dessa abordagem, tem-se o algoritmo de classificação *naive bayes* (SILVA; RIBEIRO; MATIAS, 2016).

Assim, durante a etapa de treinamento, o algoritmo *naive bayes* faz uso dos dados dessa subamostra para compreender quais os valores condicionais das variáveis independentes (atributos) que estão associadas às classes do modelo, para que em uma segunda etapa denominada de validação, utilizando os dados da base de teste, seja possível realizar predições sobre a classe de cada observação na amostra, fazendo uso dos valores das variáveis preditoras identificados pelo modelo na etapa de treinamento (WANKE *et al.*, 2014; SOMBRA *et al.*, 2020).

3.2.2 Modelo de classificação *random forest*

O modelo *random forest* também se caracteriza como um algoritmo de classificação que apresenta vasta aplicabilidade na literatura acadêmica, constitui-se por um conjunto de árvores de decisões, as quais são elaboradas para uma subamostra dos dados, originada de forma aleatória mediante a técnica de *bootstrapping* (BREIMAN, 2001; SUSS; TREITEL, 2019).

A árvore de decisão é uma estrutura hierárquica que faz uso da condicionante *if-then* (se-então), em que durante a etapa de treinamento do algoritmo são determinados os melhores valores de separação (pontos de corte) para cada ramificação (ou nó) dessa árvore, o qual por sua vez representa uma das variáveis preditoras do modelo (atributos). Assim sendo, tais valores passam a ser fundamentais, pois para cada nó pai dessa árvore são estabelecidos os nós filhos que satisfazem essa condição, localizados sempre à esquerda, e aqueles que não atendem a esse requisito, estabelecidos à direita, sendo esse processo realizado até que todas as observações sejam categorizadas (BARBOSA, 2017).

Assim, para a execução desse modelo de classificação a amostra de treinamento é utilizada como base de dados, permitindo a criação de diversas subamostras menores estabelecidas por meio do método de *bootstrapping*, e que apresentam os mesmos padrões dos dados originais de treinamento, em que para cada um desses subconjuntos é construída uma árvore de decisão, a qual contribui com um único voto para a execução da predição realizada pelo algoritmo *random forest*. Este último, por sua vez, considera o conjunto dessas árvores de decisões (floresta) para a concretização das suas classificações (BARBOSA, 2017; SILVA; RÊGO; FRASCAROLI, 2019).

Dessa forma, a indicação de companhias que apresentam maior risco de insolvência ocorre na medida em que uma proporção maior de árvores categoriza aquela instituição bancária no grupo de elevado risco, ou seja, o *random forest* fornece uma probabilidade prevista para cada indivíduo da amostra, mediante as regras de decisões estabelecidas para as variáveis independentes (SUSS; TREITEL, 2019).

3.2.3 Modelo *logit*

A regressão logística é considerada uma das técnicas estatísticas mais empregadas em modelos de predição da falência bancária, conforme é destacado por Liberman, Barbosa e Pires (2018) e Rosa e Gartner (2018). Nesse modelo é utilizada uma transformação para que a sua variável resposta apresente resultados entre 0 e 1, os quais consistem nas probabilidades de ocorrência do evento de interesse.

Silva, Ribeiro e Matias (2016) ressaltam que a principal diferença entre o modelo de regressão logística e as demais técnicas lineares aplicadas em regressões consiste no fato de que a primeira objetiva estimar a probabilidade de a variável predita fazer parte de uma categoria (normalmente 0 ou 1), em contraste com os modelos lineares que se destinam a modelar de forma direta uma estimação para a sua variável resposta. Assim sendo, o modelo funcional da regressão logística segue a equação 1, em que a probabilidade de cada observação fazer parte da categoria de maior (um) ou menor risco (zero) é representada por PA.

$$PA = \frac{1}{1 + e^{-(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n)}} \quad (1)$$

Na equação 1, os $\beta_1 .... \beta_n$ correspondem aos coeficientes das variáveis, as quais também podem ser consideradas como as características ou atributos que são representados por $x_1 ... x_n$, conforme destacado por Silva, Ribeiro e Matias (2016).

3.3 Validação e acurácia dos modelos

Para examinar a precisão dos modelos testados, seguindo a abordagem de estudos anteriores que aplicam as mesmas técnicas de classificação, foi utilizada a medida estatística de acurácia, objetivando analisar a eficácia desses métodos para a escolha daquele que demonstra maior robustez (SILVA; RIBEIRO; MATIAS, 2016; BOUGHACI; ALKHAWALDEH, 2018).

Nesse sentido, para o cálculo da acurácia em um primeiro momento é necessário fazer uso da matriz de confusão, sendo essa considerada como uma ferramenta útil durante a etapa de avaliação de modelos de classificação, ao fornecer os dados necessários para a mensuração das medidas de especificidade e sensibilidade, as quais posteriormente são utilizadas para o cálculo da acurácia dos modelos (HAN; KAMBER; PEI, 2012; BOUGHACI; ALKHAWALDEH, 2018; SILVA RÊGO; FRASCAROLI, 2019).

Acerca da especificidade, essa é considerada como a taxa verdadeira negativa, demonstrando a proporção de bancos que apresentam menor risco de insolvência e que foram classificados de maneira adequada; por sua vez, a sensibilidade compreende o índice verdadeiro positivo, indicando o percentual de instituições bancárias que possuem elevado risco de falência futura, e que foram categorizadas corretamente pelos modelos. Seguindo Silva, Ribeiro e Matias (2016), Boughaci e Alkhawaldeh (2018) e Viswanathan, Srinivasan e Hariharan (2020), a acurácia dos modelos de classificação foi mensurada conforme a equação 4; já nas equações 2 e 3 são apresentados respectivamente os cálculos das medidas de sensibilidade e especificidade.

$$sensibilidade = \frac{TP}{TP + FN} \tag{2}$$

Em que:

$sensibilidade$ = Taxa de verdadeiros positivos, indicando o percentual de instituições bancárias que possuem elevado risco de falência futura, e que foram classificadas corretamente;

TP = Verdadeiro positivo, indica o número de casos positivos (maior risco de insolvência) que foram corretamente identificados;

FN = Falso negativo, indica o número de casos positivos (maior risco de insolvência) que foram classificados de forma incorreta como casos negativos (menor risco de insolvência).

$$especificidade = \frac{TN}{TN + FP} \tag{3}$$

Em que:

$especificidade$ = Taxa de verdadeiros negativos, indica a proporção de bancos que apresentam menor risco de insolvência e que foram classificados corretamente;

TN = Verdadeiro negativo, indica o número de casos negativos (menor risco de insolvência) que foram corretamente identificados;

FP = Falso positivo, indica o número de casos negativos (menor risco de insolvência) que foram classificados de forma incorreta como casos positivos (maior risco de insolvência).

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Em que:

$Acurácia$ = Índice de acurácia do modelo de classificação;

TP = Verdadeiro positivo, indica o número de casos positivos (maior risco de insolvência) que foram corretamente identificados;

TN = Verdadeiro negativo, indica o número de casos negativos (menor risco de insolvência) que foram corretamente identificados;

FP = Falso positivo, indica o número de casos negativos (menor risco de insolvência) que foram classificados de forma incorreta como casos positivos (maior risco de insolvência);

FN = Falso negativo, indica o número de casos positivos (maior risco de insolvência) que foram classificados de forma incorreta como casos negativos (menor risco de insolvência).

3.4 Construção da variável dependente

3.4.1 Risco de insolvência

Uma das medidas amplamente difundida na literatura bancária para mensurar o risco de insolvência nesse setor é o *Z-score* (LEPETIT; STROBEL, 2013). Dessa forma, nesta pesquisa foi utilizado o *Z-score* desenvolvido por Boyd, Nicolò e Jalal (2006) na seção III. A do seu estudo, contudo seguindo a mesma abordagem de Lepetit e Strobel (2013), Vieira e Girão (2016) e Vieira *et al.* (2020) para o cálculo dessa variável, a qual foi aplicada nas referidas pesquisas para avaliar o risco de insolvência dos bancos atuantes nos países do G20 e no mercado nacional, respectivamente.

Conforme destacado em trabalhos anteriores (BOYD, GRAHAM, 1986; BOYD; NICOLÒ; JALAL, 2006; CHIARAMONTE *et al.*, 2016), o *Z-score* representa o número de desvios padrões abaixo da média, necessários para que o lucro bancário decaia, ao ponto do seu patrimônio líquido se tornar negativo, fazendo com que aquele se configure como uma empresa insolvente. Assim, quanto maior o valor do *Z-score* menor é a probabilidade de falência da companhia.

Nesse sentido, Lepetit e Strobel (2013), Vieira e Girão (2016) e Vieira *et al.* (2020) endossam que o *Z-score* representa a distância da falência em que um determinado banco se encontra, caracterizando-se como uma medida de risco e, por conseguinte, um indicador de probabilidade de insolvência, que faz uso dos dados contábeis da empresa para tanto.

Seguindo Lepetit e Strobel (2013), Vieira e Girão (2016) e Vieira *et al.* (2020), o *Z-score* foi mensurado por meio das médias móveis e do desvio padrão móvel considerando doze trimestres (três anos), em que os valores dos onze trimestres anteriores e o do trimestre contemporâneo foram utilizados para essa mensuração, conforme apresentado na equação 5.

$$Z\_score_{i,t} = \frac{\mu ROA_{i,t} + \mu ETS_{i,t}}{\sigma ROA_{i,t}} \quad (5)$$

Em que:

$Z\_score_{i,t}$ = risco de insolvência do banco i, no período t;

$\mu ROA_{i,t}$ = média móvel de doze trimestres do retorno sobre os ativos do banco i, no período t;

$\mu ETS_{i,t}$ = média móvel de doze trimestres da razão entre o patrimônio líquido e o ativo total para a firma i, no período t;

$\sigma ROA_{i,t}$ = desvio padrão móvel de doze trimestres do retorno sobre os ativos da empresa i, no período t.

Devido à necessidade de se trabalhar com as médias móveis e com o desvio padrão móvel para o cálculo da variável *Z-score*, o período de análise do estudo compreendeu do quarto trimestre de 2012 até o último trimestre de 2019, utilizando para tanto os dados contábeis a partir do primeiro trimestre de 2010, procedimento esse também realizado por Vieira e Girão (2016) e Vieira *et al.* (2020), ao fazerem uso do *Z-score* para mensurar o risco de insolvência dos bancos presentes no mercado brasileiro.

Com o intuito de categorizar os bancos que apresentam um maior risco de insolvência foi utilizado como ponto de corte o limite inferior da variável *Z-score*, o qual foi calculado para cada trimestre analisado. A escolha desse limiar se deu uma vez que os valores menores do *Z-score* denotam firmas que apresentam maior probabilidade de insolvência, assim os bancos identificados no primeiro quartil receberam o valor 1, já os demais foram categorizados com o valor 0.

Adicionalmente, corroborando essa perspectiva, Chiaramonte *et al.* (2016) identificaram em sua pesquisa, após a aplicação de quatro medidas distintas para o cálculo do *Z-score*, que em média os bancos falidos da sua base de dados obtiveram valores inferiores para a referida variável, em detrimento das firmas solventes desse setor.

Conforme destacado na seção 3.1.1 a classificação de risco indicada pelo *Z-score* foi comparada com os resultados obtidos pela medida de risco $\sigma ROA_{i,t}$ elaborada por meio do algoritmo *k-means*.

3.5 Variáveis independentes

3.5.1 Sentimento textual bancário

Com o intuito de atender a um dos objetivos deste estudo, foi elaborada uma variável de sentimento textual para os bancos presentes no mercado de capitais brasileiro, considerando que os documentos ITR e DFP dessas organizações exigidos pela CVM apresentam sentimentos textuais pessimistas e otimistas (GUPTA; SIMAAN; ZAKI, 2018; JIANG *et al.*, 2019).

Dessa maneira, para construção da variável de sentimento textual, inicialmente foi obtida a classificação setorial disponibilizada no *site* da Brasil, Bolsa, Balcão (B3), a qual apresenta a lista dos bancos de capital aberto no país. Posteriormente, foram capturados os relatórios trimestrais (ITR) e as Demonstrações Financeiras Padronizadas (DFP) dessas companhias por meio do endereço eletrônico da CVM, apanhando dessa forma os relatórios dos três primeiros trimestres de cada ano analisado, mediante os documentos ITR e o último trimestre por sua vez, correspondente aos relatórios DFP.

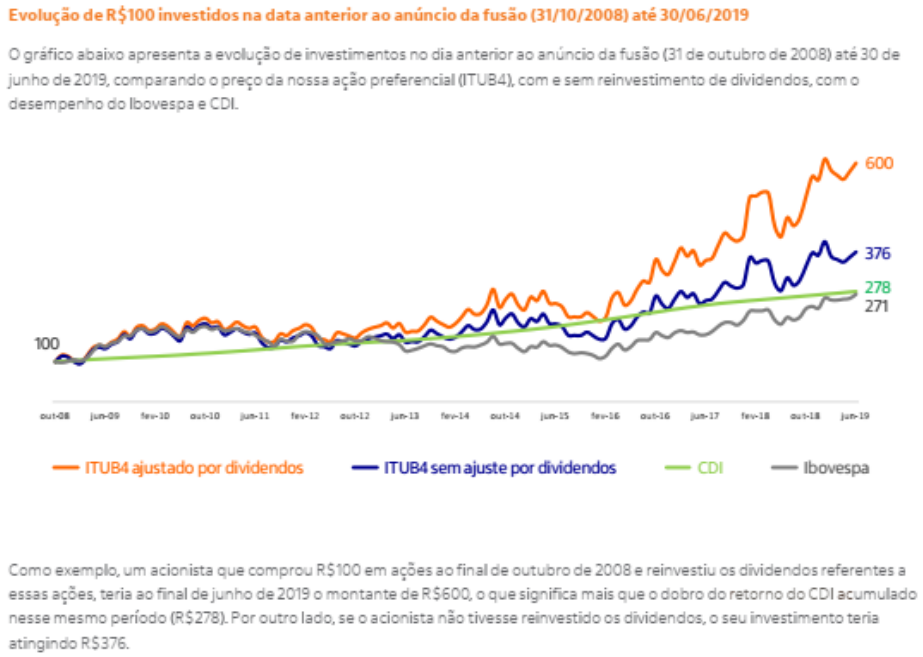
Após a obtenção dos documentos ITR e DFP, que fazem parte da base de dados do estudo, teve início a etapa de transformação dos arquivos, os quais foram capturados originalmente em formato *PDF* (*Portable Document Format*), para TXT (texto separado por tabulações). Posteriormente, estes arquivos gerados na etapa anterior passaram por um processo de pré-ajuste da amostra, sendo removidos os espaços duplos, pontuações, números, bem como os *stopwords*, que se configuram como uma lista de preposições, pronomes, conjunções e formas verbais que não apresentam relevância explicativa em documentos textuais.

Acerca dessa última prática Savoy e Gaussier (2010) consideram que a retirada de *stopwords* é necessária para o correto processamento dos textos, uma vez que tais palavras podem impactar na leitura automatizada, consistindo em um ruído durante a referida etapa.

Em seguida, foi observado que alguns termos, que apresentavam a letra “ç” não estavam sendo considerados na leitura computacional; então, optou-se por substituí-los pela letra “c”, o que se configurou em um melhor ajuste para esse procedimento, sem perdas na observação dessas palavras pela máquina, de forma semelhante ao executado por Galdi e Gonçalves (2018).

De maneira complementar foram realizados testes com documentos de controle, considerando a possibilidade de que os gráficos e as figuras presentes nos relatórios financeiros, pudessem limitar a interpretação computacional dos textos, o que poderia impedir a máquina de realizar a análise das palavras após a presença de dados visuais (MACHADO; SILVA, 2017). Conforme destacado na Figura 1, os documentos ITR e DFP apresentam em sua composição algumas imagens, sendo necessária a realização dessa etapa de teste para comprovar que a interpretação estava sendo realizada de maneira adequada pelo algoritmo.

Figura 1 – Relatório ITR do Itaú Unibanco *Holding* S.A. – 30 de junho de 2019



Fonte: Elaborado pelo autor.

Para explorar as informações textuais presentes nos relatórios ITR e DFP, foi empregada a Análise de Processamento Natural, mediante algoritmos de leitura automatizada escritos em linguagem R, sendo aplicada a técnica do *vector space model*, desenvolvida por Chisholm e Kolda (1999) e utilizada em estudos anteriores sobre essa temática (DEL GAUDIO *et al.*, 2019; GANDHI; LOUGHRAN; MCDONALD, 2019), que consideram as palavras presentes nos textos como vetores, os quais serão utilizados para a estimação do peso de cada palavra em um

determinado documento de acordo com a frequência das mesmas, conforme apresentado na equação 6.

$$P_{i,j} \begin{cases} \dfrac{(1+\log(Tf_{i,j}))}{(1+\log(a_j))} \; x \; log \; \dfrac{N}{df_i}, se \; Tf_{i,j} \geq 1 \\ 0 \qquad\qquad\qquad\qquad , se \; Tf_{i,j} = 0 \end{cases} \tag{6}$$

Em que $P_{i,j}$ consiste no peso da palavra *i* no documento *j*, por sua vez, $Tf_{i,j}$ compreende a totalidade de ocorrências de uma determinada palavra i em um relatório j, a_j diz respeito à média de frequências das palavras de um documento financeiro, *N* é o total de relatórios da amostra, e, df_i é o total de relatórios administrativos com ao menos uma ocorrência da palavra *i*.

Na equação 6, será considerado que o peso das palavras se deu mediante ponderação da frequência dos termos encontrados no total de relatórios contidos na amostra. Assim, a equação 6 também pode ser compreendida por meio de uma divisão em duas partes, cuja primeira representa o peso da palavra *i* no documento *j*, sendo dada pelo termo $\frac{(1+\log(Tf_{i,j}))}{(1+\log(a_j))}$, já a segunda considera a expressão $\frac{N}{df_i}$ a qual está associada ao peso global que a palavra representa dentro do conjunto de documentos corporativos da amostra.

A aplicação de ponderações com logaritmos foi realizada objetivando minimizar a atuação de palavras de alta frequência (*outliers*) nos documentos da base de dados, evitando que aqueles apresentem um peso maior na estimação. Acerca desse assunto, Loughran e McDonald (2011) argumentam que tal prática reduz a interferência de *outliers*, comprovando a eficácia desse método após examinarem relatórios 10-K, produzidos por firmas norte-americanas.

No Brasil, Pagliarussi, Aguiar e Galdi, (2016) corroboram com o uso desse procedimento ao defender que a aplicação de pesos é necessária na análise textual de relatórios financeiros, pois mesmo que uma palavra seja empregada com grande frequência em um determinado texto, não necessariamente esta apresenta elevada carga informacional em relação aos demais termos.

Para a estimação do sentimento textual contido nos relatórios da base de dados, foi considerado o peso das palavras positivas e negativas empregadas nesses documentos, as quais foram identificadas em uma dessas categorias, por meio do dicionário e do algoritmo de leitura

desenvolvidos por Silva e Machado (2019), sendo desconsiderados os termos que não estavam associados, no referido dicionário, a nenhuma das duas tipologias de sentimento.

A aplicação do dicionário de Silva e Machado (2019) apresenta algumas vantagens: a primeira consiste no fato de ter sido elaborado para textos de caráter financeiro, bem como por considerar durante a sua construção as especificidades das companhias bancárias, sendo adequado para a utilização em estudos sobre esse setor. Por fim, destaca-se ainda que o mesmo foi elaborado para a língua portuguesa, pois, conforme evidenciado em estudos anteriores (LOUGHRAN; MCDONALD, 2011; PAGLIARUSSI; AGUIAR; GALDI, 2016; MACHADO; SILVA, 2017), dicionários que consideram durante o seu processo de criação as particularidades do idioma e a finalidade para a qual serão utilizados, passam a ter tais características tidas como fatores positivos e que melhoram a estimação textual.

Assim, fazendo uso do dicionário destacado, e após o cálculo do peso de cada palavra otimista e pessimista nos relatórios ITR e DFP, foi realizada a mensuração do sentimento textual para identificar o tom de cada relatório da amostra, conforme a equação 7, a seguir:

$$SB_j = \frac{\sum Peso\ das\ Palavras\ Positivas - \sum Peso\ das\ Palavras\ Negativas}{\sum Peso\ das\ Palavras\ Positivas + \sum Peso\ das\ Palavras\ Negativas} \qquad (7)$$

Dessa forma, o sentimento textual SB_j do relatório j foi admitido como otimista, para resultados com valores próximos de 1; por sua vez, aquele foi determinado como neutro, quando o resultado da equação indicou valor igual a 0, e, por fim, o relatório financeiro foi considerado de tom pessimista, na medida em que os valores da estimação estiveram próximos de menos 1.

3.5.2 Variáveis de controle

As pesquisas que tratam sobre o risco de falências bancárias fazem uso de medidas contábeis e de indicadores macroeconômicos, as primeiras objetivam mensurar os fatores que possibilitam diferenciar as companhias presentes na amostra, já os últimos tendem a refletir o risco sistemático que influencia o comportamento de todos os tipos de empreendimentos, uma vez que as empresas, de maneira geral, são susceptíveis a tais ameaças (CHIARAMONTE *et al.*, 2016; VIEIRA; GIRÃO, 2016; ROSA; GARTNER, 2018; FERREIRA; ZANINI; ALVES, 2019; ANDRADE, 2020).

Considerando essa abordagem, o presente estudo utilizou a taxa de crescimento trimestral real do PIB como uma variável macroeconômica que pode afetar a probabilidade de

insolvência dos bancos, pois as condições econômicas impactam na necessidade de crédito das empresas, bem como na sua capacidade em liquidar os seus empréstimos, prejudicando assim as instituições bancárias enquanto provedoras de capital, e as tornando vulneráveis em momentos de dificuldade financeira nos países (CHIARAMONTE *et al.*, 2016; VIEIRA; GIRÃO, 2016; FERREIRA; ZANINI; ALVES, 2019).

Os dados referentes à taxa de crescimento trimestral real do Produto Interno Bruto (PIB) foram coletados por meio do endereço eletrônico do Instituto Brasileiro de Geografia e Estatística (IBGE), correspondendo à variação percentual real do PIB trimestral em relação ao trimestre imediatamente anterior.

No que se refere à utilização de dados contábeis, uma das *proxies* mais utilizadas pela literatura é o tamanho dos bancos, pois as corporações que possuem grande porte tendem a apresentar riscos menores de insolvência, ao diversificarem a sua atuação, o que lhes possibilita ganhos menos voláteis devido ao seu modelo de negócios que se faz presente em diversos segmentos bancários, ofertando, assim, uma elevada quantidade de serviços aos seus clientes (DEL GAUDIO *et al.*, 2019; FERREIRA; ZANINI; ALVES, 2019; VIEIRA *et al.*, 2020).

Contudo, também é importante destacar que as maiores firmas desse setor, em virtude da sua importância para a economia, e os riscos associados ao efeito de contágio inerente à sua falência, apresentam uma expectativa de serem socorridas pelo governo em situações de insolvência, sendo então denominadas de companhias *too-big-to-fail*, o que pode resultar em decisões por parte dos seus executivos mais voltadas à assunção de riscos, devido a essa garantia implícita de socorro governamental, com o intuito de aumentarem os lucros da empresa (VIEIRA; GIRÃO, 2016; FERREIRA; ZANINI; ALVES, 2019; ANDRADE, 2020).

Diante dessa relação dicotômica, na literatura não existe um consenso no que se refere ao impacto dessa variável no risco de insolvência dos bancos (CHIARAMONTE *et al.*, 2016; ANDRADE, 2020). Nesta pesquisa, a variável referente ao tamanho dos bancos foi mensurada por meio do logaritmo natural do ativo total, conforme os estudos de Chiaramonte *et al.* (2016), Vieira e Girão (2016), Del Gaudio *et al.* (2019), Ferreira, Zanini e Alves (2019) e Andrade (2020).

A lucratividade dos bancos também está associada ao seu risco de insolvência, representada pelo retorno sobre os ativos (ROA), partindo do pressuposto de que bancos que apresentam menor lucratividade tendem a ter maiores probabilidades de insolvência, sendo essa uma *proxy* utilizada em pesquisas anteriores. Vieira e Girão (2016), Liberman, Barbosa e Pires (2018) e Rosa e Gartner (2018), especificamente, Vieira *et al.* (2020) endossam que firmas com menos rentabilidades estão mais propensas a demonstrar maiores volatilidades em sua

lucratividade, e, por conseguinte, elevam a sua probabilidade de falência. Na equação 8 é apresentado o cálculo da variável $ROA_{i,t}$, da mesma forma que foi mensurada por Vieira e Girão (2016) e Vieira *et al.* (2020).

$$ROA_{i,t} = \frac{Lucro\ operacional_{i,t}}{Ativo\ total_{i,t}} \quad (8)$$

Em que:

$ROA_{i,t}$ = retorno sobre os ativos do banco i, no trimestre t;

$Lucro\ operacional_{i,t}$= lucro operacional da empresa i, no trimestre t;

$Ativo\ total_{i,t}$= ativo total da firma i, no trimestre t.

Por sua vez, a capitalização dos bancos foi mensurada por meio da variável $ETS_{i,t}$, pois, conforme destacado por Vieira e Girão (2016), um maior nível de capitalização limita o risco de insolvência de um banco. Dessa forma Ferreira, Zanini e Alves (2019) endossam que essa medida representa o nível de aversão ao risco de uma companhia bancária. De maneira complementar, Chiaramonte *et al.* (2016) identificaram que os bancos americanos falidos apresentam valores menores para a capitalização em detrimento das firmas solventes desse setor. Assim, a variável $ETS_{i,t}$ foi estimada de acordo com os trabalhos de Vieira e Girão (2016), Liberman, Barbosa e Pires (2018) e Rosa e Gartner (2018) destacada na equação 9.

$$ETS_{i,t} = \frac{Patrimônio\ líquido_{i,t}}{Ativo\ total_{i,t}} \quad (9)$$

Em que:

$ETS_{i,t}$= capitalização da empresa i, no período t;

$Patrimônio\ líquido_{i,t}$= patrimônio líquido do banco i, no período t;

$Ativo\ total_{i,t}$= ativo total da companhia i, no período t.

A liquidez também é apresentada na literatura como um fator preponderante para comportamentos associados a um maior risco de insolvência. Dessa forma essa variável tem como objetivo mensurar o nível de endividamento dos bancos, seja esse de curto ou longo prazo. Conforme destacado por Rosa e Gartner (2018), um incremento na liquidez tende a reduzir a probabilidade de insolvência. Essa variável foi calculada seguindo a equação 10, de maneira semelhante ao estudo de Barbosa (2017).

$$LIQ_{i,t} = \frac{Depósitos\ totais_{i,t}}{Ativo\ total_{i,t}} \quad (10)$$

Em que:

$LIQ_{i,t}$= liquidez da firma i, no período t;

$Depósitos\ totais_{i,t}$= depósitos totais da empresa i, no período t;

$Ativo\ total_{i,t}$= ativo total do banco i, no período t.

A amostra inicial do estudo foi composta por 25 bancos de capital aberto, após a coleta dos relatórios ITR e DFP no endereço eletrônico da CVM. Posteriormente, foram obtidos na base de dados Ecomomatica®, os dados contábeis correspondentes às informações consolidadas de cada banco, sendo removidas as companhias que não apresentavam dados contábeis consolidados.

Por fim, foram retirados os bancos que são listados por um curto período e que não dispunham do volume de dados necessários para o cálculo do *Z-score*, uma vez que são empregadas médias móveis e o desvio padrão móvel para o seu cômputo, totalizando assim uma amostra final com 17 bancos e 450 observações, por meio de um painel de dados não balanceado. Os procedimentos metodológicos foram executados nos *softwares* estatísticos Python e R.

3.6 Modelo econométrico

A seguir é apresentado o modelo econométrico adotado no estudo, e as variáveis de controle que foram empregadas para a sua estimação, considerando os trabalhos desenvolvidos por Chiaramonte *et al.* (2016), Vieira e Girão (2016), Liberman, Barbosa e Pires (2018), Gupta, Simaan e Zaki (2018), Del Gaudio *et al.* (2019), Ferreira, Zanini e Alves (2019) e Vieira *et al.* (2020), conforme a equação 11.

$$Risco_{i,t} = \alpha_{i,t} + \beta_1 PIB_t + \beta_2 ETS_{i,t} + \beta_3 ROA_{i,t} + \beta_4 SB_{i,t-1} + \beta_5 LIQ_{i,t} + \beta_6 TAM_{i,t} + \varepsilon_{i,t} \quad (11)$$

Em que:

$Risco_{i,t}$= variável binária obtida por meio do *cluster k-means*, assumindo valor 1, caso a empresa esteja no agrupamento de maior risco e 0, caso contrário;

PIB_t= Variação percentual do Produto Interno Bruto real brasileiro no trimestre t, em comparação ao trimestre t-1;

$ETS_{i,t}$= capitalização da companhia i, no período t;

$ROA_{i,t}$ = retorno sobre os ativos da empresa i, no trimestre t;

$SB_{i,t-1}$= tom do sentimento textual bancário do relatório i, no período t-1;

$LIQ_{i,t}$= liquidez da firma i, no período t;

$TAM_{i,t}$= tamanho do banco i, no período t;

$\varepsilon_{i,t}$= resíduos da equação;

$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$= coeficientes do modelo.

A variável de sentimento textual foi utilizada em um período anterior aos das demais variáveis pois, conforme Del Gaudio *et al.* (2019), podem ocorrer atrasos entre a disseminação das informações contidas nos relatórios financeiros e o efeito real desses dados sobre o desempenho da firma. A seguir, no Quadro 1, são apresentados os sinais esperados para as variáveis preditoras.

Quadro 1 – Sinais esperados para as variáveis explicativas e de controle

VARIÁVEL	OPERACIONALIZAÇÃO	SINAL ESPERADO	REFERÊNCIAS
Sentimento Textual Bancário (SB)	Tom do sentimento textual de cada relatório da amostra	Positivo/Negativo	Gupta, Simaan e Zaki (2018), Del Gaudio *et al.* (2019) e Gandhi, Loughran e McDonald (2019).
Tamanho (TAM)	Logaritmo natural do ativo total	Positivo/Negativo	Chiaramonte *et al*. (2016), Vieira e Girão (2016), Del Gaudio *et al*. (2019) e Andrade (2020).
Liquidez (LIQ)	Razão entre depósitos totais e ativo total	Negativo	Barbosa (2017).
Retorno sobre os ativos (*ROA*)	Razão entre o lucro operacional e ativo total	Negativo	Vieira e Girão (2016), e Vieira *et al*. (2020)
Capitalização (*ETS*)	Razão entre patrimônio líquido e ativo total	Negativo	Vieira e Girão (2016), Liberman, Barbosa e Pires (2018), e Rosa e Gartner (2018).
Produto Interno Bruto (PIB)	Variação percentual do PIB real em relação ao trimestre imediatamente anterior	Negativo	Chiaramonte *et al*. (2016), Vieira e Girão (2016), Del Gaudio *et al*. (2019), Ferreira, Zanini e Alves (2019) e Andrade (2020).

Fonte: Elaborado pelo autor.

Para a análise empírica realizada neste estudo foram empregados três tipos de modelos: o primeiro consiste no modelo clássico *logit* para dados em painel, destacado na literatura

bancária pela sua capacidade preditiva em identificar riscos de *default* (insolvência) e pela aderência que apresenta ao comportamento empírico desse tipo de evento (LIBERMAN; BARBOSA; PIRES, 2018). Os demais modelos empregados fazem uso de ferramentas supervisionadas de aprendizagem de máquina, consistindo nos modelos *naive bayes* e *random forest*. Os resultados das estimações são apresentados na seção de análise dos resultados.

4 ANÁLISE DOS RESULTADOS

Os resultados apresentados nessa seção foram obtidos após todas as variáveis serem winsorizadas a 8%, para tratar os *outliers* presentes na amostra, uma vez que as winsorizações realizadas com percentuais inferiores ao destacado (1% e 5%) ainda permitiam a presença de pontos extremos na base de dados.

4.1 Estatísticas descritivas

Nesta seção são abordadas as estatísticas descritivas referentes aos resultados da pesquisa, cujos valores associados à variável categórica dependente $Risco_{i,t}$, que compreende a medida de risco de insolvência construída por meio do algoritmo *k-means*, não foram contempladas devido a sua natureza, enquanto variável *dummy*. Destaca-se ainda que no decorrer do trabalho o termo bancos foi utilizado para se referir às instituições financeiras de capital aberto, analisadas em seu nível consolidado. A Tabela 2 detalha as estatísticas descritivas das variáveis empregadas no estudo.

Tabela 2 – Estatísticas descritivas

Variável	Média	Mediana	DP	Máximo	Mínimo
PIB	0,007	0,197	0,818	1,194	-1,486
ETS	0,104	0,096	0,034	0,180	0,061
ROA	0,003	0,003	0,004	0,009	-0,005
SB	0,005	0,007	0,068	0,124	-0,103
LIQ	0,400	0,382	0,182	0,667	0,125
TAM	17,61	16,86	2,07	21,06	15,23

Legenda: *PIB* indica a variação percentual do Produto Interno Bruto real brasileiro em comparação ao trimestre imediatamente anterior ao trimestre t; *ETS* indica o nível de capitalização do banco i, no período t, mensurada pela razão entre patrimônio líquido e ativo total; *ROA* indica a rentabilidade do banco i, no período t, mensurada pela razão entre lucro operacional e ativo total; *SB* indica ao tom do sentimento textual do documento i, no período t-1, mensurada pela ponderação dos pesos das palavras pessimistas e otimistas em cada relatório ITR ou DFP; *LIQ* indica a liquidez do banco i, no período t, mensurada pela razão entre depósitos totais e ativo total; *TAM* indica o tamanho do banco i, no período t, mensurado pelo logaritmo natural do ativo total.
Estatísticas obtidas após winsorização realizada a 8% em todas as variáveis.
Fonte: Dados da pesquisa.

No que se refere ao sentimento textual bancário, ao observar as suas medidas centrais, é possível identificar que durante o período analisado os relatórios financeiros ITR e DFP dessas instituições, apresentaram palavras que de maneira geral denotam um tom textual de neutralidade, considerando que a sua média de 0,005 e mediana de 0,007 apresentam valores marginalmente próximos de zero.

Ao observar os números máximos e mínimos é possível identificar a presença de bancos que apresentam em seus relatórios sentimentos textuais distintos, denotando o comportamento heterogêneo por parte dessas firmas financeiras, no que se refere ao tom textual dos seus documentos ITR e DFP.

Esses achados são opostos aos resultados obtidos por Del Gaudio *et al.* (2019), ao analisarem as companhias bancárias europeias entre os anos de 2012 a 2017, identificando para aquelas instituições financeiras um sentimento textual preponderantemente pessimista; contudo, é importante destacar que os autores não especificaram em sua pesquisa quais os bancos (capital aberto, fechado ou ambos) que foram observados.

Os demais estudos que tratam sobre o sentimento textual e a insolvência bancária não apresentam estatísticas descritivas em relação ao tom dos relatórios analisados para toda a sua amostra, indicando apenas as palavras positivas e negativas mais frequentes no caso do trabalho de Gupta, Simaan e Zaki (2018). Por sua vez, Gandhi, Loughran e McDonald (2019) em virtude de terem observado apenas o sentimento pessimista, apresentam a estatística descritiva dos seus dados somente para esse tipo de tom textual.

Em relação à capitalização, os bancos observados indicaram em média o valor de 0,104. Isso significa que essas companhias operam de maneira geral com certa alavancagem financeira, pois o patrimônio líquido representa 10,4% dos ativos totais dessas instituições; esse achado é corroborado ao se analisar a mediana que apresenta um valor de 0,095. Tais resultados são condizentes com o argumento de Vieira e Girão (2016), aos destacarem que devido à natureza das suas operações, os bancos tendem a utilizar principalmente o capital de terceiros para financiar as suas atividades e assim obterem lucros.

A rentabilidade dos bancos retratada por meio do retorno sobre os ativos ($ROA_{i,t}$) demonstra que essas empresas obtêm em média retornos de 0,3%, para cada trimestre. Por sua vez, considerando os valores extremos dessa variável, é perceptível que algumas companhias bancárias obtiveram retornos negativos de 0,5%, bem como uma rentabilidade máxima de 0,9%, que se configura como um percentual três vezes maior, do que a média trimestral dos demais bancos listados.

Considerando a variável $TAM_{i,t}$, uma vez que os seus valores correspondem ao logaritmo natural dos ativos totais de cada banco, é possível observar que em média as empresas bancárias analisadas, apresentaram o valor de 17, 61, e mediana de 16,86.

A variável PIB_t apresentou uma elevada variação ao longo do período analisado, indicando um desvio padrão de 0,818. Esse valor pode ser justificado devido ao aumento da

incerteza econômica e da instabilidade política no país, ocasionando resultados com alta volatilidade ao se analisar a variação percentual do Produto Interno Bruto real.

Por fim, a liquidez mensurada pela variável $LIQ_{i,t}$ apresenta, para os seus valores centrais, uma mediana de 0,382 e média de 0,40. Isso quer dizer que os depósitos totais representam em média 40% dos ativos totais desses bancos. Considerando a observação mínima de 0,125 e máxima de 0,667 são identificados comportamentos distintos para os bancos de capital aberto, pois algumas firmas operam com níveis de liquidez mais elevados em relação aos seus pares; já o percentual de 12,5% denota um nível aproximadamente três vezes abaixo da média praticada pelas demais companhias bancárias listadas.

4.2 Análise dos modelos

O critério para a escolha do modelo de previsão do risco de falência ocorreu seguindo as métricas apresentadas na seção 3.3. Especificamente, esse procedimento se fez necessário, uma vez que para a elaboração de modelos de predição, torna-se essencial o emprego de amostras distintas, etapa essa denominada de *cross-validation*, utilizada para evitar problemas de *over-fitting.* Esse último se caracteriza por uma adequação exacerbada do modelo aos dados, ocasionando estimações que apresentam bons desempenhos apenas para a amostra de treinamento, porém sem conseguir realizar previsões consistentes para dados fora desse subconjunto (BARBOSA, 2017; ROSA; GARTNER, 2018).

Nesse sentido, o algoritmo classificador *random forest* apresentou uma acurácia de 100%, em resultados não tabulados, durante a sua execução nos dados de treinamento, caracterizando uma elevada adequação desse método aos dados da subamostra e denotando um comportamento de *over-fitting*, fazendo com que os resultados obtidos na etapa de validação fossem desconsiderados para esse modelo.

Um dos fatores que contribuíram para o uso do algoritmo *random forest* neste estudo é o fato de que essa técnica apresenta um ajuste superior para dados não balanceados em relação à variável categórica dependente, em detrimento a outros métodos de aprendizagem de máquina entre os quais o *naive bayes*. Na Tabela 3 é descrita a acurácia obtida pelos modelos com os dados da subamostra de teste.

Tabela 3 – Desempenho de acurácia dos modelos

Modelo	Acurácia
Naive Bayes	0,815
Logit	0,844

Legenda: *Naive Bayes* corresponde ao método supervisionado de aprendizagem de máquina que utiliza redes bayesianas; *Logit* corresponde ao modelo clássico de regressão logística;
Fonte: Dados da pesquisa.

Para o cálculo do índice de acurácia foi realizada a comparação entre a classificação estabelecida pela previsão de cada modelo testado e a categorização presente *a priori* na base de dados. Dessa forma, foi observado que o modelo *naive bayes* apresentou uma acurácia em termos de predição de 81,5%, o modelo clássico *logit* obteve 84,4% de acurácia, demonstrando um melhor ajuste desse método para os dados do estudo.

Esse resultado permite concluir que o modelo logístico é mais robusto em comparação aos demais modelos testados, para realizar a predição de eventos de maior risco de insolvência em bancos de capital aberto, quando se utiliza a variável $\sigma ROA_{i,t}$ como uma *proxy* para categorizar tais comportamentos de interesse, sendo assim, mais consistente para a variável de risco de insolvência construída.

Dessa forma, diante dos resultados obtidos nesta etapa foi realizada a estimação do modelo *logit* para dados em painel, com o intuito de observar os sinais e o comportamento das variáveis independentes na explicação de eventos de maior risco de falência bancária.

4.3 Análise do painel e discussão dos resultados

Esta seção explana sobre os resultados da estimação da Equação 11, que considerou o modelo *logit* para dados em painel, com o intuito de mensurar se a variável de sentimento textual, que foi construída nesta pesquisa, bem como se as demais variáveis de controle podem explicar o risco de insolvência das companhias bancárias de capital aberto no Brasil.

Em relação ao modelo logístico em painel, Suss e Treitel (2019) destacam que caso a subamostra de treinamento não possua nas suas observações, uma representação de todas as companhias ou trimestres presentes na base de dados completa, durante a realização da etapa de previsão para fora dessa subamostra, o termo aleatório será igual a zero, para todas as firmas ou trimestres que não estão presentes no subgrupo de treinamento.

Ainda segundo os referidos autores, esse argumento é preponderante para que sejam aplicados modelos logísticos do tipo *pooled*, devido a especificidade desse tipo de estimação

em não considerar os efeitos aleatórios, e a necessidade de pesquisas acadêmicas em realizarem previsões para fora da subamostra de treinamento.

Nesse sentido, o presente estudo optou por utilizar o modelo *logit* do tipo *pooled*, uma vez que a determinação das observações que fazem parte das subamostras de treinamento e teste ocorreram de forma aleatória pelo próprio algoritmo, não podendo assegurar dessa maneira, que o subgrupo de treinamento possui dados referentes a todas os bancos ou trimestres presentes na base de dados completa. Assim, na Tabela 4 é apresentado o resultado da estimação do modelo *logit* do tipo *pooled*.

Tabela 4 – Resultados da estimação da equação 11 com o modelo *logit pooled*

Variável	**Coeficiente**	**Erro Padrão**	**P-Valor**
Intercepto	33,5572	7,1950	3,10e-06 ***
PIB	0,4084	0,2288	0,0742 .
ETS	-4,4887	5,7457	0,4347
ROA	-280,1271	46,6060	1,85e-09 ***
SB	5,6622	2,8365	0,0459 *
LIQ	1,4606	1,3253	0,2704
TAM	-2,1576	0,4077	1,21e-07 ***

Nível de Significância: ***: 0.1%, **:1%, *:5%, .: 10%.
Legenda: *PIB* indica a variação percentual do Produto Interno Bruto real brasileiro em comparação ao trimestre imediatamente anterior ao trimestre t; *ETS* indica o nível de capitalização do banco i, no período t, mensurada pela razão entre patrimônio líquido e ativo total; *ROA* indica a rentabilidade do banco i, no período t, mensurada pela razão entre lucro operacional e ativo total; *SB* indica o tom do sentimento textual bancário do documento i, no período t-1, mensurada pela ponderação dos pesos das palavras pessimistas e otimistas em cada relatório ITR ou DFP; *LIQ* indica a liquidez do banco i, no período t, mensurada pela razão entre depósitos totais e ativo total; *TAM* indica o tamanho do banco i, no período t, mensurada pelo logaritmo natural do ativo total.
Fonte: Dados da pesquisa.

A estimação evidenciada na Tabela 4 demonstra que a variável $\mathrm{PIB_t}$ apresentou uma relação positiva e significativa com a variável $\mathrm{Risco_{i,t}}$, tal achado é oposto ao encontrado por Vieira e Girão (2016), na maioria dos seus modelos, e por Ferreira, Zanini e Alves (2019) ao analisarem a relação entre esse indicador macroeconômico e o *Z-score* em bancos brasileiros.

Contudo, é importante destacar que os referidos estudos fizeram uso das técnicas de dados em painel de efeito fixo e o Método dos Momentos Generalizados (GMM) com e sem o uso de dados em painéis, bem como, empregaram outros *Z-score*, não se limitando a métrica desenvolvida por Boyd, Nicolò e Jalal (2006).

Adicionalmente, Ferreira, Zanini e Alves (2019) examinaram a relação entre o desvio padrão móvel do retorno sobre os ativos, calculado de maneira distinta a utilizada no presente

estudo, e a taxa de crescimento do PIB, não encontrando uma relação significativa entre essas duas variáveis.

Dessa forma, uma possível explicação para esse resultado, pode se dar ao considerar que as instituições bancárias em momentos de crescimento econômico, assumem uma postura de maior risco, aumentando a sua oferta de crédito e flexibilizando os critérios para a concessão de empréstimos, justificando assim, a relação encontrada entre as variáveis $\text{Risco}_{\text{i,t}}$ e PIB_{t}.

Por sua vez, a variável $Risco_{i,t}$ não apresentou uma relação significativa estatisticamente com as variáveis $ETS_{i,t}$ e $LIQ_{i,t}$, indicando que o risco de insolvência não é influenciado pela capitalização e pela liquidez das companhias. Tal evidência, para a variável de capitalização, é oposta a encontrada por Vieira e Girão (2016) e Ferreira, Zanini e Alves (2019) que identificaram uma relação positiva entre o *Z-score* e a capitalização dos bancos.

No entanto, Ferreira, Zanini e Alves (2019) ao analisarem a relação entre a capitalização e o desvio padrão móvel do retorno sobre os ativos, também não evidenciaram uma relação significativa. Ademais, Vieira *et al.* (2020) não identificaram uma relação significativa entre o *Z-score* e a capitalização.

Os referidos autores observaram ainda a relação entre a capitalização e o desvio padrão móvel do retorno sobre os ativos, evidenciando uma relação positiva e significativa, no entanto, esse resultado não se manteve no teste de robustez realizado por Vieira *et al.* (2020), ao adicionarem interações entre as variáveis de complexidade e *dummies* relacionadas ao tamanho dos conglomerados financeiros presentes no Brasil. Ressalta-se que Vieira *et al.* (2020) empregaram na sua pesquisa os métodos GMM e dados em painel de efeito fixo.

Uma possível explicação teórica para o resultado das variáveis $ETS_{i,t}$ e $LIQ_{i,t}$ pode se dar ao considerar a abordagem destacada por Vieira e Girão (2016), pois em sua pesquisa os autores argumentam que em teoria, os bancos de capital aberto apresentam uma menor concentração de propriedade, e por esse motivo tendem a obter valores superiores para o *Z-score*, em detrimento as instituições bancárias que apresentam uma concentração de capital mais elevada, e por conseguinte maior tendência para a assunção de riscos.

Dessa maneira, é possível considerar que os bancos presentes no mercado de capitais brasileiro operam com menores níveis de risco de insolvência, e consequentemente apresentam uma maior capitalização e liquidez, em relação as firmas bancárias não listadas, o que poderia explicar a não significância estatística obtida para essas duas variáveis, presentes nesse estudo.

O resultado da variável de sentimento textual indica que o risco de uma insolvência por parte dos bancos está associado ao sentimento textual otimista. Logo, instituições que

apresentam um modelo de gestão mais arriscado podem expressar em seus relatórios ITR e DFP uma maior predominância desse tipo de sentimento textual. Esse achado corrobora com Gupta, Simaan e Zaki (2018), ao defenderem que os bancos falidos manifestam sentimentos textuais mais otimistas em relação aos seus pares.

Conforme destacado pelos autores, esse fato pode indicar que ocorrem possíveis problemas de agência nessas companhias, uma vez que gestores, com o intuito de manter a confiança dos acionistas e investidores, podem tentar transmitir sinais positivos para essas partes interessadas, ou, ainda, esse comportamento também pode ser resultante de uma análise demasiadamente otimista por parte desses executivos, em relação ao desempenho futuro do banco; contudo, ambas as explicações não fazem parte do escopo deste estudo.

Esses resultados permitem concluir que o sentimento textual pode ser utilizado para explicar comportamentos de maior risco de insolvência bancária, por meio de informações de caráter qualitativo conforme defendido por Gupta, Simaan e Zaki (2018), Del Gaudio *et al.* (2019) e Gandhi, Loughran e McDonald (2019) contribuindo para a linha de pesquisa destinada à identificação de variáveis que explicam o aumento na probabilidade de insolvência para firmas bancárias.

A importância desse resultado está associada ao fato de os bancos serem consideradas organizações em que prevalece uma assimetria informacional entre essas instituições e os seus reguladores, bem como os demais *stakeholders*, em diversos mercados pelo mundo (DEL GAUDIO *et al.*, 2019; GANDHI; LOUGHRAN; MCDONALD, 2019; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

Assim, novas técnicas que utilizam informações qualitativas e permitem aprimorar a avaliação das condições bancárias ganham destaque, uma vez que podem reduzir essa disparidade de informações e melhorar o monitoramento dos bancos, sobretudo considerando a possibilidade de se mitigar um evento futuro de falência nesse setor, devido às consequências adversas que podem impactar na economia das nações (GUPTA; SIMAAN; ZAKI, 2018; DEL GAUDIO *et al*., 2019; GANDHI; LOUGHRAN; MCDONALD, 2019). As demais variáveis de controle apresentaram os sinais esperados que são reportados na seção 3.6.

5 CONSIDERAÇÕES FINAIS

Com o intuito de atender aos objetivos desse estudo foi empregada a técnica de aprendizagem de máquina não supervisionada *k-means,* para compreender o padrão de comportamento dos dados, e a partir deles estabelecer agrupamentos de acordo com o risco de insolvência dessas instituições, de modo semelhante à pesquisa de Viswanathan, Srinivasan e Hariharan (2020).

Dessa maneira foi observado que o algoritmo *k-means* ao fazer uso da variável $\sigma ROA_{i,t}$ obteve um percentual de similaridade de 50,82%, e 98,78%, respectivamente, para a categorização das companhias bancárias de maior e menor probabilidade de falência, ao comparar esse resultado com os valores indicados pela classificação realizada por meio do *Z-score*.

Posteriormente, foram empregados os métodos supervisionados de aprendizagem de máquina *naive bayes* e *random forest,* bem como o modelo *logit* para determinar qual dessas técnicas estatísticas é mais robusta para a predição da variável de risco de insolvência, elaborada por meio do algoritmo *k-means*.

Assim, foi observado que o modelo logístico apresentou a maior acurácia de previsão para os dados do estudo. De maneira complementar, em virtude dos objetivos desta pesquisa foi construída uma variável de sentimento textual para os bancos de capital aberto no Brasil, fazendo uso dos relatórios financeiros ITR e DFP dessas empresas, aplicando o algoritmo de leitura e o dicionário desenvolvidos por Silva e Machado (2019), para a estimação do tom textual desses documentos.

Os resultados indicam que o sentimento textual otimista consegue explicar comportamentos de maior risco de insolvência, para os bancos de capital aberto no Brasil, corroborando com os achados de Gupta, Simaan e Zaki (2018) ao analisarem companhias bancárias listadas nos Estados Unidos.

Esse achado sugere que os executivos dessas instituições financeiras apresentam uma análise otimista em relação ao desempenho futuro da firma em que atuam; também é possível supor que podem existir conflitos de agência nesses bancos, uma vez que são sinalizados aos investidores e às demais partes interessadas sentimentos textuais positivos, o que pode ser um indício de que os administradores fazem uso desses tons textuais para evitar uma possível perda de confiança desses *stakeholders* em relação ao empreendimento bancário.

Tais resultados contribuem para a atuação das entidades de supervisão do sistema financeiro de duas maneiras: a primeira consiste na indicação de uma nova medida para o risco

de falência dos bancos, já a segunda indica que informações qualitativas são relevantes na explicação desse comportamento de interesse, podendo auxiliar esses órgãos de controle nessa complexa atividade de monitoramento das instituições bancárias. De maneira complementar, esses achados podem ainda auxiliar investidores e proprietários de depósitos no processo de tomada de decisão sobre a alocação dos seus recursos, considerando as consequências sociais e econômicas que são geradas devido à falência de uma firma bancária.

Os resultados desta pesquisa se limitam aos bancos de capital aberto no Brasil e ao período estudado, uma vez que não foram observados outros tipos de companhias financeiras presentes na bolsa brasileira. Adicionalmente, também não fazem parte do escopo desta pesquisa empresas não financeiras, o que também se configura como uma limitação deste estudo.

De forma similar, é possível destacar o período amostral, o qual gerou 450 observações, devido à necessidade de se trabalhar com as médias móveis e com o desvio padrão móvel para o cálculo de algumas variáveis. Em virtude disso novos estudos podem observar períodos mais longos e se beneficiar com uma quantidade superior de observações disponíveis.

Pesquisas futuras podem ainda investigar a relação entre o sentimento textual e o risco de insolvência considerando empresas financeiras não bancárias, listadas na bolsa brasileira, com intuito de observar se o sentimento textual otimista também é mais preponderante para o risco de falência dessas firmas, ou se esse comportamento é específico para o setor bancário. Ademais, é possível, ainda, empregar outros métodos supervisionados de aprendizagem de máquina como o *support vector machine* (*SVM*) aplicado por Gupta, Simaan e Zaki (2018) e que apresentou uma elevada acurácia de previsão para comportamentos de falência dos bancos norte-americanos.

REFERÊNCIAS

ALTMAN, E. I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. **The Journal of Finance**, v. 23, n. 4, p. 589-609, 1968.

ALVES, K. L. F **Análise de sobrevivência de bancos privados no Brasil.** 2009. 83f. Dissertação (Mestrado em Engenharia de Produção) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2009.

ANDRADE, W. G. **A adoção de seguro de depósitos e seus impactos na intermediação financeira, na estabilidade bancária e no risco moral no Brasil**. 2020. 92 f. Dissertação (Mestrado em Ciências Contábeis) - Universidade de Brasília, Brasília, 2020.

ARAUJO, M. R. **Determinantes do posicionamento dos auditores sobre going concern em instituições financeiras em financial distress**. 2019. 109 f. il. Dissertação (Mestrado em Ciências Contábeis) - Universidade de Brasília, Brasília, 2019.

BARBOSA, J. H. F**. Early Warning System para distress bancário no Brasil**. 2017. xi, 186 f. il. Tese (Doutorado em Administração) - Universidade de Brasília, Brasília, 2017.

BEAVER, W. H. Financial ratios as predictors of failure. **Journal of accounting research**, p. 71-111, 1966.

BOUGHACI, D.; ALKHAWALDEH, A. A. K. Appropriate machine learning techniques for credit scoring and bankruptcy prediction in banking and finance: A comparative study**. Risk and Decision Analysis**, n. Preprint, p. 1-10, 2018.

BOYD, J. H.; GRAHAM, S. L. Risk, regulation, and bank holding company expansion into nonbanking. **Quarterly Review**, n. Spr, p. 2-17, 1986.

_____; DE NICOLÒ, G.; JALAL, A. M. Bank risk-taking and competition revisited: New theory and new evidence. **IMF Working Paper,** International Monetary Fund, 2006.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

CHIARAMONTE, L.; CROCI, E.; POLI, F. Should we trust the Z-score? Evidence from the European Banking Industry. **Global Finance Journal**, v. 28, p. 111-131, 2015.

_____ *et al*. How accurately can Z-score predict bank failure?. **Financial Markets, Institutions & Instruments**, v. 25, n. 5, p. 333-360, 2016.

CHISHOLM, E.; KOLDA, T. G. **New Term Weighting Formulas for the Vector Space Method in Information Retrieval**. Oak Ridge National Laboratory, Oak Ridge, TN, 1999.

CLIMENT, F.; MOMPARLER, A.; CARMONA, P. Anticipating bank distress in the Eurozone: An extreme gradient boosting approach. **Journal of Business Research**, v. 101, p. 885-896, 2019.

COUTO JUNIOR, C. G.; GALDI, F. C. Avaliação de empresas por múltiplos aplicados em empresas agrupadas com análise de cluster. **Revista de Administração Mackenzie**, v. 13, n. 5, p. 135-170, 2012.

DEL GAUDIO, B. L. *et al*. Mandatory disclosure tone and bank risk-taking: Evidence from Europe. **Economics Letters**, v. 186, p. 108531, 2019.

FERNANDES, F. T.; CHIAVEGATTO FILHO, A. D. P. Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. **Revista Brasileira de Saúde Ocupacional**, v. 44, 2019.

FERREIRA, J.; ZANINI, F.; ALVES, T. A Diversificação das Receitas Bancárias: Seu Impacto sobre o Risco e o Retorno dos Bancos Brasileiros. **Revista Contabilidade & Finanças - USP**, v. 30, n. 79, p. 91-106, 2019.

GALDI, F. C.; GONÇALVES, A. Pessimismo e incerteza das notícias e o comportamento dos investidores no Brasil. **Revista de Administração de Empresas**, v. 58, n. 2, p. 130-148, 2018.

GANDHI, P.; LOUGHRAN, T.; MCDONALD, B. Using annual report sentiment as a proxy for financial distress in US banks. **Journal of Behavioral Finance**, v. 20, n. 4, p. 424-436, 2019.

GUPTA, A.; SIMAAN, M.; ZAKI, M. When positive sentiment is not so positive: Textual analytics and bank failures. **SSRN Working Paper**. 2018. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2773939. Acesso em: 07 maio 2020.

HAN, J.; KAMBER, M.; PEI, J. Data mining: concepts and techniques, Waltham, MA. **Morgan Kaufman Publishers**, v. 10, p. 978-1, 2012.

JIANG, F. *et al*. Manager sentiment and stock returns. **Journal of Financial Economics**, v. 132, n. 1, p. 126-149, 2019.

KEARNEY, C.; LIU, S. Textual sentiment in finance: A survey of methods and models. **International Review of Financial Analysis**, v. 33, p. 171-185, 2014.

LEPETIT, L.; STROBEL, F. Bank insolvency risk and time-varying Z-score measures. **Journal of International Financial Markets, Institutions and Money**, v. 25, p. 73-87, 2013.

LIBERMAN, M.; BARBOSA, K.; PIRES, J. Falência Bancária e Capital Regulatório: Evidência para o Brasil. **Revista Brasileira de Economia**, v. 72, n. 1, p. 80-116, 2018.

LIU, B.; ZHANG, L. A survey of opinion mining and sentiment analysis. In: _____. **Mining text data**. Springer, Boston, MA, 2012. p. 415-463.

LOUGHRAN, T.; MCDONALD, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. **The Journal of Finance**, v. 66, n. 1, p. 35-65, 2011.

MACHADO, M. A. V.; SILVA, M. D. O. P. Análise do Sentimento Textual dos Relatórios de Desempenho Trimestral das Indústrias Brasileiras. **Sociedade, Contabilidade e Gestão**, v. 12, n. 1, p. 6-25, 2017.

PAGLIARUSSI, M. S.; AGUIAR, M. O.; GALDI, F. C. Sentiment Analysis in Annual Reports From Brazilian Companies Listed at the BM&FBOVESPA. **BASE - Revista de Administração e Contabilidade da UNISINOS**, v. 13, n. 1, p. 53-64, 2016.

RASCHKA, S. **Python Machine Learning**. 1. ed. Birmingham: Packt Publishing Ltd., 2015.

ROSA, P. S.; GARTNER, I. R. Financial distress em bancos brasileiros: um modelo de alerta antecipado. **Revista Contabilidade e Finanças**, v. 29, n. 77, p. 312-331, 2018.

SAVOY, J.; GAUSSIER, E. Handbook of Natural Language Processing: **Information Retrieval**. 2. ed. Boca Raton, Florida: CRC Press, Taylor and Francis Group, 2010.

SENA, B. H. S.; SILVA, C. A. T.; ARRIAL, R. T. Classificação do conteúdo de documentos contábeis usando aprendizagem de máquina: o caso dos fatos relevantes. **Revista de Educação e Pesquisa em Contabilidade**, v. 4, n. 2, p. 23-42, 2010.

SILVA, D. R. B.; RÊGO, T. G.; FRASCAROLI, B. F. **Sovereign risk ratings' country classification using machine learning**. Trabalho apresentado no XLVI Encontro Nacional de Economia. São Paulo, 2019.

SILVA, M. D. O. P.; MACHADO, M. A. V. Índice de Sentimento Textual: Uma Análise Empírica do Impacto das Notícias Sobre Risco Sistemático. **Revista Contemporânea de Contabilidade**, v. 16, n. 40, p. 24-42, 2019.

SILVA, P.; H.; N.; BESARRIA, C.; N.; SILVA, M. D. O. P. **Mensurando o Sentimento de Incerteza da Política Econômica**: Uma Análise a Partir da Comunicação do Banco Central do Brasil. Trabalho apresentado no XLVI Encontro Nacional de Economia. São Paulo, 2019.

SILVA, R. A.; RIBEIRO, E. M. S.; MATIAS, A. B. Aprendizagem estatística aplicada à previsão de default de crédito. **Revista de Finanças Aplicadas**, v. 7, n. 2, p. 1-19, 2016.

SOMBRA, T. R. *et al*. Utilização de redes bayesianas através do algoritmo naive bayes para classificação carcaças de ovinos. **Brazilian Journal of Development**, v. 6, n. 3, p. 10476-10498, 2020.

SUSS, J.; TREITEL, H. Predicting bank distress in the UK with machine learning. **Bank of England Working Paper**, p. 831, 2019.

VARELLA, J. L.; QUADRELLI, G. Redes Neurais e Análise de Potência. **Revista de Tecnologia Aplicada**, v. 6, n. 3, p. 33-45, 2017.

VIEIRA, C. A. M.; GIRÃO, L. F. A. P. Diversificação das receitas e risco de insolvência dos bancos brasileiros. **Revista de Contabilidade e Organizações**, v. 10, n. 28, p. 3-17, 2016.

_____ *et al*. Complexidade e Risco dos Conglomerados Financeiros Operantes no Brasil. **BASE - Revista de Administração e Contabilidade da UNISINOS**, v. 17, n. 2, p. 277-308, 2020.

VISWANATHAN, P. K.; SRINIVASAN, S.; HARIHARAN, N. Predicting Financial Health of Banks for Investor Guidance Using Machine Learning Algorithms. **Journal of Emerging Market Finance**, p.226-261, 2020.

WANKE, B. S. L. *et al*. **Aplicação do Classificador Naive Bayes Para Identificação de Falhas de um Manipulador Robótico**. Trabalho apresentado no VIII CONEM – Congresso Nacional de Engenharia Mecânica. Uberlândia, 2014.