UNIVERSIDADE FEDERAL DA PARAÍBA – UFPB
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS – CCSA
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO – PPGA
CURSO DE MESTRADO EM ADMINISTRAÇÃO – CMA

ALLISSON SILVA DOS SANTOS

PREVISÃO DE INSOLVÊNCIA CORPORATIVA:
UMA ANÁLISE DE EMPRESAS BRASILEIRAS DE CAPITAL ABERTO POR MEIO DE APRENDIZADO DE MÁQUINA

JOÃO PESSOA – PB
2021

ALLISSON SILVA DOS SANTOS

PREVISÃO DE INSOLVÊNCIA CORPORATIVA:
UMA ANÁLISE DE EMPRESAS BRASILEIRAS DE CAPITAL ABERTO POR MEIO DE APRENDIZADO DE MÁQUINA

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

Orientador: Prof. Dr. Cássio da Nóbrega Besarria

Coorientador: Prof. Dr. Márcio André Veras Machado

JOÃO PESSOA – PB
2021

Catalogação na publicação
Seção de Catalogação e Classificação

S237p Santos, Allisson Silva Dos.
Previsão de insolvência corporativa: uma análise de empresas brasileiras de capital aberto por meio de aprendizado de máquina / Allisson Silva Dos Santos. - João Pessoa, 2021.
73 f.

Orientação: Cássio da Nóbrega Besarria.
Coorientação: Márcio André Veras Machado.
Dissertação (Mestrado) - UFPB/CCSA/PPGA.

1. Risco de insolvência corporativa. 2. Aprendizado de máquina. 3. Métricas de desempenho. 4. Empresas brasileiras de capital aberto. I. Besarria, Cássio da Nóbrega. II. Machado, Márcio André Veras. III. Título.

UFPB/BC CDU 346.5

Elaborado por Gracilene Barbosa Figueiredo - CRB-15/794

ALLISSON SILVA DOS SANTOS

PREVISÃO DE INSOLVÊNCIA CORPORATIVA:
UMA ANÁLISE DE EMPRESAS BRASILEIRAS DE CAPITAL ABERTO POR MEIO DE APRENDIZADO DE MÁQUINA

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

Aprovada em: 09 de dezembro de 2021

Banca examinadora:

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

Prof. Dr. Antônio Vinicius Barros Barbosa
Examinador Externo – UFPB

Prof. Dr. Adriano Leal Bruni
Examinador Externo – UFBA

AGRADECIMENTOS

Gratidão! Esse é o momento de afirmar o quanto sou grato às pessoas que apoiaram minha jornada antes e durante a jornada deste Mestrado. Tenho certeza que essas ainda estarão comigo após essa jornada. Primeiramente, agradeço à minha mãe por ter me dado todo apoio maternal, principalmente nos momentos de maior tensão perante os desafios que o Mestrado promoveu. Ela presenciou quase todo meu percurso em nossa casa, já que as aulas ocorreram remotamente durante os dois anos de formação.

Agradeço à minha irmã, que me estressa e ao mesmo tempo me apoia. Ela que é responsável por aumentar minha estima dizendo que eu vou ser o membro rico da família (financeiramente). Em relação à riqueza de saberes, estou tentando ampliar a cada dia. Quem sabe em algum momento eu me torne um ser rico em saberes e financeiramente, mas até lá ela pode continuar tentando aumentar minha estima.

Gratidão a Felipe que esteve comigo do início ao fim, do processo seletivo de Mestrado até a defesa da dissertação. Cada aprovação nas etapas do processo seletivo se tornou motivo de comemoração pra gente. A escuta dele foi essencial em cada momento desafiante dessa trajetória. Espero ter o ouvido dele bem próximo a mim nos próximos anos, já que o fim desse Mestrado não é nem de perto o fim da minha trajetória acadêmica.

Agradeço aos companheiros que conheci no Mestrado e que espero levar para a minha vida toda, em especial a Gabrielle, Taciana, Anderson, Caritsa, Alex, Edna e Denner. Sem eles, o curso não teria graça. Foi muito bom compartilhar os momentos bons e difíceis com vocês. Espero que continuemos compartilhando as cenas dos próximos capítulos de nossas vidas e que as confraternizações que a gente não teve durante a pandemia venham em 2022, 2023, etc.

Eu não poderia esquecer do incentivo que eu recebi por parte dos professores do Instituto Federal da Paraíba, durante a graduação em Administração. Eles incentivaram a minha participação no Mestrado em alguns momentos com algumas frases: “Vai terminar a graduação quando? Você tem perfil acadêmico, vá para o Mestrado logo!”; “Corra para o Mestrado, viu!”. Em especial, aos professores Washington Medeiros, Alysson André Oliveira, Maria Luiza Santos, Fernanda Nóbrega, Odilon Saturnino e Rebeca Cordeiro.

Preciso enfatizar o apoio dado pelo CNPq – Conselho Nacional de Pesquisa, que disponibilizou uma bolsa de Mestrado para que eu pudesse, de maneira tranquila, finalizar o curso. Sem esse apoio dado pela instituição eu possivelmente teria dificuldades para estar entregando este manuscrito com o zelo que tanto merece.

Por último, gostaria de agradecer aos ótimos professores que conheci no PPGA da UFPB. Primeiramente, ao professor Cássio por sua paciência e dedicação no trato com os alunos. Quando eu estiver lecionando quero levar a sua calmaria pra sala de aula. Ao professor Márcio, pelos conselhos sobre a trajetória profissional que recebi nos momentos que tive oportunidade de ouvi-lo. Além disso, pela disponibilização de acesso ao DataCamp, que me ajudou a compreender o pouco que conheço sobre aprendizado de máquina e manuseio do R. Agradeço à professora Carol Kruta pela parceria que tivemos durante o estágio supervisionado (foi ensino e diversão ao mesmo tempo). Os demais professores e servidores da secretária se sintam abraçados, pois eu sou grato pelo cuidado que tiveram comigo e com os meus colegas de turma. Gratidão!

RESUMO

Este trabalho teve por objetivo analisar a efetividade de modelos de aprendizado de máquina na identificação do risco de insolvência corporativa, considerando o contexto do mercado acionário brasileiro, com dados do período de 2010 a 2020. Empresas de todos setores foram consideradas neste estudo e, de maneira particular, as instituições financeiras foram analisadas conjuntamente e separadamente, diante de suas especificidades. Considerando a média anual, 299 empresas brasileiras de capital aberto tiveram seus dados analisados pelo estudo. O risco de insolvência foi mensurado de duas formas: pelo modelo Z'' Score e pela utilização do modelo de aprendizado de máquina não supervisionado *K-means*. Os modelos de aprendizado supervisionado de máquina que foram testados são o *random forest*, *naive bayes*, *logit*, K-NN, SVM (linear, polinomial e de base radial), *bagging* e *boosting*. Para a análise efetuada com todas as empresas, o *random forest* apresentou os melhores resultados, com acurácia entre 90,94% e 95,44%. A partir do momento que a análise alcança apenas as instituições financeiras, o melhor modelo preditivo passou a ser o SVM polimonial, o *logit* ou o *boosting*, dependendo da métrica de desempenho. Se a métrica decisiva para escolha do melhor modelo for o AUC, o *random forest* demonstra diferenças marginais comparadas aos melhores modelos, sugerindo que pode ser utilizado em empresas financeiras sem perda de performance. Os resultados desse trabalho apoiam as decisões de investidores, que capturam informações para definir racionalmente os ativos mais vantajosos; de executivos, que se preocupam com a credibilidade de organizações; de instituições provedoras de crédito, que irão fornecer empréstimos e financiamentos com mais assertividade; de pesquisadores, que irão contribuir com melhorias para a predição do risco de insolvência; e demais *stakeholders*, que utilizarão dos achados de acordo com a necessidade informacional.

Palavras-chave: Risco de insolvência corporativa. Aprendizado de máquina. Métricas de desempenho. Empresas brasileiras de capital aberto.

ABSTRACT

This research aimed to analyze the effectiveness of machine learning models in predicting corporate bankruptcy, considering the context of the Brazilian stock market, with data from 2010 to 2020. I considered companies from all sectors. Moreover, I analyzed financial firms jointly and separately, given their specificities. Considering the annual average, 299 Brazilian publicly traded companies had their data analyzed by the study. I measured the risk of financial distress in two ways: using the Z'' Score model, and using the K-means, an unsupervised machine learning model. I tested the following supervised machine learning models: random forest, naive bayes, logit, K-NN, SVM (linear, polynomial and radial basis), bagging and boosting. When I carried out the analysis with all companies, the random forest showed the best results, with an accuracy between 90.94% and 95.44%. When I analyzed only financial firms, the best predictive models were as following: polymonial SVM, logit or boosting, depending on the performance metrics. If the decisive metric for choosing the best model is AUC, random forest shows marginal differences compared to the best models, which means that it can be used in financial firms without loss of performance. The results of this work support the decisions of investors, who capture information to rationally define the most advantageous assets; of executives, who are concerned with the credibility of organizations; of credit provider institutions, which will provide loans and financing with more assertiveness; of researchers, who will contribute with improvements for the prediction of the risk of insolvency; and other stakeholders, who will use the findings according to their informational needs.

Keywords: Corporate insolvency risk. Machine learning. Performance metrics. Publicly traded Brazilian companies.

LISTA DE GRÁFICOS

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização e problema de pesquisa

Desde a obra clássica de Altman (1968), o estudo de empresas que apresentam dificuldades financeiras tem repercutido na literatura com modelos preditivos de falência corporativa. Essas pesquisas de natureza acadêmica buscam alcançar uma melhor capacidade preditiva. Geralmente, os autores incluem o estado de falência como linha divisória para distinguir as firmas que são saudáveis das que são insolventes.

Há modelos que utilizam de indicadores macroeconômicos, de dados de demonstrações financeiras, ou ainda, de informações mercadológicas, como variáveis independentes. As principais abordagens preditivas, comparadas em trabalhos científicos são: análise discriminante, modelos *logit* e *probit*, inteligência artificial e *hazard models* (ROSA; GARTNER, 2017).

Com a recessão mundial motivada pela crise de crédito global, a temática que corresponde à dificuldade financeira de empresas intensificou sua relevância na academia e no ambiente mercadológico em 2009 (MSELMI; LAHIANI; HAMZA, 2017). Essa crise revelou deficiências relevantes no campo financeiro e no nível de insuficiência de regulamentações financeiras, resultando na inatividade de organizações (GADGIL, 2021). Nesse contexto, surge a necessidade de investigar os riscos envolvidos em operações empresariais em prol dos *stakeholders*.

Uma das nomenclaturas utilizadas para empresas que podem não estar financeiramente saudáveis é a insolvência corporativa. Considerada como o indicativo de que as cláusulas dos contratos estabelecidas com credores de uma empresa não são cumpridas ou são honradas com dificuldade, a situação de insolvência corporativa pode resultar no estágio de falência do negócio (BAE, 2012). O estado de insolvência de uma entidade é caracterizado pela insuficiência de recursos para a quitação de dívidas determinadas, líquidas e vencidas (VODA *et al.*, 2021). Quando esse estado é percebido, estratégias podem ser tomadas para que, minimamente, não haja piora da situação.

A gestão financeira possui um papel importante e aplicável dentro das organizações, pois reúne informações relevantes diante da existência de riscos. Quando a tomada de decisão recebe dados advindos da predição, o aparecimento de falência corporativa pode ser amenizado, pois é possível identificar os casos insolventes antes que as relações de capital caiam abaixo de

um nível crítico ou os relatórios apresentem números negativos (SUSS; TREITEL, 2019; VODA *et al.*, 2021).

O estado de insolvência corporativa apresenta efeitos variados e significativos sobre a economia. Com o número crescente de grandes empresas, as dificuldades financeiras podem perturbar bruscamente o ambiente financeiro global. Para evitar problemas como esse, existem estratégias corporativas que podem ser combinadas a uma análise rigorosa de dados qualitativos e quantitativos e, assim, alcançar a identificação dos riscos financeiros de uma organização.

Para instituições financeiras, explorar o risco de crédito a partir das cláusulas contratuais, tanto de pessoas jurídicas quanto de pessoas físicas, tem sido uma das preocupações mais importantes e desafiantes dos últimos anos (ORESKI; ORESKI, 2014; TELES *et al.*, 2020). Diante disso, a literatura tem se preocupado em utilizar métodos e em definir variáveis robustas para melhorar a previsão de dificuldade financeira em organizações (ARORA; SINGH, 2020; BARBOZA; KIMURA; ALTMAN, 2017; STUPP; FLACH; MATTOS, 2018; VODA *et al.*, 2021).

Com o avanço da tecnologia, as técnicas de inteligência artificial têm ganhado aderência pelos pesquisadores e estão ocupando o lugar de métodos estatísticos tradicionais, sendo soluções alternativas com resultados promissores (SHI; LI, 2019). Um subconjunto da inteligência artificial é denominado de aprendizado de máquina, e dispõe de esforços para desenvolver estratégias, métodos e algoritmos que permitem o aprendizado de sistemas computacionais, por meio de dados fornecidos, com a execução de tarefas de *design* e de realização de testes preditivos (TELES *et al.*, 2020).

Para prever riscos financeiros, diante de um grande volume de dados, pesquisas têm utilizado algumas técnicas de aprendizado de máquina, tais como: *Support Vector Machine* (SVM), *random forests*, redes neurais artificiais, processos gaussianos e aprendizado adaptativo (TELES *et al.*, 2020). A título de exemplo, Barboza, Kimura e Altman (2017) utilizaram dados de empresas norte-americanas e aplicaram os modelos SVM linear e radial, análise discriminante linear, *bagging*, *random forest, boosting*, redes neurais *e logit.* Os modelos de aprendizado de máquina, exceto o SVM linear, apresentaram desempenho melhor que os modelos tradicionais *logit* e análise discriminante linear, com aproximadamente 10% a mais de acurácia, e o *random forest* apresentou o melhor resultado. No contexto brasileiro, ainda existe uma carência de estudos que procurem respostas comparativas de modelos de aprendizado de máquina para a identificação do risco de insolvência.

Esta pesquisa utiliza de dez modelos de aprendizado de máquina, um não supervisionado e nove supervisionados, que podem ser cruciais na gestão e na tomada de

decisões. Ressalta-se que os métodos não eliminam o risco de insolvência, mas o identifica antes de aparecer. A partir da predição, os *stakeholders* poderão tomar decisões mais estratégicas.

Dessa forma, o presente estudo almeja responder a seguinte questão-problema: no contexto do mercado acionário brasileiro, quão efetivos são os modelos de aprendizado de máquina na identificação do risco de insolvência corporativa? A resposta para essa pergunta é encontrada com o apoio de dados de onze anos (2010 a 2020) de empresas brasileiras de capital aberto.

Além de responder esse questionamento, o estudo é capaz de definir as empresas com maior ou menor risco de insolvência e decidir quais são as variáveis indispensáveis para a execução de predições. Melhorias na capacidade de previsão do risco de insolvência podem estimular a ocorrência de maiores retornos e a minimização de resultados negativos para os *stakeholders*.

1.2 Objetivo geral

O propósito deste estudo é analisar a efetividade de modelos de aprendizado de máquina na identificação do risco de insolvência corporativa, considerando o contexto do mercado acionário brasileiro.

1.3 Objetivos específicos

a) Classificar as companhias brasileiras de capital aberto quanto ao seu risco de insolvência;

b) Identificar variáveis que interferem no risco de insolvência de empresas brasileiras de capital aberto;

c) Comparar o desempenho dos modelos de aprendizado de máquina por diferentes métricas.

1.4 Justificativa

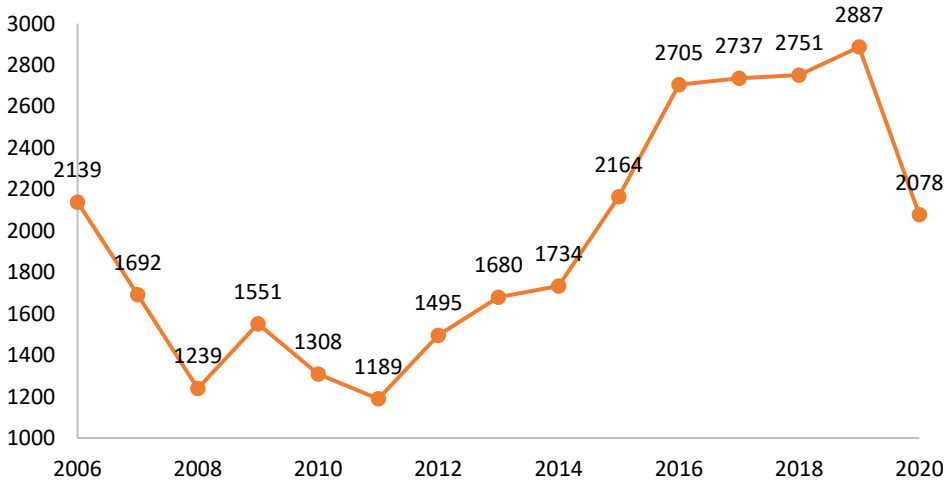
Em países emergentes, onde o ambiente empresarial possui como característica constante a instabilidade, as dificuldades financeiras são recorrentes (THINH *et al.*, 2020).

Porém, a literatura internacional tem se preocupado predominantemente em analisar determinantes de insolvência corporativa em mercados desenvolvidos (HSU; LI; FAN, 2006; NADEEM; SILVA; KAYANI, 2016).

Todos os anos, um quantitativo de organizações encerra suas atividades e decreta falência (KEYA *et al.*, 2021). De acordo com o levantamento de dados da Serasa Experian (2021), o número de empresas que decretaram falência e tiveram recuperações judiciais deferidas e concedidas no Brasil tem crescido de 2011 a 2019, conforme pode ser visualizado no Gráfico 1.

Um fato que chama a atenção no Gráfico 1 é que, no cenário pandêmico da Covid-19, em 2020, o número de organizações que decretaram falência ou com recuperações judiciais deferidas e concedidas no Brasil caiu expressivamente em comparação ao ano de 2019, com uma queda de 28,02%. A queda pode ser explicada pela facilitação de prazos, juros mais baixos e as novas linhas de crédito, que surgiram no momento difícil enfrentado pelos negócios durante a pandemia (SERASA EXPERIAN, 2021). Mesmo com esse decaimento, nos anos anteriores (2011 a 2019), o número de empresas com problemas dessa tipologia só aumentou. Assim, torna-se relevante estudar soluções e características de empresas brasileiras ligadas às dificuldades financeiras, para subsidiar as decisões corporativas em busca de melhores resultados.

Gráfico 1 - Falências decretadas e Recuperações Judiciais deferidas e concedidas no Brasil



Fonte: Serasa Experian (2021)

Quando uma organização se encontra em estado de insolvência, há uma alta possibilidade de falência, em virtude da incapacidade de cumprir o plano de recuperação financeira (VODA *et al.*, 2021). Ao considerar o alto risco incorporado no mercado de capitais,

prever a saúde financeira das organizações listadas em bolsas de valores é importante para governos, investidores, instituições financeiras e demais organizações, visto que toda parte interessada quer saber em qual negócio está envolvida (THINH *et al.*, 2020; ABDULLAH, 2021). Esta pesquisa é responsável por ampliar o corpo teórico sobre a efetividade preditiva de modelos de aprendizado de máquina, na busca da definição do comportamento econômico-financeiro de empresas em um cenário de mercado emergente.

Apesar de existirem modelos consolidados na literatura (ALTMAN, 1968; ALTMAN; HARTZEL; PECK, 1998; OHLSON; 1980), o padrão dos dados pode ser melhor definido por um algoritmo de aprendizado de máquina não supervisionado (VISWANATHAN; SRINIVASAN; HARIHARAN, 2020). Nesta pesquisa, além da utilização do método Z'' Score de Altman, Hartzel e Peck (1998), houve a aplicação do método de *clusterização K-Means*, para identificar dois grupos de nível de risco (alto e baixo) de insolvência corporativa. A partir da identificação dos *clusters*, a predição do risco de insolvência foi feita com modelos de aprendizado supervisionado. Ressalta-se que a combinação de técnicas de aprendizado de máquina não supervisionados e supervisionados garante a homogeneidade dentro de grupos, com a variância mínima, preenchendo lacunas em termos de robustez e heterogeneidade entre aglomerados.

Do ponto de vista prático, os resultados podem melhorar as análises de investidores, pesquisadores e executivos, ao validar modelos de aprendizado de máquina capazes de identificar o risco de insolvência corporativa das organizações brasileiras. Além disso, a predição do risco de insolvência pode apoiar a recuperação econômica de um negócio cuja continuidade esteja ameaçada. Uma melhoria na precisão da previsibilidade da dificuldade financeira é capaz de proporcionar maiores retornos e também minimizar os efeitos negativos para as partes interessadas (DUARTE; BARBOZA, 2020).

Os algoritmos de aprendizado de máquina podem servir para: (a) organizações credoras em busca de alavancar suas decisões de investimento; (b) formuladores de políticas que poderiam reconhecer e examinar suas decisões estratégicas; (c) investidor de nível individual que deseja adquirir racionalmente ativos financeiros; e (d) empresa fornecedora de recursos, que poderia solicitar pagamento antecipado às empresas insolventes (ABDULLAH, 2021). Nesse sentido, uma diretriz elaborada para avaliar o risco de insolvência financeira é sempre bem-vinda.

Ademais, incorporar as informações mais relevantes da análise de indicadores econômico-financeiros e desenvolver modelos para avaliar a insolvência corporativa podem ser consideradas maneiras descomplicadas de prever dados de futuras demonstrações financeiras

(VODA *et al.*, 2021). Dessa forma, os *stakeholders* terão subsídios para tomar melhores decisões, reduzir os riscos inerentes do negócio e aumentar o valor da empresa.

1.5 Estrutura do trabalho

Este trabalho científico encontra-se estruturado em cinco seções, além desta introdução. A seção seguinte evidencia a revisão da literatura sobre risco de insolvência corporativa. A terceira seção refere-se aos procedimentos metodológicos da pesquisa, que consiste em apresentar: (a) a mensuração do risco de insolvência corporativa; (b) as variáveis preditivas; (c) as técnicas de aprendizado de máquina supervisionado; e (d) os procedimentos para coleta de dados. A quarta seção corresponde à análise e discussão dos resultados do estudo. A última seção refere-se às considerações finais, que dispõe de limitações e sugestões para novas pesquisas.

2 REVISÃO DA LITERATURA

2.1 Insolvência Corporativa

Mudanças na economia global e nacional, surgimento de novas abordagens legislativas e diferentes tipos de organizações com características específicas demonstram a necessidade de revisitar estudos e indicadores debatidos de mensuração do risco de insolvência corporativa. Afinal, a discussão sobre causas de dificuldades financeiras e da incapacidade de pagamento de dívida empresarial nunca deixa de ser relevante, visto que corporações entram em processo de falência a todo momento (TARAN, 2017).

As dificuldades financeiras são compreendidas como condições pelo qual empresas enfrentam problemas financeiros unidos a restrições de liquidez (LESÁKOVÁ; GUNDOVÁ; VINCZEOVÁ, 2020). Essas dificuldades não prejudicam somente a sustentabilidade operacional de uma organização e os direitos e interesses de seus *stakeholders*, mas também podem influenciar na sociedade e na economia nacional (JAN; 2021).

A insolvência corporativa pode ser entendida como o indicativo de que as promessas feitas aos credores de uma organização são deixadas de lado ou são honradas com dificuldade, podendo chegar ao estágio de falência do negócio. Esse indicativo de dificuldade financeira pode gerar implicações, tais como: redução drástica do valor de mercado; fornecedores só aceitarem pagamento à vista; e a possibilidade de o cliente cancelar suas solicitações na expectativa de não conseguir usufruir de produtos e/ou serviços dentro das condições negociadas de qualidade e prazo (BAE; 2012).

A insolvência corporativa é um campo de estudo que tem ganhado maior importância desde a recessão mundial motivada pela crise de crédito global em 2009. Esse campo de estudo discute a situação em que os pagamentos acordados contratualmente não são supridos pelo histórico de atividades da empresa. Essa situação é capaz de ilustrar o risco de crédito que os financiadores possuem com empresas em dificuldade financeira (MSELMI; LAHIANI; HAMZA, 2017).

Apesar de não resultar necessariamente em falência das organizações, a insolvência corporativa indica que um desempenho financeiro persistentemente fraco afeta negativamente o patrimônio dos investidores (HABIB *et al.*, 2018). Dessa forma, as partes interessadas no negócio precisam reter a atenção para indicadores que possam transparecer sinais de perigo no desempenho financeiro e, assim, pensar em ações para proteger seus interesses.

Para prever a condição de problemas financeiros, modelos estão surgindo, com o intuito de apoiar a gerência na mitigação de riscos e tomadas de decisão (LESÁKOVÁ; GUNDOVÁ; VINCZEOVÁ, 2020). Os modelos de dificuldade financeira geralmente são construídos com base em dados das demonstrações financeiras, indicando condições financeiras positivas ou negativas (MUÑOZ-IZQUIERDO *et al.*, 2020).

Estratégias corporativas podem ser combinadas a uma análise rigorosa de dados qualitativos e quantitativos, para alcançar a identificação dos riscos financeiros de uma organização (TELES *et al.*, 2020). Arora e Singh (2020) identificaram que dados não estruturados, como notícias, em conjunto com os dados estruturados, como índices financeiros, podem desempenhar um papel importante na previsão de dificuldades financeiras.

A previsão de risco de insolvência corporativa pode apoiar a recuperação econômica de uma empresa cuja continuidade operacional encontra-se ameaçada. As causas para a insuficiência financeira de uma empresa são as mais diversas, podendo ser tanto de natureza endógena quanto exógena. Um diagnóstico contábil-financeiro para apuração do estado de insolvência, com indicadores de liquidez e de solvência financeira, é importante para reconhecer o estado do patrimônio de uma entidade caracterizada por insuficiência de recursos (VODA *et al.*, 2021).

Autores pioneiros na elaboração de métodos de previsão de insolvência podem ser citados, como Altman (1968), que utiliza as seguintes variáveis: capital de giro sobre ativos totais, lucros retidos sobre ativos totais, *Earnings Before Interest and Taxes* (EBIT) sobre ativos totais, valor de mercado do patrimônio líquido sobre a dívida total e vendas sobre ativos totais; e Ohlson (1980), que incentiva a utilização da regressão logística e descobre que a baixa lucratividade, a baixa liquidez e a alta dívida ampliam a possibilidade de inadimplência empresarial.

Com a quarta revolução industrial, as tecnologias de inteligência artificial e *big data* aprimoraram a análise de insolvência corporativa. Instituições financeiras, a fim de garantir estabilidade na concessão de crédito e de minimizar os riscos de dificuldade financeira, estão cada vez mais utilizando técnicas preditivas, a partir de modelos de aprendizado de máquina. Esses modelos preditivos estão sendo avaliados por métricas que envolvem a quantidade de acertos na predição (JIN *et al.*, 2019; SHI; LI, 2019).

O motivo pelo qual a previsão de insolvência corporativa é relevante está na possibilidade de mudar seu estado, visto que, quando uma entidade se encontra insolvente, a probabilidade de falência é extremamente alta. Na avaliação de empresas, é preciso considerar a interdependência entre os números de demonstrações contábeis. A partir de informações

preditivas, gestores podem trabalhar com planos de recuperação financeira, com o intuito de retornar ao estado solvente (VODA *et al.*, 2021). Diante da evolução da inteligência artificial, estudos têm utilizado de técnicas de aprendizado de máquina, como: *support vector machine*, *random forests*, redes neurais artificiais, processos gaussianos e aprendizado adaptativo (TELES *et al*., 2020).

Em casos de dificuldades financeiras nas organizações, três estágios normalmente podem ser esperados: (a) queda acentuada do fluxo de caixa; (b) escassez de capital e (c) a ocorrência de insolvência corporativa. Se os problemas persistirem sem solução durante o momento inicial de dificuldade financeira, as organizações poderão entrar em processo de liquidação, ir à falência ou ser adquiridas por terceiros (JAN; 2021).

Problemas financeiros possuem custos que podem exceder o valor empresarial, fazendo com que as empresas sejam dissolvidas. Uma alternativa para enfrentamento de custos é adotar um planejamento de reorganização, selecionando opções de financiamento com menor custo. É possível que as organizações levantem dinheiro de investidores externos e se reestruturem reorganizando as dívidas e vendendo ativos. As companhias podem ser consideradas com dificuldades financeiras ou em estado de falência quando: (a) a empresa se encontra em situação de falência decretada ou com pedido de recuperação judicial deferido; (b) o *Earnings Before Interest, Taxes, Depreciation and Amortization* (EBITDA) é inferior às despesas financeiras por dois anos consecutivos; e (c) o Patrimônio Líquido (PL) da empresa é negativo por dois anos consecutivos (MANZANEQUE; MERINO; PRIEGO, 2016).

O cenário mercadológico é um aspecto que ganha relevância na avaliação situacional de companhias. Conforme o ambiente de negócios se renova e perpassa por novos dilemas, novos fatores podem determinar o risco de insolvência corporativa. Para gerir os possíveis transtornos na saúde financeira das empresas e buscar o desenvolvimento, é importante observar os sistemas econômicos, oportunidades de financiamento, políticas fiscais, condições de trabalho, mercados, legislação e culturas nacionais (TARAN, 2017).

Em países emergentes, onde o ambiente empresarial possui como característica constante instabilidade, as dificuldades financeiras são recorrentes. Essas dificuldades prejudicam a economia e a sociedade de um país como um todo. Nesse sentido, prever a saúde financeira das organizações listadas em bolsas de valores é importante para governos e investidores, dado o alto risco incorporado no mercado de capitais (THINH *et al.*, 2020).

A predição de insolvência pode ajudar a evitar ou reduzir perdas financeiras por parte das empresas. Essas perdas dão origem a custos que afetam diretamente os funcionários, gerentes, acionistas e fornecedores. Numa análise aprofundada, empresas com alto risco de

insolvência podem afetar a sociedade em geral, quando os impostos arrecadados dessas organizações são utilizados para o bem-estar geral da população (STUPP; FLACH; MATTOS, 2018).

2.2 Evidências empíricas da previsão de insolvência corporativa

Evidências empíricas têm demonstrado a contribuição de variáveis financeiras e não financeiras para a previsão do risco de insolvência corporativa (BARBOZA; KIMURA; ALTMAN, 2017; KEYA *et al*., 2021; STUPP; FLACH; MATTOS, 2018). Esta subseção é responsável por relatar objetivos, resultados e conclusões de pesquisas com problemáticas similares a este estudo.

Uma pesquisa interessada em modelos de previsão de dificuldades financeiras foi realizada pelos autores Hsu, Li e Fan (2006), com uma amostra de 62 empresas e 248 observações, sendo 31 não falidas e 31 com dificuldades financeiras, assim classificadas pelo Taiwan Economic Journal. Para a realidade de Taiwan, a dificuldade financeira pode ser prevista com precisão de até 91,53%, diminuindo após o primeiro ano da ocorrência de dificuldade. Os autores utilizaram de redes neurais e de regressão logística. As redes neurais são sistemas artificiais que simulam pensamentos de pessoas, para identificar a relação entre entrada e saída por meio do processo de aprendizagem. Já a regressão logística considera a probabilidade condicional de determinado evento está entre 0 e 1.

Hsu, Li e Fan (2006) descobriram que, no estágio inicial, a lucratividade é o fator mais importante, e que as empresas com pouca lucratividade podem apresentar problemas financeiros no curto prazo. Porém, no longo prazo, a gestão da operação (como contas a receber) e o capital intelectual (como patentes e despesas com P&D) têm um impacto significativo na situação financeira de uma empresa.

Geng, Bose e Chen (2015) estudaram o fenômeno da crise financeira para 107 empresas da China que receberam o rótulo de 'tratamento especial' de 2001 a 2008 pela Bolsa de Valores de Xangai e de Shenzhen. As técnicas de mineração de dados construíram modelos de alerta para crises financeiras, com base em 31 indicadores financeiros. Os modelos de aprendizado de máquina testados por esses autores foram: árvores de decisão, redes neurais, SVM e múltiplos classificadores baseados em *majority voting*.

As árvores de decisão geralmente são utilizadas para investigar recursos e identificar padrões, e estabelecem nós entre a raiz (base de dados) e as folhas (valor gerado como resposta) que se relacionam por uma hierarquia. O SVM é um modelo que busca a adequação de uma

linha ou de um hiperplano, entre as diferentes classes, com o intuito de maximizar a distância até alcançar os pontos das classes. Já os múltiplos classificadores baseados em *majority voting* consideram os resultados de cada classificador em um grupo de classificadores, para gerar maior estabilidade na escolha final. Entre esses modelos, as redes neurais apresentaram melhor desempenho e as variáveis financeiras relacionadas à lucratividade demonstraram ser cruciais para as análises (GENG; BOSE; CHEN, 2015).

Barboza, Kimura e Altman (2017) utilizaram dados de empresas norte-americanas que obedecem ao período temporal de 1985 a 2013 e aplicaram os modelos SVM linear e radial, análise discriminante linear, *bagging, random forest, boosting*, redes neurais *e logit.* A análise discriminante linear é um método que possui a propriedade de distância quadrada simétrica e se preocupa com a classificação, redução de dimensão e visualização de dados, removendo características dependentes e redundantes ao reduzir o espaço dimensional. O *bagging* se trata da agregação de *bootstrapping*, pelo qual há a elaboração de novos subconjuntos aleatórios de dados por meio de amostragem, com substituição, gerando estimativas de intervalo de confiança (BREIMAN, 1996). O *random forest* também utiliza de *bootstrapping* por meio de múltiplas árvores e subamostras, reduzindo a variação que existe na aplicação de árvores de decisão (SCHONLAU; ZOU, 2020). O *boosting* utiliza da ideia de adicionar novos modelos ao conjunto, com modelos fracos sendo treinados, buscando velocidade, melhoria da precisão e diminuição de chances de ocorrer *overfitting* (FRIEDMAN, 2001).

Os modelos de aprendizado de máquina, exceto o SVM linear, apresentaram desempenho melhor que os modelos *logit* e análise discriminante linear, com aproximadamente 10% a mais de acurácia, em média. Os atributos utilizados foram as variáveis definidas pelo modelo clássico de Altman (1968), margem operacional, mudança no ROE (*Return on Equity*), mudança no *price-to-book*, crescimento em número de empregados, crescimento em vendas e crescimento de ativos (BARBOZA; KIMURA; ALTMAN, 2017).

Na realidade das empresas listadas na *Bourse Istanbul* (BIST), foi realizado um estudo para os modelos de avaliação de risco construídos por meio de indicadores financeiros. Foram utilizadas cinco técnicas preditivas para verificar a precisão das previsões: SVM, árvores de decisão, redes neurais artificiais com sistemas de inferência neuro-fuzzy adaptativos (ANFIS), análise discriminante e K-NN (K - *Nearest Neighbors*). A classificação K-NN consiste em decidir o valor K adequado e estabelecer a função de distância para identificar K vizinhos mais próximos. Essa classificação é realizada por meio da distância entre os dados nas amostras de treinamento (ZHANG, 2019). Os resultados indicaram que o modelo de árvores de decisão alcançou a melhor acurácia nas predições (ERDOGAN; KONAKLI, 2018).

Para o contexto brasileiro, Stupp, Flach e Mattos (2018) analisaram a influência da adoção das normas internacionais de contabilidade na predição de insolvência corporativa. Os autores dividiram 94 empresas em dois grupos: empresas solventes e insolventes. Os dados coletados para a realização da pesquisa compreendem o período de 31 de dezembro de 2004 a 31 de dezembro de 2013, e com eles foram definidas 29 variáveis preditivas. O método utilizado para a categorização das observações foi a de análise discriminante. O principal resultado desta pesquisa é que houve uma melhoria considerável na previsão de insolvência após a chegada das normas internacionais de contabilidade no Brasil, pois a média de acerto aumentou de 73,5% para 82,1%.

Ciaccio e Cialone (2019) realizaram uma pesquisa, com o intuito de desenvolver um modelo de previsão de insolvência de empresas comerciais da Itália, sendo aplicada em pequenas, médias e grandes empresas. Os resultados demonstraram que os modelos utilizados pelos autores, como *logit*, *gradient boosting* e *random forest* apresentaram bons valores de precisão, sensibilidade e AUC. O modelo com uma melhor performance foi o *gradient boosting*, com o maior AUC (98,1%). Essa pesquisa ainda fez comparações com os resultados de Barboza, Kimura e Altman (2017), demonstrando que os valores de AUC são maiores para os modelos que utilizaram os dados de empresas da Itália. Porém, ressalta-se que os dados utilizados por essas duas pesquisas são diferentes, o que pode influenciar nas medidas dos modelos.

Diante da realidade empresarial do Reino Unido, um estudo foi efetuado para desenvolver um sistema de alerta antecipado de angústia bancária por meio de aprendizado de máquina. Os modelos utilizados foram *logit*, K-NN, *boosting*, *random forest* e SVM. Foi implementado um procedimento de multivalidação randomizado de duplo bloco para avaliar o desempenho fora da amostra. Os resultados indicaram que o modelo *random forest* apresentou o melhor desempenho (SUSS; TREITEL, 2019).

Teles *et al.* (2020) utilizaram o embasamento sobre insolvência corporativa para explorar a gestão do risco de crédito com dois modelos de aprendizado de máquina. Os resultados que foram obtidos pelos autores indicaram que o algoritmo *random forest* é promissor para uso na gestão de risco de crédito, e as vantagens da abordagem *random forest* diante do algoritmo SVM estão em sua velocidade e simplicidade operacional. Em contrapartida, o modelo SVM apresentou o benefício de uma maior precisão preditiva.

Petropoulos *et al.* (2020) empregaram uma série de técnicas de modelagem para prever insolvência em uma amostra de instituições financeiras dos Estados Unidos. As técnicas utilizadas foram: análise discriminante, *logit*, redes neurais, *random forest* e SVM. Os

resultados demonstraram que o método *random forest* teve um bom desempenho preditivo e, logo em seguida, o método de redes neurais. Além disso, os autores evidenciaram que indicadores relacionados a lucro e capital constituem os atributos com maior contribuição marginal para a previsão de insolvência bancária.

Viswanathan, Srinivasan e Hariharan (2020) realizaram uma pesquisa que classifica os bancos com base em sua saúde financeira, ajudando a identificar bancos com perfis de risco mais elevados. Os autores utilizaram dados de 44 bancos indianos em categorias distintas de saúde financeira no espaço temporal de 12 anos (2005-2017). Foram utilizadas três técnicas para verificar a acurácia das previsões: análise discriminante, *classification and regression tree* (CART) e *random forest*. O modelo CART se trata de um método binário de árvores de decisão. A acurácia para os modelos análise discriminante e *random forest* apresentou valores similares: 95,36% e 95,93%, respectivamente. Como a diferença é marginal, esses dois modelos podem ser elencados como os melhores e recomendados para a realização de predições de insolvência corporativa.

Arora e Singh (2020) estudaram a predição da falência de empresas dos Estados Unidos usando índices financeiros e dados de notícias. Os dados para realização das análises foram extraídos dos relatórios financeiros organizacionais, dos jornais *online*, de relatórios e de artigos encontrados a partir do Google News. As notícias foram elencadas como pessimistas e otimistas e, juntamente com os índices financeiros das organizações, foram utilizadas na aplicação de três modelos de aprendizado de máquina: SVM, *random forest* e *logit*. A métrica de acurácia apontou o *random forest* como modelo com maior capacidade preditiva, com 90% de acerto na classificação da amostra.

Para os dados da Indonésia, Rahayu e Suhartanto (2020) testaram técnicas de aprendizado de máquina e verificaram quais delas possuem um melhor desempenho de previsão para empresas não saudáveis financeiramente. Os modelos utilizados pelos autores foram: *random forest*, *AdaBoost*, *logit*, análise discriminante linear, análise discriminante quadrática, *naive bayes*, SVM linear, SVM de base radial, árvores de decisão, K-NN e redes neurais. O *AdaBoost* é um método adaptável, pelo qual as classificações em sequência são ajustadas para favorecer as instâncias classificadas negativamente por classificações anteriores. O método classificador *naive bayes* é capaz de prever probabilidades *e* considera que a ausência ou presença de um determinado atributo de uma classe não está relacionada à ausência ou presença de qualquer outra característica quando a variável de classe é definida. (JADHAV; CHANNE, 2016). Os resultados dos modelos *random forest* e *AdaBoost* são similares, revelando melhor

desempenho pelas métricas (acurácia, precisão, *recall* e f-1) definidas para fins de avaliação dos modelos.

Keya *et al.* (2021) utilizaram de alguns modelos de aprendizado de máquina, tais como: *AdaBoost*, árvores de decisão, *bagging* e *random forest*, para detectar a falência entre organizações da Polônia. A precisão entre os modelos variou entre 94 a 97%. A maior precisão apresentada foi a do modelo *bagging*. Os autores utilizaram o total de 64 atributos para realizar as predições.

Para empresas listadas em Bangladesh, Abdullah (2021) verificou se o aprendizado de máquina pode ser implementado na previsão de insolvência financeira. O autor utilizou os modelos denominados como redes neurais, SVM, *naive bayes*, K-NN e *ensemble*. O método *ensemble* combina múltiplos algoritmos de aprendizagem para garantir um desempenho superior a métodos singulares. Foram utilizadas as seguintes métricas para avaliação dos modelos: perda logarítmica (logLoss), AUC, *recall* de precisão AUC (prAUC), precisão, kappa, sensibilidade e especificidade. O modelo de rede neural artificial apresentou 88% de precisão e de taxa de sensibilidade e uma AUC de 96%. Esses resultados foram os mais expressivos entre os modelos para essas três métricas de desempenho. Para a métrica logLoss, o melhor resultado foi o do método *ensemble*.

Sehgal *et al.* (2021) realizaram um estudo com o objetivo de identificar determinantes microeconômicos críticos de dificuldades financeiras e projetar um modelo parcimonioso de previsão de dificuldades financeiras para uma economia emergente. Os autores utilizaram dos modelos *logit*, rede neural artificial e SVM, para comparar a taxa de acurácia de previsão de insolvência corporativa de técnicas alternativas para o setor empresarial da Índia. Os resultados da pesquisa indicaram variáveis altamente significativas para a previsão: retorno sobre o capital empregado, fluxos de caixa para o passivo total, índice de giro de ativos, ativos fixos para ativos totais, índice de dívida para patrimônio líquido e uma medida do tamanho da empresa. Os modelos SVM e rede neural artificial demonstraram melhor desempenho do que o modelo *logit*.

Nessa perspectiva, considerar um conjunto de atributos na previsão do risco de insolvência corporativa e validar modelos de aprendizado de máquina são ações que podem apoiar a alocação eficiente de recursos financeiros empresariais. Diante dos estudos citados nesta subseção, para facilitar a visualização dos modelos preditivos utilizados pelos diversos autores, foi elaborado o Quadro 1.

Quadro 1 – Melhores modelos identificados pelas pesquisas

Autores	**Técnicas**										**Melhor(es) Modelo(s)**
	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**	
Hsu, Li e Fan (2006)			X							X	Para previsões com até dois anos antes, o modelo de redes neurais apresentou melhor resultado. Para previsões com três anos antes, o *logit* se mostrou melhor.
Geng, Bose e Chen (2015)					X					X	Redes neurais
Barboza, Kimura e Altman (2017)	X		X		X		X	X	X	X	*Random forest*, *bagging* e *boosting* apresentaram os melhores resultados
Erdogan e Konakli (2018)				X	X					X	Árvores de decisão
Stupp, Flach e Mattos (2018)										X	Análise discriminante
Ciaccio e Cialone (2019)	X		X						X	X	*Gradient boosting*
Suss e Treitel (2019)	X		X	X	X				X		*Random forest*
Teles *et al.* (2020)	X				X						O *random forest* possui vantagens em velocidade e simplicidade operacional. O SVM apresentou o benefício de uma maior precisão preditiva.
Petropoulos *et al*. (2020)	X		X		X					X	*Random forest*
Viswanathan, Srinivasan e Hariharan (2020)	X									X	Análise discriminante e *random forest* possuem marginalmente a mesma contribuição
Arora e Singh (2020)	X		X		X						*Random forest*
Rahayu e Suhartanto (2020)	X	X	X	X	X		X		X	X	*Random forest* e *AdaBoost*
Keya *et al.* (2021)	X							X	X	X	*Bagging*
Abdullah (2021)		X		X	X					X	Redes neurais
Sehgal *et al.* (2021)		X			X					X	SVM e redes neurais

Legenda: M1 representa o *random forest*, M2 se refere ao *naive bayes*, M3 representa o *logit*, M4 se refere ao K-NN, M5 representa o SVM linear, M6 representa o SVM polinomial, M7 se refere ao SVM de base radial, M8 representa o *bagging*, M9 se refere ao *boosting* e M10 representa outro(s) modelo(s).

Fonte: Elaborado pelo autor (2021)

3 PROCEDIMENTOS METODOLÓGICOS

3.1 Mensuração do risco de insolvência corporativa

A mensuração do risco de insolvência corporativa foi efetuada de duas formas: a primeira, com a aplicação do modelo Z'' Score de Altman, Hartzel e Peck (1998), desenvolvido para empresas localizadas em países emergentes; e a segunda, com a utilização do modelo de aprendizado de máquina não supervisionado *K-means.*

3.1.1 Mensuração do Z'' Score de Altman, Hartzel e Peck (1998)

Para mensurar o risco de insolvência das empresas, primeiramente, foi considerado o modelo Z'' Score de Altman, Hartzel e Peck (1998). O modelo elaborado por esses autores é adequado para empresas que operam em países emergentes (SHAHWAN; HABIB, 2020), adequando-se, assim, ao mercado acionário brasileiro. Para a criação do modelo, Altman, Hartzel e Peck (1998) consideraram um sistema de pontos que se baseia em uma revisão financeira fundamental advinda de um modelo de risco quantitativo e de avaliações sobre riscos de crédito específicos. Diante de suas conclusões, o valor da pontuação Z'' passou a ser estimado conforme Equação 1:

$$Z''Score_{it} = 3{,}25 + 6{,}56\,X_{1_{it}} + 3{,}26\,X_{2_{it}} + 6{,}72\,X_{3_{it}} + 1{,}05\,X_{4_{it}} \qquad (1)$$

Em que: *Z'' Score* representa o risco de insolvência da empresa *i* no momento *t*; X_1 representa o capital de giro sobre o total de ativos da empresa *i* no momento *t*; X_2 representa os lucros acumulados sobre o total de ativos da empresa *i* no momento *t*; X_3 representa os lucros antes de juros e impostos sobre o total de ativos da empresa *i* no momento *t*; e X_4 representa o patrimônio líquido sobre o valor contábil da dívida total da empresa *i* no momento *t.*

O método Z'' Score de Altman, Hartzel e Peck (1998) surgiu a partir do modelo criado por Altman (1968), pelo qual foi verificado que não atendia às necessidades de empresas localizadas em países emergentes, apresentando precisão e confiabilidade não satisfatórias. Diante disso, o modelo de Altman, Hartzel e Peck (1998) considera ajustes para o tipo de indústria e posição competitiva das empresas públicas e privadas e apresenta resultados

robustos para mercados emergentes. Os dois modelos elaborados, em 1968 e 1998, foram desenvolvidos a partir da análise discriminante multivariada. As variáveis que estão presentes no modelo foram escolhidas com base na relevância da literatura da época.

Para testar a viabilidade do modelo, Altman (1968) utiliza de testes como o "F", para alcançar a capacidade discriminante individual das variáveis, e valores de erro tipo 1 e tipo 2, para alcançar a precisão do modelo. Altman (2005) relata que aplicou, desde a década de 90, o modelo para organizações de mercados emergentes, e os resultados apresentados foram robustos para países como Brasil e Argentina e para muitos dos países do Sudeste Asiático.

Atiya (2001) elucida a importância de cada uma das variáveis presentes no modelo de Altman, Hartzel e Peck (1998). Primeiramente, o total de ativos sugere o tamanho da organização. Em seguida, o capital de giro, representado pelo ativo circulante menos o passivo circulante, pode ser visto como uma indicação da capacidade da empresa de pagar suas dívidas de curto prazo. Caso o valor do capital de giro seja muito negativo, a empresa pode ficar inadimplente em alguns pagamentos.

Os lucros acumulados significam o acúmulo dos ganhos desde o início da empresa. Os lucros antes de juros e impostos significam que lucros altamente negativos (positivos) indicam que a empresa está perdendo (ganhando) competitividade, e que possui maior (menor) preocupação em sobreviver no mercado de atuação. Por último, a lógica por trás do quarto componente do modelo de Altman, Hartzel e Peck (1998) é de que a empresa pode emitir e vender novas ações no mercado para pagar sua dívida, pois em um cenário de grande capitalização de mercado, comparado à dívida total, a empresa indica uma alta capacidade de realizar essa venda junto aos investidores, caso julgue necessário (ATIYA, 2001).

Na interpretação do resultado obtido pela Equação 1, quanto maior for o valor de Z'' Score, menor será o risco atrelado ao risco de insolvência corporativa. De acordo com Altman (2005), uma organização com pontuação acima de 5,85 tem baixo risco de insolvência, enquanto uma pontuação abaixo de 4,15 representa uma empresa com alto risco de insolvência. Já uma pontuação entre 4,15 e 5,85 indica que a organização está em uma área de um potencial risco de insolvência.

Com o intuito de transformar o resultado do Z'' Score em uma variável *dummy,* a amostra foi dividida em dois grupos: organizações saudáveis e não saudáveis. As organizações são saudáveis, quando recebem valor de Z'' igual ou acima de 4,15, e são não saudáveis, quando recebem o valor de Z'' abaixo de 4,15 (ALTMAN, 2005).

3.1.2 Aplicação do *K-means*

Nesta pesquisa, além de utilizar o Z'' Score de Altman, Hartzel e Peck (1998), houve a adoção de aprendizado de máquina não supervisionado para a operacionalização do risco de insolvência corporativa. As técnicas de aprendizado não supervisionado garantem a homogeneidade dentro de grupos, com a variância mínima, preenchendo lacunas em termos de robustez e heterogeneidade de aglomerados (GENTLEMAN; CAREY, 2008).

O aprendizado de máquina não supervisionado é referenciado como a descoberta de classes e não há treinamento definido para sua aplicação. Após a seleção de amostras a serem utilizadas, é necessária a seleção de recursos para a formação do *clustering* e a escolha de um algoritmo para utilizar (GENTLEMAN; CAREY, 2008). O método escolhido para agrupamento de *clusters* foi o *K-means*.

O algoritmo *K-means* é considerado o modelo de agrupamento não supervisionado mais disseminado na área de ciência de dados, por sua aplicação ocorrer de maneira simples e eficaz. Na *clusterização*, o K se refere ao número de clusters, sendo realizado um procedimento de classificação simples de um conjunto de objetos. O *K-means* representa cada um dos clusters pela média ponderada de seus pontos, ou seja, pelo seu centróide (CAMBRONERO; MORENO, 2006).

A formação dos centróides possui a vantagem de disponibilizar um significado gráfico e estatístico imediato. Cada *cluster* é adaptado pelo valor do centróide (CAMBRONERO; MORENO, 2006). Para a concretização do agrupamento dos clusters pelo *K-means*, foi utilizada apenas uma variável, nesta dissertação representada pelo desvio padrão móvel do retorno sobre os ativos (σROA) de doze trimestres (DAMASCENO, 2021). O σROA é elencado como uma variável precursora do risco de insolvência, sendo considerado um indicativo de volatilidade da lucratividade corporativa (VIEIRA *et al.*, 2020).

3.2 Variáveis preditivas

3.2.1 Coeficiente Intelectual de Valor Agregado

Para mensurar o desempenho do capital intelectual de uma empresa, Pulic (2000) propôs o método denominado *Value Added Intellectual Coefficient* (VAIC). A literatura tem considerado o VAIC como uma boa medida agregada de capacidade do capital intelectual

(ARDALAN; ASKARIAN, 2014; BRANDT; MACHAIEWSKI; GEIB, 2018; DžENOPOLJAC; JANOŁEVIC; BONTIS, 2016). Esse coeficiente é capaz de fornecer à gerência informações sobre desempenho, valor e competitividade da organização, por meio de indicadores financeiros tradicionais (PARDO-CUEVA; HERRERA; GÓMEZ, 2018).

O VAIC mensura indiretamente os recursos intangíveis por meio da medição da eficiência do capital humano, do capital investido e do capital estrutural, conforme Equação 2. Quanto maior a soma desses indicadores, maior será o nível de eficiência da geração de valor de uma organização (BRANDT; MACHAIEWSKI; GEIB, 2018).

$$VAIC_{it} = HCE_{it} + SCE_{it} + CEE_{it} \quad (2)$$

Em que: *VAIC* é o coeficiente intelectual de valor agregado da empresa *i* no momento *t*; *HCE* se refere à eficiência do capital humano da empresa *i* no momento *t*; *SCE* representa a eficiência do capital estrutural da empresa *i* no momento *t;* e *CEE* se refere à eficiência do capital empregado da empresa *i* no momento *t*.

O HCE será obtido por meio da divisão do valor agregado (VA) pelo total de salários dos funcionários (HC) de uma empresa, conforme Equação 3. O VA, por sua vez, representa a soma entre lucro operacional, custos com funcionários, depreciação e amortização.

$$HCE_{it} = VA_{it}/HC_{it} \quad (3)$$

Em que: *HCE* se refere à eficiência do capital humano da empresa *i* no momento *t*; *VA* representa o valor agregado da empresa i no momento *t*; *HC* é o total de salários dos funcionários da empresa *i* no momento *t*.

O SCE será mensurado pela divisão entre o capital estrutural (SC) e o VA de uma empresa, conforme Equação 4. Para obter o valor de SC, é preciso subtrair os custos de HC do VA, sendo visualizado como os custos necessários para criar valor além dos recursos humanos.

$$SCE_{it} = SC_{it}/VA_{it} \quad (4)$$

Em que: *SCE* se refere à eficiência do capital estrutural da empresa *i* no momento *t*; *SC* representa o capital estrutural da empresa *i* no momento *t*; e o *VA* é o valor agregado da empresa *i* no momento *t*.

Por último, o CEE será representado pela divisão entre VA e ativos líquidos (CE) da empresa, conforme Equação 5. No modelo VAIC, o CE representa o capital empregado que foi investido em uma empresa no passado.

$$CEE_{it} = VA_{it}/CE_{it} \quad (5)$$

Em que: *CEE* representa a eficiência do capital empregado da empresa *i* no momento *t*; *VA* refere-se ao valor agregado da empresa *i* no momento *t*; *CE* representa os ativos líquidos da empresa *i* no momento *t*.

A literatura tem julgado o VAIC como a técnica mais atraente para mensurar o desempenho do capital intelectual (CI), pois: é uma medida padronizada e coerente do desempenho do CI; é aceita como uma ferramenta útil para avaliação dos resultados empresariais (FIJAŁKOWSKA, 2014); e representa um poderoso preditor para o desempenho financeiro de empresas (CHU; CHAN; WU, 2011).

3.2.2 As demais variáveis preditivas

Neste estudo, foram utilizadas nove variáveis preditivas. Além do VAIC, as variáveis adotadas na aplicação dos modelos de aprendizado de máquina foram: tamanho da empresa, ROA (*Return On Assets*), alavancagem financeira, liquidez corrente, giro do ativo, a razão do capital de giro pelo ativo total, a razão do fluxo de caixa pela dívida total e a razão do fluxo de caixa pelas vendas. Variáveis como essas foram utilizadas em estudos anteriores com problemática similar (BARBOZA; KIMURA; ALTMAN, 2017; CHEN; CHEN; LIEN, 2020; LIN; LIANG; CHEN, 2011; SHAHWAN; HABIB, 2020; WANG; CHEN; CHU, 2018).

O tamanho da organização pode ser um dos principais determinantes que influenciam o risco de insolvência corporativa, e grandes empresas tendem a possuir menos risco em suas operações, devido à sua capacidade de diversificação (CHAKRABORTY; GAO; SHEIKH, 2018). Em relação ao ROA, as empresas que apresentam maior rentabilidade dependem menos de dívidas para financiar suas atividades e, assim, quanto menor a dívida, menor o risco de dificuldades financeiras (MYERS; MAJLUF, 1984).

Sobre a alavancagem, em nações emergentes pelo qual a estrutura de capital depende fortemente de financiamento, espera-se que a alavancagem financeira seja maior para empresas com alto risco de insolvência corporativa. A determinação da variável liquidez corrente como variável preditora, está na identificação da capacidade de cumprimento das obrigações em curto

prazo, pois quanto maior for o valor desse indicador, maior será sua capacidade de quitar suas dívidas com terceiros (SANTANA FILHO *et al.*, 2019).

O giro do ativo relaciona o quanto a empresa vendeu em determinado período, em relação ao valor do ativo total. Com o número encontrado dessa relação, é possível perceber se a organização utiliza adequadamente seus bens na produção de capital financeiro. Sobre a razão entre o capital de giro e o total de ativos, tem-se que o capital de giro pode ser visto como uma indicação da capacidade da empresa de pagar suas dívidas de curto prazo, e essa capacidade é comparada ao tamanho da organização representado pelo total de ativos (ATIYA, 2001).

A razão entre o fluxo de caixa da empresa e sua dívida total representa uma medida de reserva do caixa para o pagamento de suas dívidas. Por último, a razão entre o fluxo de caixa da empresa e as vendas apresenta o resultado da relação de reserva do caixa em torno do total de vendas efetuadas (GOMBOLA *et al.*, 1987). Diante das nove variáveis preditivas apresentadas, para analisar a predição do risco de insolvência corporativa das empresas da amostra, foi estimada a Equação 6.

$$RISCO_{it} = \beta_0 + \beta_1 VAIC_{it} + \beta_2 TAM_{it} + \beta_3 ROA_{it} + \beta_4 ALAV_{it} + \beta_5 LIQC_{it} + \beta_6 GIRO_{it} + \beta_7 CGAT_{it} + \beta_8 FCDT_{it} + \beta_9 FCV_{it} + \varepsilon_{it} \quad (6)$$

Em que: *RISCO* representa o risco de insolvência, substituído pelo Z'' Score ou σROA da empresa *i* no momento *t*; *VAIC* é a pontuação do valor agregado de capital intelectual da empresa *i* no momento *t*; *TAM* representa o tamanho da empresa *i* no momento *t*; *ROA* é o retorno sobre os ativos da empresa *i* no momento *t*; *ALAV* representa a alavancagem financeira da empresa *i* no momento *t*; *LIQC* representa a liquidez corrente da empresa *i* no momento *t*; *GIRO* representa o giro do ativo da empresa *i* no momento *t*; *CGAT* representa o capital de giro dividido pelo ativo total da empresa *i* no momento *t; FCDT* é o fluxo de caixa dividido pela dívida total da empresa *i* no momento *t*; *FCV* representa o fluxo de caixa dividido pelo total de vendas da empresa *i* no momento *t;* β_0 é uma constante; β_n e α_j representam os coeficientes das variáveis independentes; ε refere-se ao tamanho do erro.

Além disso, a operacionalização de cada variável preditiva selecionada para este estudo pode ser observada com maior detalhamento no Quadro 2.

Quadro 2 – Variáveis preditoras do modelo

VARIÁVEL	OPERACIONALIZAÇÃO	REFERÊNCIAS
Value Added Intellectual Coefficient (VAIC)	VAIC = HCE (Eficiência do Capital Humano) + SCE (Eficiência do Capital Estrutural) + CEE (Eficiência do Capital Empregado)	Ardalan e Askarian (2014); Shahwan e Habib (2020)
Tamanho (TAM)	TAM = ln(Ativo Total)	Chakraborty, Gao e Sheikh (2018); Dong *et al.* (2014); Shahwan e Habib (2020)
Retorno sobre os ativos (ROA)	ROA = Lucro Líquido / Ativo Total	Myers e Majluf (1984); Chen, Chen e Lien (2020)
Alavancagem financeira (ALAV)	ALAV = Passivo Oneroso / Ativo total	Shahwan (2015); Shahwan e Habib (2020)
Liquidez Corrente (LIQC)	LIQC = Ativo Circulante / Passivo Circulante	Wang, Chen, Chu (2018);
Giro do Ativo (GIRO)	GIRO = Receita Líquida / Ativo Total	Chang, Hsu (2016); Barboza, Kimura e Altman (2017)
Capital de Giro / Ativo total (CGAT)	CGAT = Capital de Giro / Ativo total	Barboza, Kimura e Altman (2017); Zięba, Tomczak e Tomczak (2016); Antunes, Ribeiro e Pereira (2017)
Fluxo de caixa / Dívida total (FCDT)	FCDT = Fluxo de Caixa / Dívida Total	Lin, Liang e Chen (2011); Wang e Wu (2017)
Fluxo de caixa / Vendas (FCV)	FCV = Fluxo de Caixa / Vendas	Lin, Liang e Chen (2011); Li, Sun e Sun (2009)

Fonte: Elaborado pelo autor (2021)

3.3 Modelos de aprendizado de máquina supervisionados

O aprendizado de máquina supervisionado é percebido como uma das tecnologias da inteligência artificial mais disruptivas da última década, impactando muitas áreas de estudo, inclusive as que tratam de finanças corporativas. Essa técnica incorpora uma ampla gama de algoritmos e ferramentas de modelagem usadas para processamento de dados, com o objetivo de reconhecer padrões para tratar de problemas que costumam ser despercebidos (CARLEO *et al.*, 2019; MUHAMEDYEV, 2015). O treinamento dos dados foi executado com 75% das observações sem dados faltantes, e o teste foi executado com o restante das observações. Essas observações foram separadas para treinamento e teste de maneira aleatória.

Os modelos de aprendizagem de máquina que foram utilizados para a análise empírica deste estudo são: *random forest*, *naive bayes*, *logit*, K-NN, SVM (linear, polinomial e de base radial), *bagging* e *boosting*. Esses métodos foram escolhidos por representarem alguns dos algoritmos mais relevantes de aprendizado de máquina supervisionado e por serem de fácil aplicação. Além disso, modelos como esses estão sendo utilizados em estudos com

problemáticas que envolvem insolvência corporativa (BARBOZA; KIMURA; ALTMAN, 2017; RAHAYU; SUHARTANTO, 2020; SUSS; TREITEL, 2019).

O modelo *random forest* é um algoritmo de aprendizagem que utiliza de múltiplas árvores. Esse método é construído por subamostras e não na amostra original, por meio do método chamado de *bootstrapping.* Ao utilizar dessa técnica de aprendizagem em conjunto, há uma redução da variação em comparação com as árvores de decisão única, devido às árvores usadas para elaborar a floresta possuírem características amplamente distintas daquelas usadas individualmente. O método é capaz de aumentar a precisão, ao reduzir distorções e permitir o crescimento das árvores, com uma variação suportável a partir da junção de resultados individuais. É um método simples e que geralmente demonstra bom desempenho (COURONNÉ; PROBST; BOULESTEIX, 2018; SCHONLAU; ZOU, 2020).

O algoritmo de execução do *random forest* precisa seguir seis procedimentos: 1) extrair os dados originais para que a árvore de decisão seja construída em apenas uma amostra do conjunto de dados original; 2) extrair o subconjunto das variáveis independentes durante a construção da árvore de decisão; 3) construir uma árvore de decisão com base nos dados do subconjunto onde o subconjunto de linhas e colunas é usado como conjunto de dados; 4) prever no conjunto de dados de teste ou validação; 5) repetir as etapas 1 a 3 em um número n de vezes, onde n é o número de árvores construídas; e 6) a previsão final sobre o conjunto de dados do teste é a média de previsões de todas as n árvores (AYYADEVARA, 2018).

A técnica de *random forest* consiste em conjunto de árvores simples T_1, ..., T_N pelo qual a função de classificação pode ser expressa por $h(X,\Phi_1)$, $h(X,\Phi_N)$, onde *h* é uma função, *X* é um preditor e $\Phi 1,\ldots,\Phi N$ são vetores aleatórios independentes, igualmente distribuídos (GREGOVA *et al.,* 2020). O método é capaz de aumentar a precisão, ao reduzir distorções e permitir o crescimento das árvores, com uma variação suportável a partir da junção de resultados individuais. A sua maior limitação é a necessidade de uma grande quantidade de árvores, que pode tornar o algoritmo lento e ineficiente para execução de predições.

O modelo *naive bayes* é considerado uma das especificações mais válidas para modelar sistemas complexos. Essa técnica é capaz de prever probabilidades de que um determinado conjunto de dados pertence a um determinado rótulo de classe. Um classificador *naive bayes* considera que a ausência ou presença de um determinado atributo de uma classe não está relacionada à ausência ou presença de qualquer outra característica quando a variável de classe é definida. Esse modelo é rápido para aplicações com *big data* e não é afetado por problemas de dimensionalidade (JADHAV; CHANNE, 2016).

As vantagens de utilizar o método *naive bayes* é que necessita de pouco tempo computacional para treinamento, geralmente demonstra um bom desempenho e remove os recursos considerados irrelevantes. Como base para a técnica *naive bayes,* tem-se o Teorema de Bayes, utilizado para calcular a probabilidade de P(C|X), P(C), P(X) e P(X|C) (JADHAV; CHANNE, 2016).

O algoritmo *naive bayes* cria uma estrutura de decisão em forma de árvore, pelo qual começa com um único nó raiz e vai sendo fragmentada até alcançar a classe pretendida. O método possui dois momentos: construção e simplificação. Na fase da construção é feita a avaliação dos pontos de separação em potencial e a geração das partições. No momento da simplificação, após definição do melhor ponto de separação de cada nó, são geradas partições pela aplicação criteriosa de separação (SILVA *et al*., 2017).

A regressão logística é um modelo que tenta verificar a dependência unilateral das variáveis, pelo qual as variáveis dependentes examinadas são ordinárias, binárias ou categóricas e as variáveis independentes podem pertencer a qualquer classificação. Em comparação à regressão linear, a regressão logística possui menor rigor em suas restrições, visto que não é exigida a normalidade das variáveis ou homoscedasticidade de grupos individuais. O *logit* é um modelo facilmente implementado e simples de ser treinado e interpretado, mas também é um algoritmo vulnerável ao *overfitting* e não resolve problemas não lineares (MUHAMEDYEV, 2015).

Ao investigar a dificuldade financeira, cada empreendimento pertencerá a uma categoria, dependendo do valor que a variável dependente receberá. A modelagem da empresa saudável/não saudável baseia-se na probabilidade condicional da variável dependente (Y), dependendo das variáveis independentes (X). Dessa forma, a regressão logística define que uma organização não é saudável, se a probabilidade prevista for maior que o valor limite estabelecido. Caso o valor previsto estiver abaixo dos valores-limite determinados, a organização pertence ao grupo de empreendimentos saudáveis (GREGOVA *et al.*, 2020).

A classificação K-NN é considerada um dos dez principais algoritmos de mineração de dados, e encontram dois desafios: definir o valor K adequado e determinar a função de distância para identificar K vizinhos mais próximos. Esse método realiza a classificação por meio da distância entre os dados nas amostras de treinamento, sendo a distância um elemento suficiente para realizar inferência. Por conta disso, é chamado de aprendizado baseado em instâncias, devido à ausência de parâmetros de aprendizado. Além disso, sua implementação é simples, rápida e funciona bem com dados não lineares (ZHANG, 2019).

As vantagens que o K-NN possui são: (a) facilidade de implementação e depuração; (b) modelo eficaz quando a análise dos vizinhos for importante; e (c) algumas técnicas de redução de ruído funcionam apenas com o K-NN, sendo provedoras de melhoria na predição do classificador. Em relação às desvantagens, o K-NN pode apresentar: (a) tempo de execução e desempenho ruins, caso o conjunto de treinamento for grande; (b) sensibilidade a recursos redundantes que podem contribuir para a similaridade; e (c) em tarefas muito complexas, outros modelos de aprendizado de máquina podem apresentar melhores resultados, como o SVM e as redes neurais (CUNNINGHAM; DELANY, 2007).

O método SVM é um modelo de aprendizado de máquina que tenta realizar a adequação de uma linha, ou de um hiperplano, entre as diferentes classes, com o objetivo de maximizar a distância até alcançar os pontos das classes. Esse modelo é recomendado pela literatura, devido à sua capacidade de aprender padrões de classificação de dados com precisão equilibrada e reprodutibilidade. Um SVM pode ser linear ou não linear, sendo eficaz em espaços de alta dimensão. É por meio do truque de kernel que há possibilidades de trabalhar com a não linearidade (PISNER; SCHNYER, 2020).

O truque de kernel é a suposição de que é possível criar uma fronteira de decisão em uma nova dimensão, minimizando e facilitando os cálculos, em comparação a estar mapeando os pontos para a nova dimensão. O modelo SVM é interessante para a classificação de dados dispersos de forma não regular e tem resultados positivos em espaços com muitas dimensões, convergindo para o melhor hiperplano possível, pelo qual o algoritmo não se perde, como geralmente ocorre na execução de redes neurais. Desvantagens do modelo que podem ser elencadas são: (a) o resultado ser de difícil interpretação; e (b) conforme o conjunto de dados vai aumentando, o tempo para conclusão dos dados cresce rapidamente e a dificuldade para interpretar também (PISNER; SCHNYER, 2020). Como na maioria dos casos os dados não são lineares, além do SVM linear, este estudo utilizou do SVM polinomial e de base radial.

O método *bagging* para árvores de classificação foi sugerido por Breiman (1996) e Breiman (1998), com o intuito de estabilizar as árvores e reduzir possibilidades de *overfitting* de uma classe. Esse método é popularmente conhecido como agregação de *bootstrapping,* que considera classificadores independentes e utilizadores de partes dos dados. Ou seja, há a criação de novos subconjuntos aleatórios de dados por meio de amostragem, com reposição, gerando estimativas de intervalo de confiança. Após o tratamento feito com os dados, os mesmos são combinados por meio do cálculo da média do modelo, proporcionando resultados mais eficientes. Esse método cria árvores distintas e combinam suas saídas para melhorar o poder preditivo em relação a árvores de decisão (BREIMAN, 1996).

Na maior parte das vezes um classificador combinado gera melhores resultados do que classificadores individuais, por combinar as vantagens dos classificadores individuais na resolutiva final. A decisão do melhor classificador pode ser feita por simples votação por maioria, média de probabilidades, produto de probabilidades, etc (SKURICHINA; DUIN, 2002). Dessa forma, o *bagging* pode ser útil para construir um classificador melhor, mas somente haverá a maximização de desempenho da classificação apenas em situações muito instáveis. Se a classificação apresentar bastante estabilidade, o método pode ser inútil (SKURICHINA; DUIN, 1998).

Por último, o método *boosting* segue a ideia de adicionar, de maneira sequencial, novos modelos ao conjunto. Os conjuntos de dados de treinamento e classificadores são obtidos sequencialmente no algoritmo, em contraste ao método *bagging*, pelo qual os dados de treinamento e classificadores são obtidos independentemente e aleatoriamente da etapa anterior do algoritmo. A cada etapa executada, os dados de treinamento, na aplicação do método *boosting*, recebem novos pesos de tal maneira que os objetos classificados erroneamente ganham pesos maiores em um novo conjunto de treinamento modificado (SKURICHINA; DUIN, 2002).

A cada iteração efetuada, um novo modelo considerado fraco é treinado. O modelo fraco é aquele pelo qual a taxa de erro é um pouco melhor do que a suposição aleatória inicial. Já que é feita de forma sequencial, até o fim da execução do algoritmo, há uma constante melhora dos erros, mesmo que pequena. A partir disso, algumas vantagens podem ser elencadas: velocidade, melhoria da precisão e diminuição de chances de ocorrer *overfitting* (FRIEDMAN, 2001). O método *boosting* utilizado neste estudo foi o *gradient boosting*.

Para operacionalização dos métodos de aprendizado de máquina, os dados foram padronizados, para garantir melhor adequação aos modelos. O desempenho dos modelos foi avaliado a partir das seguintes métricas: acurácia, sensibilidade, especificidade, MAE (*Mean Absolut Error*), RMSE (*Root Mean Squared Error*), AUC (*Area Under the ROC Curve*) e ROC (*Receiver Operating Characteristic*). Os valores encontrados pelas métricas consideram os valores de verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) e falso negativo (FN). FP e FN representam os resultados com a classificação errada, enquanto VP e VN evidenciam os resultados com a classificação correta (DEVI; RADHIKA, 2018).

3.5 Coleta e tratamento dos dados

Os dados foram coletados a partir das demonstrações financeiras das empresas listadas na Brasil, Bolsa, Balcão [B]³. Para essa coleta, utilizou-se do *site* público da Comissão de Valores Mobiliários – CVM, com a finalidade de capturar dados da DVA - Demonstração de Valor Adicionado, referentes à remuneração de funcionários, e os demais dados foram capturados pela plataforma Thomson Reuters Eikon, que contempla as informações de empresas brasileiras de capital aberto.

Os dados analisados obedecem ao período de 2010 a 2020. Esse período foi escolhido pelo motivo das informações referentes à remuneração dos funcionários só estarem disponíveis para consulta e download na CVM a partir do primeiro trimestre de 2010, apesar da adoção da DVA ser obrigatória para empresas de capital aberto, por intermédio da Lei nº 11.638, a partir de 28 de dezembro de 2007.

Empresas de todos os ramos foram considerados neste estudo e, de maneira particular, as instituições financeiras foram analisadas conjuntamente e separadamente, diante de suas especificidades, tais como: normatização própria, supervisão por entidades reguladoras específicas, o dinheiro como a principal matéria-prima e concentração em operações de curto prazo (SILVA *et al.*, 2019). Ressalta-se que empresas foram excluídas da amostra por não obedecer ao critério de apresentar todos os dados necessários para a execução dos testes. A quantidade inicial e final para cada ano pode ser verificada na Tabela 1.

Tabela 1 – Quantidades iniciais e finais da amostra do estudo

ANO	EMPRESAS			OBSERVAÇÕES		
	QUANT. INICIAL	QUANT. FINAL	%	QUANT. INICIAL	QUANT. FINAL	%
2010	264	140	53	1023	394	39
2011	269	154	57	1063	443	42
2012	273	159	58	1075	457	43
2013	275	169	61	1096	483	44
2014	280	175	63	1110	513	46
2015	290	175	60	1152	514	45
2016	296	177	60	1171	528	45
2017	306	182	59	1209	531	44
2018	315	187	59	1241	545	44
2019	353	203	58	1321	591	45
2020	364	211	58	1427	611	43
Média anual	299	176	59	1172	510	44

Fonte: Dados da pesquisa

Ao final, considerando a média anual, 299 empresas brasileiras de capital aberto tiveram seus dados analisados pelo estudo. Além disso, 5610 observações sem dados faltantes foram utilizadas. Apesar da pandemia ocasionada pelo Covid-19, desconsiderar os dados do ano de 2020 não apresenta diferenças significativas nos resultados (verificar Apêndice A), assim como desconsiderar dados de empresas financeiras também não apresenta diferenças significativas (verificar Apêndice B). Toda análise estatística foi feita com o apoio dos *softwares* R (versão 4.0.2) e R Studio (versão 1.4.1106). Para capturar as saídas da aplicação dos modelos de aprendizado de máquina, foram utilizadas as seguintes bibliotecas: caret, e1071, ipred, randomForest, stats, pRoc e ROCR.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Os dados originais, referentes às variáveis, apresentam valores extremos que podem poluir os resultados da amostra em quesitos pontuais. Dessa forma, os dados perpassaram pelo processo de winsorização ao nível de 5%, com o intuito de eliminar os efeitos negativos que os *outliers* podem levar à análise. Nesse sentido, os dados que desviam do padrão foram cortados, sendo igualados aos dados limítrofes (menor e maior dado do conjunto).

4.1 Estatísticas descritivas

Neste momento, os dados winsorizados ao nível de 5% das variáveis preditivas são analisados a partir de medidas estatísticas básicas. Esses dados estão presentes na aplicação dos modelos supervisionados de aprendizado de máquina. As estatísticas do Z'' Score e o Risco mensurado pelo desvio padrão móvel do ROA foram desconsiderados neste momento, por se tornarem variáveis categóricas, após aplicação dos pontos de corte. A estatística descritiva pode ser visualizada na Tabela 2.

O VAIC apresentou valores mínimo e máximo de -0,853 e 10,159, respectivamente. A média e a mediana encontradas foram similares, sendo nessa ordem, de 2,918 e 2,186. Esses valores correspondem aos recursos intangíveis que envolvem o capital humano, o capital investido e o capital estrutural (BRANDT; MACHAIEWSKI; GEIB, 2018).

Tabela 2 – Estatística descritiva das variáveis preditivas

	Min	1º Q	Mediana	Média	3º Q	Máximo	Desvio Padrão
VAIC	-0,853	1,613	2,186	2,918	3,624	10,159	2,533
TAM	19,12	20,690	21,860	21,900	23,130	24,720	1,587
ROA	-0,050	-0,004	0,006	0,003	0,015	0,037	0,021
ALAV	0,028	0,169	0,309	0,314	0,435	0,671	0,178
LIQC	0,384	1,110	1,661	1,797	2,327	4,018	0,941
GIRO	0,017	0,075	0,134	0,154	0,217	0,392	0,103
CGAT	-0,390	0,024	0,130	0,128	0,273	0,468	0,206
FCDT	0,002	0,036	0,108	0,152	0,223	0,534	0,146
FCV	0,019	0,166	0,490	0,776	1,024	3,140	0,839

Legenda: VAIC é a pontuação do valor agregado de capital intelectual da empresa *i* no momento *t*; TAM representa o tamanho da empresa *i* no momento *t*; ROA é o retorno sobre os ativos da empresa *i* no momento *t*; ALAV representa a alavancagem financeira da empresa *i* no momento *t*; LIQC representa a liquidez corrente da empresa *i* no momento *t*; GIRO representa o giro do ativo da empresa *i* no momento *t*; CGAT representa o capital de giro dividido pelo ativo total da empresa *i* no momento *t*; FCDT é o fluxo de caixa dividido pela dívida total da empresa *i* no momento *t*; FCV representa o fluxo de caixa dividido pelo total de vendas da empresa *i* no momento *t*.
Estatísticas obtidas após winsorização de 5%.
Fonte: Dados da pesquisa

Para o conjunto de variáveis de interesse, o desvio padrão do VAIC é o maior, com o valor de 2,533. Ao trabalhar com um conjunto de dados de doze empresas brasileiras listadas na [B]³, do comércio varejista, Brandt, Machaiewski e Geib (2018) encontraram um maior desvio padrão para o VAIC dentre as variáveis de interesse do estudo, com um valor de 3,093, corroborando os resultados desta pesquisa. Esse achado reflete a discrepância no total de investimentos em capital intelectual que as organizações brasileiras de capital aberto possuem.

Para a variável TAM, a média e a mediana encontrada apresentam valores bem aproximados, de 21,9 e 21,86, respectivamente. Ressalta-se que antes da winsorização, os dados passaram por uma transformação logarítmica, o que contribuiu para essa proximidade. Os valores de mínimo e de máximo correspondem a 19,12 e 24,72, respectivamente. As organizações com valor próximo a 24,72 tendem a possuir menos risco em suas operações, devido à sua capacidade de diversificação (CHAKRABORTY; GAO; SHEIKH, 2018).

No critério rentabilidade, ao avaliar os resultados do ROA, percebe-se que a média para as empresas brasileiras é de 0,3%, e mediana de 0,6%. Em relação aos valores de mínimo e máximo, existem empresas que indicam uma maior rentabilidade (3,7%), e outras que indicam uma menor rentabilidade, de valor negativo (-5%), ou seja, empresas com prejuízo em determinado trimestre. Empresas com ROA negativo (positivo) tendem a possuir maior (menor) risco de enfrentar dificuldades financeiras, por depender de capital externo (interno) para a sua sobrevivência (MYERS; MAJLUF, 1984). Além disso, o desvio padrão da variável é de 0,021, sendo o menor entre a lista de variáveis preditivas.

A alavancagem financeira (ALAV) possui média de 0,314, mediana de 0,309, máximo de 0,671 e mínimo de 0,028. Dessa forma, as empresas brasileiras, em média, possuem 31,4% de passivo oneroso em relação ao ativo total. As empresas menos endividadas se concentram entre 2,8% e 16,9% de passivo oneroso, enquanto as mais endividadas entre 43,5% a 67,1%. A média de alavancagem financeira para empresas brasileiras listadas na [B]³, encontrada por Barros *et al.* (2014), foi de 0,253, ao trabalhar com dados de 2002 a 2011. Apesar de Barros *et al.* (2014) desconsiderarem as empresas financeiras da amostra, é possível sugerir que as organizações podem estar se endividando mais nos últimos anos.

A liquidez corrente (LIQC) recebeu média de 1,797 entre as empresas, indicando boa capacidade de quitar dívidas a curto prazo. Organizações com maior dificuldade de quitar suas dívidas de curto prazo são as que estão entre o valor de mínimo (0,384) e o valor do primeiro quartil (1,11). Na perspectiva contrária, organizações com menor dificuldade de quitar suas dívidas de curto prazo são as que receberam valor entre o terceiro quartil (2,327) e o máximo

(4,018). Quanto maior for o valor desse indicador, maior será a capacidade empresarial de quitar suas dívidas com terceiros (SANTANA FILHO *et al.*, 2019).

Em relação ao giro do ativo (GIRO), a média apresentada foi de 15,4%, representando o quanto a organização vende em relação ao total de ativos que ela possui. As empresas que apresentam um menor resultado para essa relação encontram-se entre os valores de 1,7% a 7,5%. Já as organizações que mais vendem, diante de seu porte mensurado pelo ativo total, são as que possuem observações com valores entre 21,7% a 39,2%, e são elencadas como as que utilizam mais adequadamente seus bens na produção de capital financeiro.

O capital de giro comparado ao ativo total (CGAT) possui valores de -39% a 46,8%. Para a média, foi encontrado o valor de 12,8%, próximo da mediana (13%). As organizações com menores CGAT encontram-se entre o valor de mínimo e 2,4%. Já as empresas com maiores CGAT, encontram-se entre 27,3% e valor de máximo. O capital de giro é visto como uma indicação da capacidade da organização de pagar suas dívidas de curto prazo. Quanto maior for sua capacidade em relação ao total de ativos, menor tenderá a ser o risco de inadimplência (ATIYA, 2001).

O fluxo de caixa comparado à dívida total (FCDT) recebeu média de 0,152 e mediana de 0,108, representando o percentual da diferença entre entradas e saídas de caixa em relação à dívida total efetuada em determinado período. As observações com valores menores de FCDT estão entre 0,002 e 0,036. Em contrapartida, as observações com valores maiores estão entre 0,223 e 0,534. A razão entre o fluxo de caixa e a dívida total representa uma medida de reserva do caixa para o pagamento de suas dívidas (GOMBOLA et al., 1987). Empresas com maior valor de FCDT tendem a ter menos complicações em relação a quitar dívidas com terceiros.

Por último, o fluxo de caixa comparado ao valor de vendas (FCV) recebeu valores entre 0,019 e 3,14, representando o mínimo e o máximo, respectivamente. As observações pertencentes ao 1º quartil da amostra recebeu valores de até 0,166. As observações com maiores valores são as que estão entre 1,024 e valor máximo. A média recebeu o valor de 0,776. Essas observações consideram o quanto do resultado da diferença entre entradas e saídas de caixa representa em relação ao total de vendas (GOMBOLA *et al.*, 1987).

Analisar a correlação de variáveis é importante, para a aplicação de técnicas estatísticas tradicionais, como o *logit* (BARBOZA; KIMURA; ALTMAN, 2017). Dessa forma, na Tabela 3 são apresentados os valores de correlação e nível de significância entre os atributos desta pesquisa.

Percebe-se que as correlações entre GIRO e ALAV e LIQC e TAM não são estatisticamente significativas, ao nível de 5%. Em suma, a maioria das correlações apresentou

significância estatística, mas não com valores fortes. A maior correlação significativa está entre CGAT e LIQC, com o valor de 0,82 e p valor <0,001, sendo a única relação que pode gerar problemas de multicolinearidade.

Tabela 3 – correlação de Pearson entre as variáveis preditivas

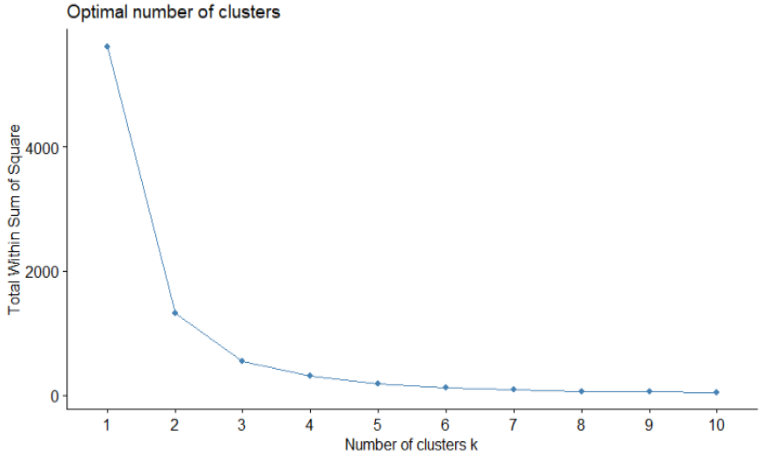
	VAIC	TAM	ROA	ALAV	LIQC	GIRO	CGAT	FCDT	FCV
VAIC	1								
TAM	0,23***	1							
ROA	0,36***	0,21***	1						
ALAV	0,07***	0,21***	-0,26***	1					
LIQC	0,07***	0,03	0,34***	-0,28***	1				
GIRO	-0,14***	-0,21***	0,17***	-0,03	-0,06**	1			
CGAT	0,05***	0,06***	0,42***	-0,20***	0,82***	0,14***	1		
FCDT	0,09***	-0,03*	0,32***	-0,16***	0,48***	0,08***	0,48***	1	
FCV	0,21***	0,12***	0,03*	0,1***	0,31***	-0,43***	0,2***	0,57***	1

Legenda: VAIC é a pontuação do valor agregado de capital intelectual da empresa *i* no momento *t*; TAM representa o tamanho da empresa *i* no momento *t*; ROA é o retorno sobre os ativos da empresa *i* no momento *t*; ALAV representa a alavancagem financeira da empresa *i* no momento *t*; LIQC representa a liquidez corrente da empresa *i* no momento *t*; GIRO representa o giro do ativo da empresa *i* no momento *t*; CGAT representa o capital de giro dividido pelo ativo total da empresa *i* no momento *t*; FCDT é o fluxo de caixa dividido pela dívida total da empresa *i* no momento *t*; FCV representa o fluxo de caixa dividido pelo total de vendas da empresa *i* no momento *t*.
Nota: * p<0,05; ** p<0,01; *** p<0,001
Fonte: Dados da pesquisa

4.2 Agrupamento do risco de insolvência

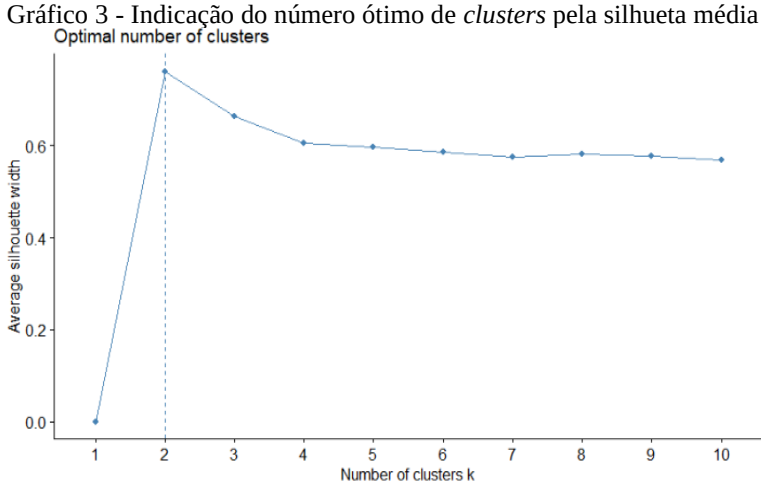
As observações foram agrupadas de duas formas em relação ao risco de insolvência: por meio da *dummy* gerada pelo ponto de corte do Z'' Score de Altman, Hartzel e Peck (1998) e pela aplicação do aprendizado de máquina *K-means*, com o uso da variável desvio padrão móvel do ROA de 12 trimestres (DAMASCENO, 2021). Para validar o número ótimo de *clusters*, foram utilizados dois métodos: (a) a soma dos quadrados dentro do cluster e (b) a silhueta média. Os resultados podem ser vistos nos Gráficos 2 e 3, indicando o número ótimo de dois *clusters*. Dessa forma, foram considerados dois grupos: de maior e de menor risco.

Gráfico 2 - Indicação do número ótimo de *clusters* pela soma dos quadrados



Fonte: Dados da pesquisa

Gráfico 3 - Indicação do número ótimo de *clusters* pela silhueta média



Fonte: Dados da pesquisa

Na Tabela 4, é possível visualizar o *matching* alcançado por meio dessas duas medidas de risco. O *matching* considerou a quantidade relativa de observações tidas como risco comum pelas duas variáveis, em relação à quantidade total de observações do Z'' Score. É perceptível que o *matching* entre as observações de menor risco é expressivo, coincidindo em 94,3% de observações elencadas igualmente pelas medidas Z'' Score e σROA. Esse resultado demonstra que há uma boa aceitabilidade das duas medidas para organizações que apresentam baixo risco de insolvência.

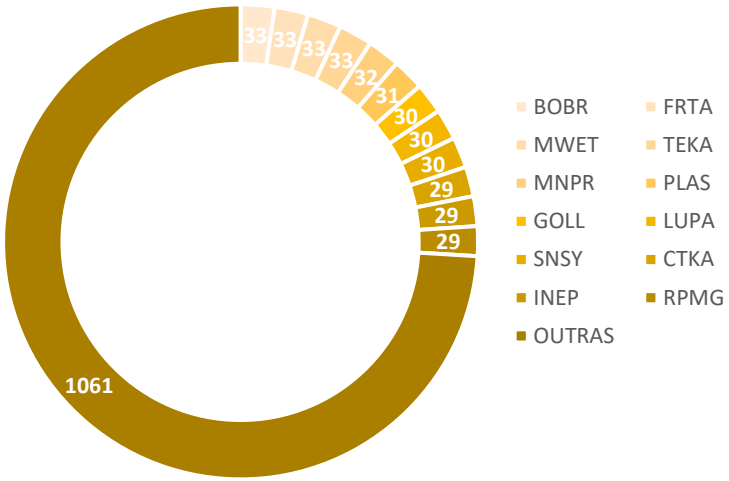
Tabela 4 – *Matching* entre as variáveis de risco de insolvência

Risco	**Z'' Score**	**σROA**	***Matching***	***Matching* (%)**
Maior risco	1433	834	596	41,59%
Menor risco	4177	4776	3939	94,30%

Fonte: Dados da pesquisa

De maneira contrária, para as observações que representam maior risco, apenas 41,59% foram elencadas igualmente pelas duas medidas de risco. Diante desse resultado de baixo *matching*, foram consideradas as duas métricas de risco para realizar as predições no decorrer do trabalho. No Gráfico 4, a partir do agrupamento efetuado pelo Z'' Score, é possível visualizar as organizações com o maior quantitativo de observações que se enquadram no grupo de maior risco.

Gráfico 4 – Observações de empresas com o maior risco de insolvência medido pelo Z'' Score

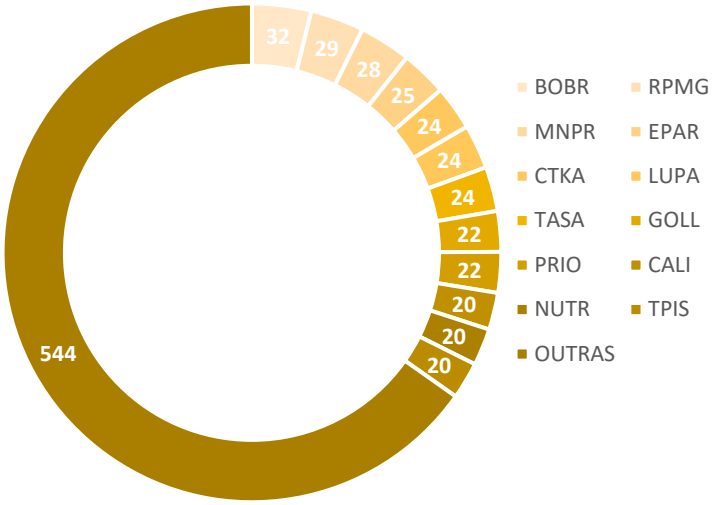


Fonte: Dados da pesquisa

Para a identificação das empresas, foram utilizados os códigos de negociação da [B]³, desconsiderando a numeração que indica a natureza do tipo de ação. As doze empresas com maior risco de insolvência tiveram de 29 a 33 observações, de um total de 44 possibilidades dentro do período estudado. Ressalta-se que o Z'' Score identificou 109 empresas com ao menos uma observação indicativa de maior risco de insolvência. No Gráfico 5, a partir do agrupamento efetuado pelo *K-means*, utilizando da variável σROA, é possível visualizar o maior quantitativo de observações que se enquadram como maior risco.

As doze empresas com maior risco de insolvência, mensurado pelo σROA, tiveram de 20 a 32 observações com esse indicativo. O σROA detectou 86 organizações com ao menos uma observação de maior risco de insolvência. Ao comparar os resultados dos Gráficos 4 e 5, observa-se que as organizações BOBR (Bombril S/A), MNPR (Minupar Participações S/A), GOLL (Gol Linhas Aereas Inteligentes S/A), LUPA (Lupatech S/A), CTKA (Karsten S/A) e RPMG (Refinaria De Petroleos Manguinhos S.A.) estão presentes entre as doze empresas com o maior número de observações insolventes, pelos dois métodos utilizados.

Gráfico 5 – Observações de empresas com maior risco de insolvência medido pelo σROA



Fonte: Dados da pesquisa

Corroborando esses resultados, na Figura 1, é possível ter uma outra visualização das organizações com observações de maior risco de insolvência, por meio da representação visual de nuvem de palavras.

Figura 1 – Nuvens de palavras das empresas com o maior risco de insolvência

Z'' Score	σROA
	

Fonte: Dados da pesquisa

Quanto maior a frequência das organizações com dificuldades financeiras durante os trimestres, maior o tamanho das palavras que constam na Figura 1. Esse agrupamento de empresas com maior risco de insolvência serve de insumo informacional para executivos, acionistas e demais *stakeholders* tomarem melhores decisões. Stupp, Flach e Mattos (2018) elencaram as empresas Karsten S/A, Bombril S/A, Lupatech S/A, CALI (Construtora Adolpho

Lindenberg S/A), INEP (Inepar S/A Indústria e Construções) e SNSY (Sansuy S/A) como insolventes, reforçando os achados de organizações com maior risco de insolvência, destacados nos Gráficos 4 e 5.

4.3 Avaliação dos modelos de aprendizado de máquina supervisionados

Para execução e avaliação dos modelos de aprendizado de máquina, foram utilizadas as duas medidas de risco elencadas no estudo: σROA e Z'' Score. Os modelos foram avaliados a partir das seguintes métricas: erro tipo 1 e tipo 2, sensibilidade, especificidade, acurácia, AUC, MAE, RMSE e ROC. Além disso, foi analisado o grau de importância das variáveis na aplicação do modelo que apresentou melhores resultados.

4.3.1 Matriz de Confusão

Os primeiros critérios utilizados para avaliação dos modelos de aprendizado de máquina são os resultados da matriz de confusão. Essa matriz possui a responsabilidade de fornecer os valores de Verdadeiro Positivo (VP), Falso Positivo (FP), Falso Negativo (FN) e Verdadeiro Negativo (VN). Ao realizar o agrupamento dos valores em cada situação de acertos e erros da previsão, é possível calcular os valores dos erros (tipo 1 e tipo 2), e consequentemente, os valores de sensibilidade e de especificidade. Os resultados correspondentes à matriz de confusão estão disponíveis na Tabela 5.

Ao verificar os valores de erro tipo I do modelo que possui a *dummy* do σROA como *proxy* para o risco de falência, nota-se uma menor taxa para os modelos *random forest* e *naive bayes* (44,71%). Na perspectiva contrária, o SVM linear demonstrou um maior percentual (68,27%). Para o modelo que utiliza da *dummy* do Z'' Score como *proxy* para o risco de falência, houve uma diminuição considerável do erro tipo 1 para os nove modelos, pelo qual o modelo *random forest* recebeu a menor taxa (10,34%) e o SVM polinomial a maior taxa (22,07%). Barboza, Kimura e Altman (2017) também evidenciaram altas taxas de erro tipo I para os modelos SVMs, tanto do método linear quanto da base radial. Corroborando os achados desta dissertação, o modelo *random forest* também tem sido apontado pela literatura como um dos melhores modelos para previsão de falência (BARBOZA; KIMURA; ALTMAN, 2017; SUSS; TREITEL, 2019; PETROPOULOS *et al.*, 2020; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020; ARORA; SINGH, 2020; RAHAYU; SUHARTANTO, 2020).

Tabela 5 – Matriz de confusão da aplicação dos modelos preditivos

Variável dependente: σROA								
Modelo	**VP**	**FP**	**FN**	**VN**	**Erro TI**	**Erro TII**	**Sensibilidade**	**Especificidade**
K-NN	98	110	45	1149	52,88%	3,77%	47,12%	96,23%
Naive Bayes	115	93	102	1092	44,71%	8,54%	55,29%	91,46%
Logit	70	138	39	1155	66,35%	3,27%	33,65%	96,73%
Random Forest	115	93	27	1167	44,71%	2,26%	55,29%	97,74%
SVM Base Radial	75	133	21	1173	63,94%	1,76%	36,06%	98,24%
SVM Polinomial	76	132	21	1173	63,46%	1,76%	36,54%	98,24%
SVM Linear	66	142	34	1160	68,27%	2,85%	31,73%	97,15%
Bagging	107	101	30	1164	48,56%	2,51%	51,44%	97,49%
Boosting	94	114	33	1161	54,81%	2,76%	45,19%	97,24%
Variável dependente: Z'' Score								
Modelo	**VP**	**FP**	**FN**	**VN**	**Erro TI**	**Erro TII**	**Sensibilidade**	**Especificidade**
K-NN	302	56	28	1016	15,64%	2,68%	84,36%	97,32%
Naive Bayes	281	77	49	995	21,51%	4,69%	78,49%	95,31%
Logit	289	69	36	1008	19,27%	3,45%	80,73%	96,55%
Random Forest	321	37	23	1021	10,34%	2,2%	89,66%	97,8%
SVM Base Radial	308	50	17	1027	13,97%	1,63%	86,03%	98,37%
SVM Polinomial	279	79	9	1035	22,07%	0,86%	77,93%	99,14%
SVM Linear	285	73	28	1016	20,39%	2,68%	79,61%	97,32%
Bagging	314	44	34	1010	12,29%	3,26%	87,71%	96,74%
Boosting	313	45	26	1018	12,57%	2,59%	87,43%	97,41%

Fonte: Dados da pesquisa

O erro tipo II, que se refere ao falso negativo, recebeu menores percentuais de erros, comparado ao erro tipo I. O modelo que apresentou menores taxas de erro tipo II foi o SVM (polinomial ou de base radial), na aplicação dos dois cenários de variável dependente. O modelo que apresentou maiores taxas do erro tipo II foi o *naive bayes*, também para os dois cenários de variável dependente. As métricas de sensibilidade e especificidade indicam que os melhores modelos são os que apresentaram menores taxas de erros (I ou II), ou seja, são os modelos que já foram mencionados.

4.3.2 Acurácia, MAE e RMSE

As próximas métricas utilizadas para a mensuração de desempenho dos modelos são a acurácia, MAE e RMSE e estão disponíveis na Tabela 6. A acurácia é a porcentagem de classificações corretas, obtida a partir dos dados da Tabela 5, ou seja, dos valores de: Verdadeiro Positivo (VP), Falso Positivo (FP), Falso Negativo (FN) e Verdadeiro Negativo (VN). Quanto maior o valor da acurácia, melhor.

O modelo que apresentou maior precisão, nos dois casos de variável dependente, foi o *random forest*. Suss e Treitel (2019), Petropoulos *et al.* (2020) e Arora e Singh (2020) também constataram uma alta performance preditiva para o modelo *random forest*. Diferentemente, Teles *et al*. (2020) encontraram que o modelo SVM apresentou uma maior precisão preditiva,

mas alertaram que o *random forest* é um modelo que oferece baixo custo em termos de velocidade e simplicidade operacional.

Tabela 6 – Métricas de desempenho dos modelos preditivos

Variável dependente: σROA				
Modelo	**Acurácia**	**MAE**	**RMSE**	**Ranking**
K-NN	0,8894	0,1106	0,3325	6º
Naive Bayes	0,8609	0,1391	0,3729	9º
Logit	0,8738	0,1262	0,3553	8º
Random Forest	0,9144	0,0856	0,2926	1º
SVM RB	0,8902	0,1098	0,3314	5º
SVM Polinomial	0,8909	0,1091	0,3303	4º
SVM Linear	0,8745	0,1255	0,3543	7º
Bagging	0,9066	0,0934	0,3057	2º
Boosting	0,8951	0,1049	0,3238	3º
Variável dependente: Z'' Score				
Modelo	**Acurácia**	**MAE**	**RMSE**	**Ranking**
K-NN	0,9401	0,0599	0,2448	5º
Naive Bayes	0,9101	0,0899	0,2998	9º
Logit	0,9251	0,0749	0,2737	8º
Random Forest	0,9572	0,0427	0,2069	1º
SVM RB	0,9522	0,0478	0,2186	2º
SVM Polinomial	0,9372	0,0628	0,2505	6º
SVM Linear	0,928	0,0720	0,2684	7º
Bagging	0,9444	0,0556	0,2359	4º
Boosting	0,9486	0,0514	0,2266	3º

Fonte: Dados da pesquisa

O MAE é a medida que calcula o erro médio absoluto entre os valores reais e previstos. Já o RMSE, é a métrica que calcula a raiz quadrática média dos erros entre os valores reais e previstos. Como as duas medidas são relacionadas aos erros, quanto menor o valor, melhor. Para os dois casos que se diferem em termos de variável dependente, o MAE apresentou menor valor para o *random forest* e maior valor para o *naive bayes*.

Para o RMSE, a aplicação dos modelos com variável dependente σROA e Z'' Score resultou no menor valor para o modelo *random forest*. O maior valor de RMSE foi alocado no modelo *naive bayes* para as duas situações de dependência. Essa performance do *random forest* corrobora com os achados de Suss e Treitel (2019), Petropoulos *et al.* (2020), Arora e Singh (2020) e com as métricas anteriormente mencionadas.

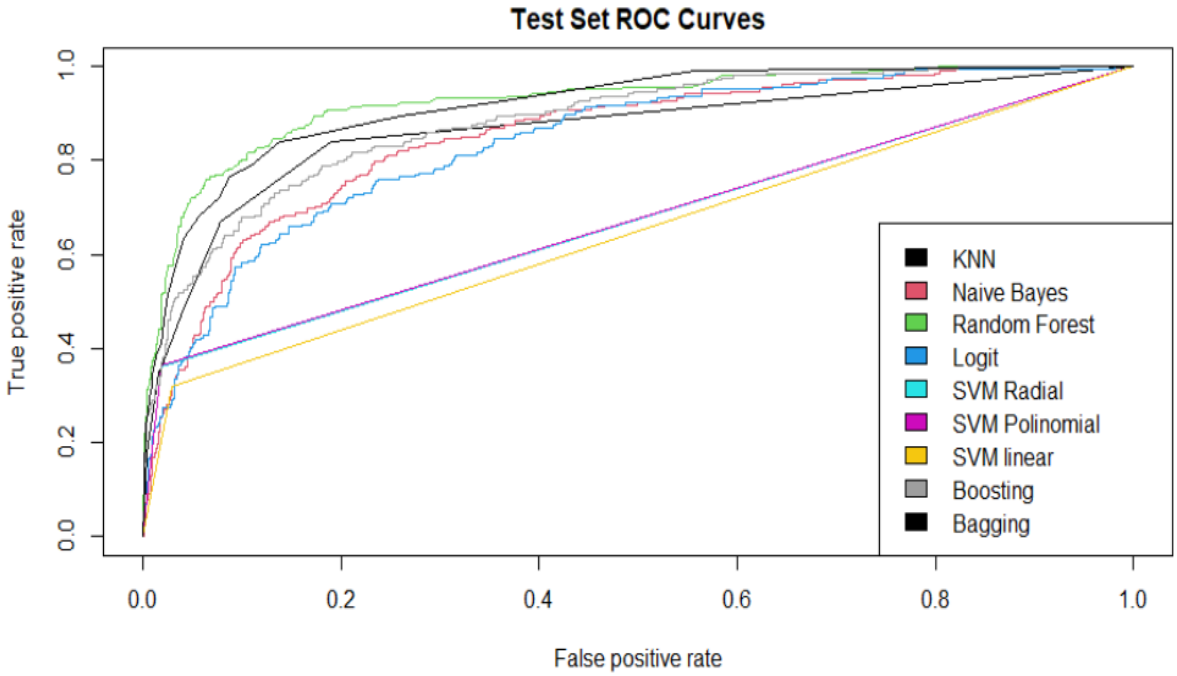
4.3.3 Curva ROC e AUC

A curva ROC ilustra o desempenho do classificador, demonstrando a taxa de verdadeiros positivos, à medida que a taxa de falsos positivos se altera. A tendência considerada

para este método é de que quanto maior, melhor. Nos Gráficos 6 e 7, estão exibidas as curvas para cada modelo de aprendizado de máquina utilizado nesta pesquisa.

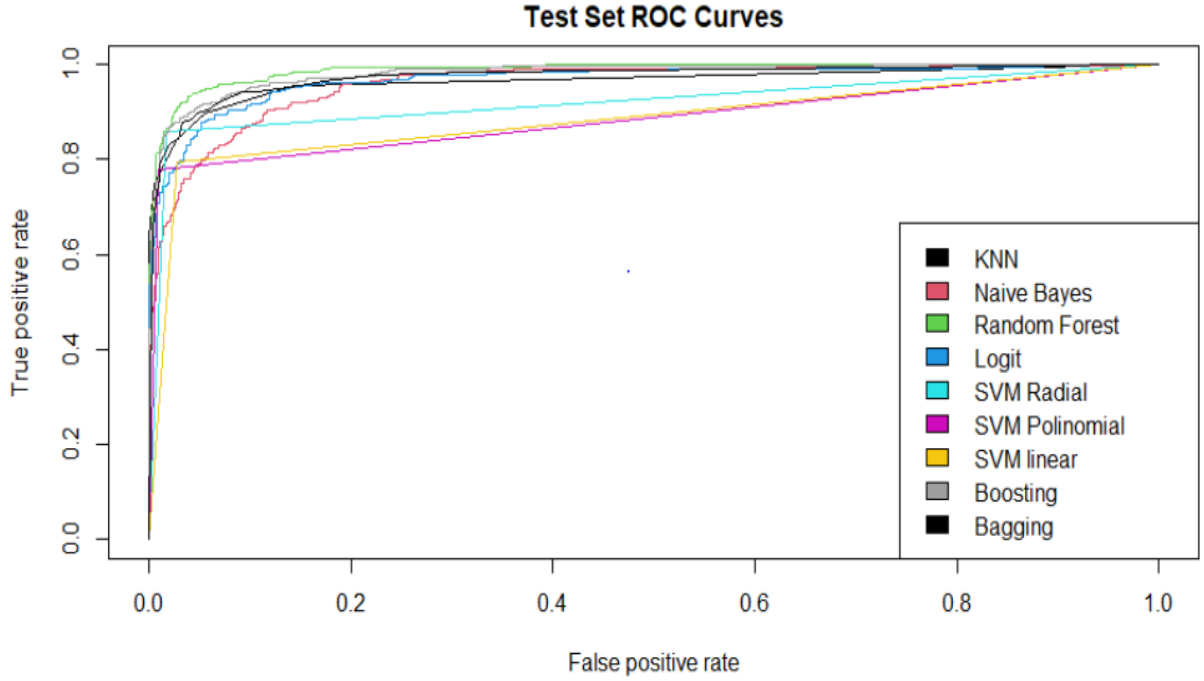
Tanto no Gráfico 6 quanto no Gráfico 7, o modelo com melhor resultado é o *random forest*. Ou seja, independente da *proxy* para falência utilizada, esse é o modelo mais recomendado para prever o risco de insolvência corporativa de empresas brasileiras com ações negociadas na [B]^{3}. Os modelos de aprendizado de máquina que apresentaram os piores resultados foram os SVMs. Esse resultado corrobora o observado na literatura, que tem sugerido o *random forest* como um dos melhores modelos na avaliação da curva ROC (BARBOZA; KIMURA; ALTMAN, 2017; VISWANATHAN; SRINIVASAN; HARIHARAN, 2020).

Gráfico 6 – Curva ROC da aplicação dos modelos com a variável dependente σROA



Fonte: Dados da pesquisa

Gráfico 7 – Curva ROC da aplicação dos modelos com a variável dependente Z'' Score



Fonte: Dados da pesquisa

Ressalta-se que as curvas ROC apresentaram melhores resultados na aplicação da predição com a variável dependente Z'' Score. Para verificar essa informação com mais profundidade, podem ser utilizados os valores AUC, visto que são fornecidos números a serem avaliados. Quanto maior o valor de AUC, melhor será o modelo, variando de 0 a 1. O modelo com AUC abaixo de 0,5 pode ser considerado ruim. Na Tabela 7, é possível visualizar os valores AUC de cada modelo de aprendizado de máquina.

Tabela 7 – Resultados do AUC (*area under the ROC curve*)

Variável dependente: σROA									
	K-NN	Naive Bayes	Logit	Random Forest	SVM RB	SVM Polinomial	SVM Linear	Bagging	Boosting
AUC	0,8656	0,8519	0,8388	0,923	0,6715	0,6739	0,6444	0,9175	0,8878
Ranking	4º	5º	6º	1º	8º	7º	9º	2º	3º
Variável dependente: Z'' Score									
	K-NN	Naive Bayes	Logit	Random Forest	SVM RB	SVM Polinomial	SVM Linear	Bagging	Boosting
AUC	0,9648	0,9603	0,9685	0,988	0,922	0,8854	0,8846	0,9478	0,9818
Ranking	4º	5º	3º	1º	7º	8º	9º	6º	2º

Fonte: Dados da pesquisa

Os modelos de aprendizado de máquina apresentaram resultados melhores com o uso da variável dependente Z'' Score. Mas, em todos os cenários, o AUC apresentou valor maior que 0,5, pelo qual representa uma boa predição dos modelos testados. Nos dois casos

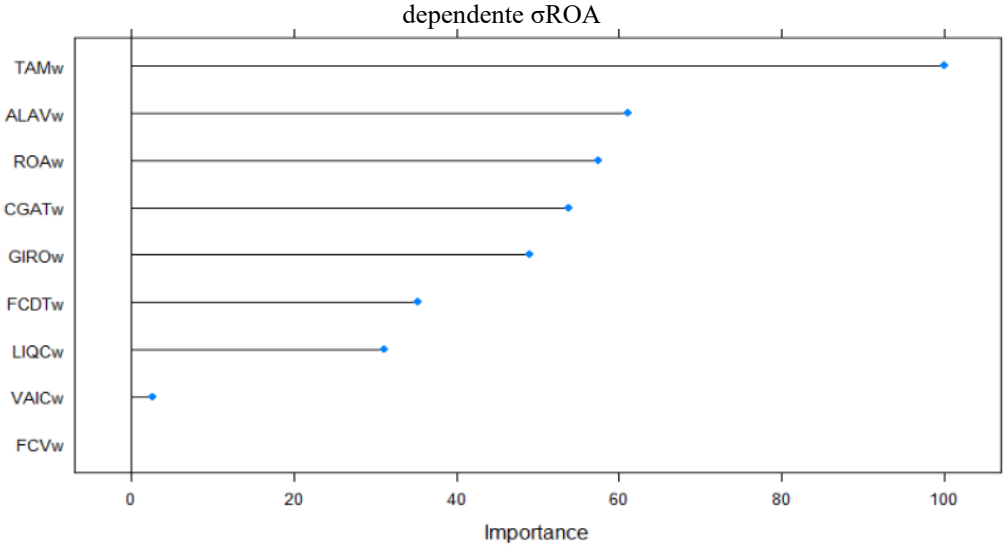
apresentados, o menor AUC foi para o SVM linear. Corroborando as evidências obtida neste trabalho, Barboza, Kimura e Altman (2017) também evidenciaram valores baixos para o SVM linear, sugerindo que ele não é o melhor modelo para realizar predições de risco de insolvência.

4.3.4 Importância das variáveis preditivas

O modelo que apresentou melhores resultados para os modelos preditivos foi o *random forest*. O algoritmo desse modelo permite verificar quais são as variáveis mais importantes para a predição, sendo visível nos Gráficos 8 e 9. O "w" no final da denominação de cada variável, representa a utilização de dados winsorizados. Para a previsão que possui como variável dependente o σROA, o TAM se mostrou mais importante. Corroborando esse achado, Chakraborty, Gao e Sheikh (2018) relatam que o tamanho da empresa pode ser um dos principais determinantes que influenciam o risco de insolvência corporativa.

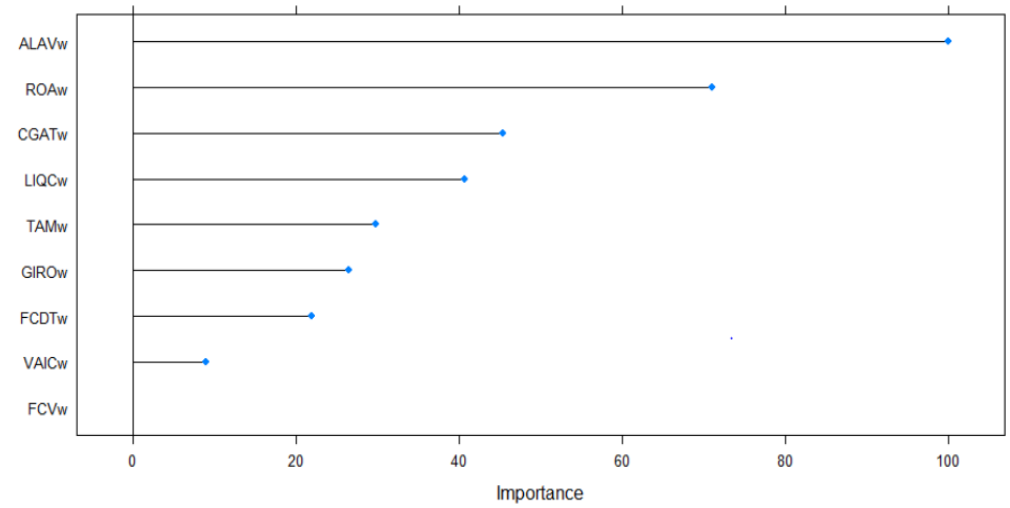
Para a predição que possui como variável dependente o Z'' Score, a ALAV demonstrou ser mais importante. Nos dois casos preditivos, entre as três variáveis mais importantes estão a ALAV e o ROA. A alavancagem financeira pode ser uma maneira de aumentar a lucratividade da empresa, mas, ao mesmo tempo, uma exposição ao risco de insolvência, podendo ser prejudicial ao funcionamento da organização, caso o endividamento não seja bem gerido. O ROA é uma medida que indica a performance da empresa, pelo qual a que apresenta maior rentabilidade pode possuir menor dependência de dívidas para financiar suas atividades, ou seja, pode gerar um menor endividamento. Se uma empresa possui valores pequenos ou negativos de ROA, sugere-se que pode existir um risco de insolvência corporativa que carece de atenção (MYERS; MAJLUF, 1984).

Gráfico 8 – Importância das variáveis preditivas na aplicação do modelo *random forest* com a variável dependente σROA



Fonte: Dados da pesquisa

Gráfico 9 – Importância das variáveis preditivas na aplicação do modelo *random forest* com a variável dependente Z'' Score



Fonte: Dados da pesquisa

Por fim, as duas variáveis menos importantes para o modelo *random forest*, independente da *proxy* utilizada para risco de falência, é o VAIC e o FCV. Divergente desse resultado, Hsu, Li e Fan (2006), identificaram que, para a realidade de Taiwan, o capital intelectual foi uma das variáveis mais significativas para a definição da situação financeira de uma empresa, analisando no horizonte de longo prazo. Para Pardo-Cueva, Herrera e Gómez (2018) o VAIC é um coeficiente capaz de transmitir informações à gerência sobre desempenho, valor e competitividade da organização, mas para detectar risco de insolvência em empresas brasileiras, foi constatado como a segunda variável menos importante deste estudo.

4.3.5 Análise das empresas financeiras

Diante das particularidades que existem nas organizações de natureza financeira, tais como normatização própria, supervisão por entidades reguladoras específicas, o dinheiro como a principal matéria-prima e concentração em operações de curto prazo (SILVA *et al.*, 2019), esta subseção tratou de realizar as análises do modelo preditivo apenas com empresas dessa tipologia. A filtragem feita resultou no total de 476 observações. O total de observações destinadas às aplicações de testes dos modelos foi 125.

Na Tabela 8, é possível visualizar o *matching* alcançado pelas duas medidas de risco. O *matching* entre as observações de menor risco continua sendo expressivo, coincidindo em 86,1% de observações. Para as observações de maior risco, o *matching* alcançado foi de 47,95%, representando uma pequena melhora em relação ao resultado da Tabela 4, para empresas de todos os setores, mas ainda é um resultado baixo.

Tabela 8 - *Matching* entre as variáveis de risco de insolvência

Risco	**Z'' Score**	**σROA**	***Matching***	***Matching* (%)**
Maior risco	73	91	35	47,95%
Menor risco	403	385	347	86,10%

Fonte: Dados da pesquisa

Os resultados da aplicação da matriz de confusão podem ser vistos na Tabela 9. Ao analisar os percentuais do erro tipo I dos modelos que possuem a *dummy* do σROA como *proxy* para risco de falência, a menor taxa foi destinada ao método *naive bayes* (30,43%). Já a maior taxa pertenceu ao SVM linear e ao *logit*, com valor igual (60,87%). No que concerne aos resultados dos modelos *naive bayes* e SVM linear, esses achados também foram detectados na aplicação para todo tipo de empresa, conforme Tabela 5. Para os modelos que utilizam da *dummy* do Z'' Score como *proxy* para risco de falência, o modelo SVM linear recebeu a menor taxa (22,22%), enquanto o SVM polinomial e o *logit* a maior taxa (50%).

O erro tipo II recebeu menores percentuais de erros, comparado ao erro tipo I. O modelo que apresentou menor taxa de erro tipo II foi o K-NN (1,04%), na aplicação do cenário do σROA como *proxy* para risco de falência. O modelo que apresentou maior taxa de erro tipo II foi o *naive bayes* (6,25%). Para o cenário do Z'' Score como *proxy* para risco de falência, o modelo que apresentou menor taxa de erro tipo II foi o SVM polinomial (2,97%), enquanto a maior taxa o SVM linear (6,93%). Para os percentuais de sensibilidade e especificidade, quanto maior, melhor, e os modelos mais bem avaliados por essas medidas são os que apresentaram menores taxas de erros (I ou II).

Tabela 9 – Matriz de confusão da aplicação dos modelos preditivos para empresas financeiras

Variável dependente: σROA								
Modelo	**VP**	**FP**	**FN**	**VN**	**Erro T1**	**Erro T2**	**Sensibilidade**	**Especificidade**
K-NN	11	12	1	95	52,17%	1,04%	47,83%	98,96%
Naive Bayes	16	7	6	90	30,43%	6,25%	69,57%	93,75%
Logit	9	14	4	92	60,87%	4,17%	39,13%	95,83%
Random Forest	10	13	3	93	56,52%	3,13%	43,48%	96,88%
SVM Base Radial	12	11	2	94	47,83%	2,08%	52,17%	97,92%
SVM Polinomial	14	9	2	94	39,13%	2,08%	60,87%	97,92%
SVM Linear	9	14	3	93	60,87%	3,13%	39,13%	96,88%
Bagging	11	12	5	91	52,17%	5,21%	47,83%	94,79%
Boosting	11	12	4	92	52,17%	4,17%	47,83%	95,83%
Variável dependente: Z'' Score								
Modelo	**VP**	**FP**	**FN**	**VN**	**Erro T1**	**Erro T2**	**Sensibilidade**	**Especificidade**
K-NN	10	8	5	96	44,44%	4,95%	55,56%	95,05%
Naive Bayes	9	9	6	95	50,00%	5,94%	50,00%	94,06%
Logit	14	4	5	96	22,22%	4,95%	77,78%	95,05%
Random Forest	11	7	4	97	38,89%	3,96%	61,11%	96,04%
SVM Base Radial	10	8	5	96	44,44%	4,95%	55,56%	95,05%
SVM Polinomial	9	9	3	98	50,00%	2,97%	50,00%	97,03%
SVM Linear	14	4	7	94	22,22%	6,93%	77,78%	93,07%
Bagging	11	7	5	96	38,89%	4,95%	61,11%	95,05%
Boosting	11	7	4	97	38,89%	3,96%	61,11%	96,04%

Fonte: Dados da pesquisa

Os valores de acurácia, MAE e RMSE, podem ser vistos na Tabela 10. Para os casos em que a variável dependente é o σROA, o modelo que apresentou maior acurácia foi o SVM Polinomial (0,9076) e uma menor acurácia foi para o *logit* (0,8487). Em relação aos casos que a variável dependente é o Z'' Score, o modelo que apresentou maior acurácia foi o *logit* (0,9244), enquanto a menor acurácia foi o modelo *naive bayes* (0,8739).

Corroborando os valores de acurácia, o menor valor MAE para os modelos com a variável dependente σROA e Z'' Score foi para o SVM polinomial (0,0924) e *logit* (0,0756), respectivamente. Na perspectiva contrária, o *logit* apresentou maior taxa de erro (0,1513), para a aplicação com a variável dependente σROA, e o *naive bayes* (0,1261) para a aplicação com a variável dependente Z'' Score. Os valores de RMSE seguem a mesma sequência de valores menores/maiores encontrado pelo MAE. Para fortalecer esses achados, nos Gráficos 10 e 11, estão exibidas as curvas ROC para cada modelo de aprendizado de máquina utilizado para observações de empresas financeiras.

De acordo com a curva ROC do Gráfico 10, o modelo com melhor resultado, na maior parte do tempo, é o *boosting*. Ou seja, para a variável dependente σROA, esse é o modelo mais recomendado para prever o risco de insolvência corporativa das empresas financeiras brasileiras com ações negociadas na [B] [3]. O modelo que apresentou o pior resultado foi o SVM linear. Já

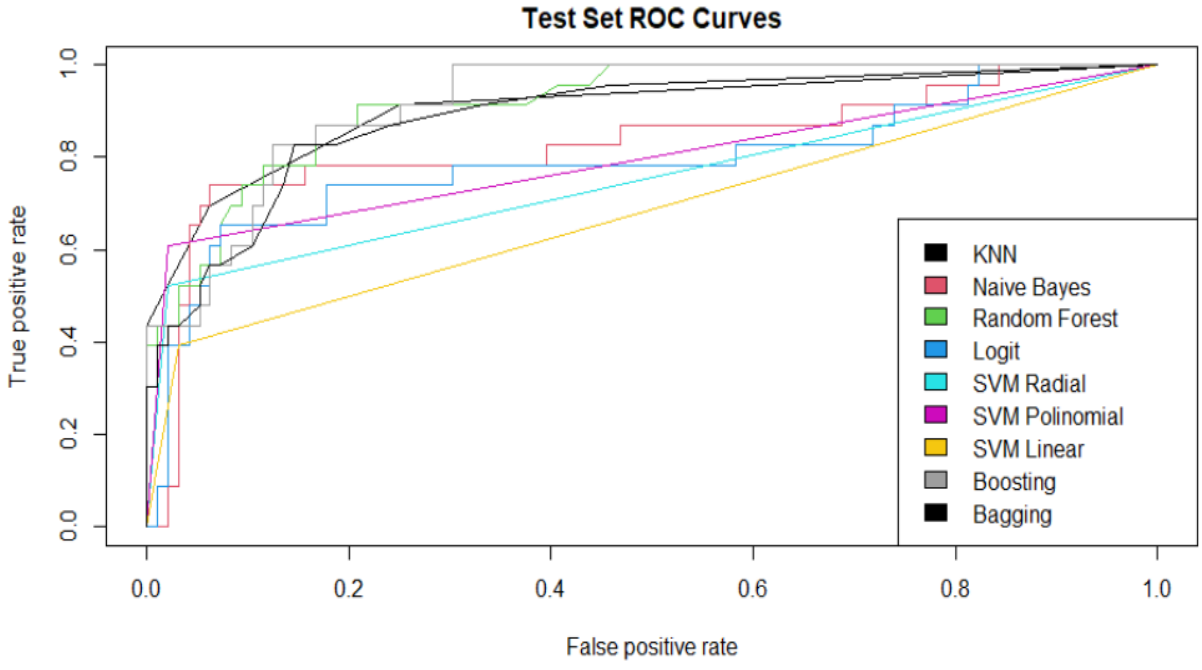
para a curva ROC do Gráfico 11, para modelos com variável dependente Z'' Score, o melhor modelo foi o *logit*, e o pior modelo foi o SVM polinomial.

Tabela 10 – Métricas de desempenho dos modelos preditivos para empresas financeiras

Variável dependente: σROA				
Modelo	**Acurácia**	**MAE**	**RMSE**	**Ranking**
K-NN	0,8908	0,1092	0,3305	2º
Naive Bayes	0,8908	0,1092	0,3305	2º
Logit	0,8487	0,1513	0,3889	9º
Random Forest	0,8655	0,1345	0,3667	5º
SVM RB	0,8908	0,1092	0,3305	2º
SVM Polinomial	0,9076	0,0924	0,3040	1º
SVM Linear	0,8571	0,1429	0,3780	7º
Bagging	0,8571	0,1429	0,3780	7º
Boosting	0,8655	0,1345	0,3667	5º
Variável dependente: Z'' Score				
Modelo	**Acurácia**	**MAE**	**RMSE**	**Ranking**
K-NN	0,8908	0,1092	0,3305	7º
Naive Bayes	0,8739	0,1261	0,3550	9º
Logit	0,9244	0,0756	0,2750	1º
Random Forest	0,9076	0,0924	0,3040	2º
SVM RB	0,8908	0,1092	0,3305	7º
SVM Polinomial	0,8992	0,1008	0,3176	5º
SVM Linear	0,9076	0,0924	0,3040	2º
Bagging	0,8992	0,1008	0,3176	5º
Boosting	0,9076	0,0924	0,3040	2º

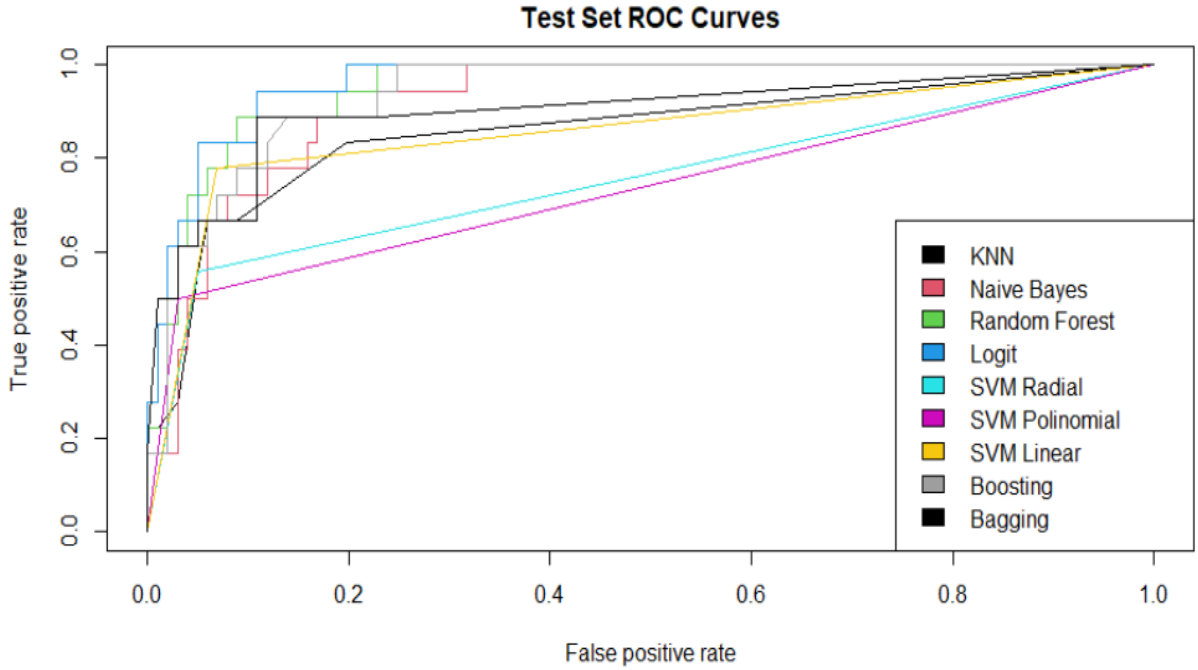
Fonte: Dados da pesquisa

Gráfico 10 – Curva ROC da aplicação dos modelos para a variável dependente σROA com observações de empresas financeiras



Fonte: Dados da pesquisa

Gráfico 11 – Curva ROC da aplicação dos modelos para a variável dependente Z'' Score com observações de empresas financeiras



Fonte: Dados da pesquisa

Para verificar os resultados transmitidos pelas curvas ROC com mais cautela, os valores de AUC podem ser visualizados na Tabela 11. Corroborando a análise feita a partir das curvas ROC, o *boosting* apresentou melhor resultado (0,9457) para predições com a variável dependente σROA e o SVM linear demonstrou o pior resultado (0,68). Para predições com a variável dependente Z'' Score, o *logit* apresentou o melhor resultado (0,962) e o SVM polinomial apresentou o pior resultado (0,7351).

Tabela 11 – Resultados do AUC (*area under the ROC curve*)

Variável dependente: σROA									
	K-NN	**Naive Bayes**	**Logit**	**Random Forest**	**SVM RB**	**SVM Polinomial**	**SVM Linear**	**Bagging**	**Boosting**
AUC	0,9035	0,8293	0,7899	0,9158	0,7505	0,7939	0,68	0,8893	0,9457
Ranking	3º	5º	7º	2º	8º	6º	9º	4º	1º
Variável dependente: Z'' Score									
	K-NN	**Naive Bayes**	**Logit**	**Random Forest**	**SVM RB**	**SVM Polinomial**	**SVM Linear**	**Bagging**	**Boosting**
AUC	0,8581	0,9197	0,962	0,9505	0,753	0,7351	0,8542	0,8999	0,9554
Ranking	6º	4º	1º	2º	8º	9º	7º	5º	3º

Fonte: Dados da pesquisa

Diante das particularidades que existem para as empresas do setor financeiro, o modelo que melhor prevê situações de risco de insolvência não é o mesmo para todo tipo de organização. Quando a análise preditiva englobar vários tipos de empresa, o *random forest* pode

demonstrar maior força preditiva. A partir do momento que as observações coletadas para realização de análises forem predominantemente de financeiras, o *boosting*, SVM polinomial ou o *logit* podem apresentar resultados mais robustos. Ressalta-se que, se a métrica decisiva para escolha do melhor modelo for o AUC, o *random forest* apresenta diferenças marginais comparadas aos melhores modelos, o que significa que pode ser utilizado em empresas financeiras sem perda de performance. Diante desses achados, pesquisadores e executivos serão capazes de identificar o risco de insolvência corporativa das organizações brasileiras com maior precisão (DUARTE; BARBOZA, 2020).

5 CONSIDERAÇÕES FINAIS

O gerenciamento do risco de insolvência corporativa é importante para o sucesso do conjunto de partes interessadas. Dessa forma, trabalhos que geram contribuições para essa área são sempre bem-vindos. Desde Altman (1968), a análise discriminante tem sido elencada como uma técnica adequada para separar empresas solventes de insolventes. Corroborando essa técnica, este estudo avaliou a efetividade de modelos de aprendizado de máquina na identificação do risco de insolvência corporativa, considerando o contexto do mercado acionário brasileiro. A avalição foi feita em dois momentos: primeiro com dados de empresas de todos os setores empresariais; e segundo considerando apenas dados de instituições financeiras, diante de suas especificidades.

O modelo Z'' Score de Altman, Hartzel e Peck (1998) demonstrou um bom alinhamento às predições dos métodos supervisionados, sendo um modelo já consolidado na literatura. Em comparação, o σROA demonstrou valores inferiores de alinhamento preditivo. Sobre os modelos de aprendizado supervisionado, o *random forest* predominantemente apresentou os melhores resultados. Quando a análise leva em consideração apenas dados de empresas financeiras, o melhor modelo preditivo passou a ser o SVM polimonial, o *logit* ou o *boosting*, dependendo do tipo da métrica de performance. Porém, se a métrica decisiva para escolha do melhor modelo for o AUC, o *random forest* apresenta diferenças marginais comparadas aos melhores modelos, o que significa que pode ser utilizado em empresas financeiras sem perda de performance.

Os resultados de efetividade dos modelos são considerados estratégicos para empresas, investidores e demais partes interessadas, pois medidas podem ser adotadas para minimizar o estado de maior risco de insolvência e para reconsiderar se os ativos são bons para alocar em carteiras de investimentos. Até então, o *random forest* pode ser considerado o modelo com melhor desempenho para realizar predições de risco de insolvência. Estudos futuros podem investigar o desempenho de modelos de aprendizado de máquina que não foram contemplados por esta pesquisa e compará-lo com o desempenho do *random forest*.

Para realizar boas predições de risco de insolvência corporativa, sugere-se não dispensar as variáveis alavancagem financeira e ROA, visto que apresentaram alto nível de importância no resultado preditivo do *random forest*. Ressalta-se que o aprendizado de máquina não é perfeito. A aplicação dos modelos SVMs (linear, polinomial e de base radial) apresentaram resultados baixos para o AUC, sendo os piores modelos para a variável dependente σROA. Dessa forma, há um alerta de que os modelos de aprendizado devem ser bem selecionados para

atender às finalidades pretendidas. Além disso, independente do modelo de aprendizado supervisionado, as maiores falhas preditivas encontram-se na identificação de empresas com maior risco de insolvência.

Apesar da acurácia para todos os casos testados ser maior que 84%, um elemento que pode ser elencado como limitação é o uso de poucas variáveis preditivas, pois há pesquisas que têm se preocupado em definir um portfólio amplo, incorporando mais de 25 atributos (STUPP; FLACH; MATTOS, 2018; VODA *et al.* 2021). Outra limitação está no foco em acertos que os modelos preditivos apresentam, deixando de lado o custo de operacionalização de cada modelo preditivo, conforme Teles *et al.* (2020).

Este estudo também não analisa indicadores qualitativos que podem surgir de relatórios administrativos e de governança corporativa ou de redes sociais de fácil investigação, como o Twitter, o que pode ser feito em pesquisa futuras. Além disso, recomenda-se adicionar outras variáveis preditivas, inclusive de natureza macroeconômica, como o PIB, crescimento da indústria, inflação, taxa de juros, e qualitativas, advindas de relatórios e publicações de especialistas, com o intuito de verificar se há melhorias significativas que justifiquem o incremento na quantidade de variáveis independentes. Paralelamente, se mostra oportuna a concentração de esforços para definir quais modelos oferecem menor custo em velocidade e maior simplicidade operacional.

Outro fator importante está na indecisão sobre o melhor modelo preditivo para empresas financeiras, que pode ser o SVM polimonial, o *logit*, o *boosting*, ou o *random forest*, dependendo da métrica a ser avaliada. Uma opção capaz de modificar esse cenário é o aprimoramento de modelos híbridos de aprendizado de máquina, que mescla características de dois ou mais modelos em um só e adiciona robustez às análises pretendidas, o que pode ser realizado em trabalhos futuros.

Espera-se que este estudo consiga encorajar parcerias entre estudiosos da área computacional e de finanças, para desenvolver sistemas preventivos do risco de insolvência, com o papel de alertar os gestores, investidores e *stakeholders* sobre a saúde financeira de empresas. Além disso, esses sistemas também poderiam fornecer conteúdos ligados à resolução de problemas financeiros.

Esta pesquisa englobou todas as empresas brasileiras de capital aberto, que possuem os dados disponíveis para realização da análise. A execução de predições que envolvem organizações de qualquer ramo proporciona a identificação de um modelo preditivo generalista, servindo de aplicação para os dados de qualquer espaço empresarial. Ao mesmo tempo, sugere-se a execução de pesquisas que tratem individualmente de cada ramo empresarial, pois aspectos

relacionados ao mercado em que a organização atua podem influenciar na detecção de risco de insolvência. Ademais, este estudo só compreende as empresas negociadas na [B]³. A maior parte das organizações brasileiras não pertencem à natureza de capital aberto e se encaixam nas classificações de pequenas e médias empresas. Essa limitação pode ser solucionada por estudos futuros que consigam os dados necessários de pequenas e médias empresas para a sua execução.

Esta dissertação é capaz de apoiar as decisões de investidores, a partir do momento que conseguem informações para definir racionalmente os ativos mais vantajosos. A classificação de empresas com maior/menor risco de insolvência pode servir de incremento no processo de avaliação de empresas, sendo um fator decisivo para a definição do preço-alvo e do valor comprovado do ativo.

Os achados contribuem para a realidade de executivos, pois eles se preocupam com a credibilidade de empresas. Quando as organizações são bem avaliadas no mercado, há grandes chances de obter melhores resultados corporativos. Executivos que não possuem uma boa avaliação na organização gerenciada podem estar perdendo o interesse de possíveis *stakeholders*, pelo qual estratégias precisam ser tomadas, na tentativa de mudar o olhar da sociedade perante à empresa. Ativos com menor risco de insolvência podem estar influenciando a geração de uma visão positiva e crível.

As instituições provedoras de crédito irão fornecer empréstimos e financiamentos com mais assertividade. A partir de uma previsão de insolvência corporativa com maior acurácia, a gestão de risco de crédito pode repensar sobre o conjunto de medidas adotadas para a diminuição de ameaças e para a projeção da empresa em tratar de possíveis prejuízos e riscos. Dessa forma, haverá garantia de tomadas de decisões mais seguras e maiores chances de ampliação do valor corporativo.

Por fim, os pesquisadores terão a oportunidade de comparar esses resultados com outras pesquisas; de apurar as informações que os novos achados fornecem; e de contribuir com melhorias para a predição do risco de insolvência com a adesão de novas pesquisas sobre a temática, a partir de sugestões de pesquisas futuras elencadas neste estudo. Além disso, outros *stakeholders* utilizarão dos achados baseados em sua necessidade.

REFERÊNCIAS

ABDULLAH, Mohammad. The implication of machine learning for financial solvency prediction: an empirical analysis on public listed companies of bangladesh. **Journal of Asian Business And Economic Studies**, p. 1-18, 30 jun. 2021. Emerald. http://dx.doi.org/10.1108/jabes-11-2020-0128.

ALTMAN, Edward L. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. **The Journal of Finance**, v. 23, n. 4, p. 589-609, set. 1968.

ALTMAN, Edward L. An emerging market credit scoring system for corporate bonds. **Emerging Markets Review**, v. 6, n. 4, p. 311-323, dez. 2005. Elsevier BV. http://dx.doi.org/10.1016/j.ememar.2005.09.007.

ALTMAN, Edward L.; HARTZELL, John; PECK, Matthew. Emerging market corporate bonds — a scoring system. **The New York University Salomon Center Series on Financial Markets and Institutions**, p. 391-400, 1998. Springer US. http://dx.doi.org/10.1007/978-1-4615-6197-2_25.

ANTUNES, Francisco; RIBEIRO, Bernardete; PEREIRA, Francisco. Probabilistic modeling and visualization for bankruptcy prediction. **Applied Soft Computing**, v. 60, p. 831-843, nov. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.asoc.2017.06.043.

ARDALAN, Behzad; ASKARIAN, Haleh. The impact of intellectual capital on the risk of financial distress of listed companies in Tehran stock exchange, Iran. **Indian Journal of Fundamental and Applied Life Sciences,** v. 4, p. 840-853, 2014.

ARORA, Isha; SINGH, Navjot. Prediction of Corporate Bankruptcy using Financial Ratios and News. **International Journal of Engineering and Management Research**, v. 10, n. 5, p. 82-87, 28 out. 2020. Vandana Publications. http://dx.doi.org/10.31033/ijemr.10.5.15.

ATIYA, A.F. Bankruptcy prediction for credit risk using neural networks: a survey and new results. **Ieee Transactions on Neural Networks**, v. 12, n. 4, p. 929-935, jul. 2001. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/72.935101.

AYYADEVARA, V Kishore. Random Forest. **Pro Machine Learning Algorithms**, p. 105-116, 2018. Apress. http://dx.doi.org/10.1007/978-1-4842-3564-5_5.

BAE, Jae Kwon. Predicting financial distress of the South Korean manufacturing industries. **Expert Systems with Applications**, v. 39, n. 10, p. 9159-9165, ago. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2012.02.058.

BARBOZA, Flavio; KIMURA, Herbert; ALTMAN, Edward. Machine learning models and bankruptcy prediction. **Expert Systems with Applications**, v. 83, p. 405-417, out. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2017.04.006.

BARROS, M. E.; MENEZES, J. T.; COLAUTO, R. D.; TEODORO, J. D. Gerenciamento de Resultados e Alavancagem financeira em Empresas Brasileiras de Capital Aberto. **Journal of Accounting, Management and Governance**, Brasília-DF, v. 17, n. 1, 2014. Disponível em: https://www.revistacgg.org/contabil/article/view/557. Acesso em: 5 out. 2021.

BRANDT, Valnir Alberto; MACHAIEWSKI, Stela; GEIB, Vanderleia. Capital intelectual e sua relação com os índices de rentabilidade de empresas do comércio varejista listadas na BM&FBOVESPA. **Base - Revista de Administração e Contabilidade da Unisinos**, v. 15, n. 4, p. 255-263, 31 dez. 2018. UNISINOS - Universidade do Vale do Rio Dos Sinos. http://dx.doi.org/10.4013/base.2018.154.01.

BRASIL. Altera e revoga dispositivos da Lei n. 6.404, de 15 de dezembro de 1976, e da Lei n. 6.385, de 07 de dezembro de 1976. **Lei Nº 11.638, de 28 de dezembro de 2007**. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/lei/l11638.htm. Acesso em: 10.06.2021.

BREIMAN, Leo. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123-140, ago. 1996. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/bf00058655.

BREIMAN, Leo. Arcing classifier (with discussion and a rejoinder by the author). **The Annals of Statistics**, v. 26, n. 3, p. 801-849, 1 jun. 1998. Institute of Mathematical Statistics. http://dx.doi.org/10.1214/aos/1024691079.

CAMBRONERO, Cristina García; MORENO, Irene Gómez. Algoritmos de aprendizaje: k-nn & k-means. **Inteligencia en Redes de Telecomuncicación**, 2006.

CARLEO, Giuseppe; CIRAC, Ignacio; CRANMER, Kyle; DAUDET, Laurent; SCHULD, Maria; TISHBY, Naftali; VOGT-MARANTO, Leslie; ZDEBOROVÁ, Lenka. Machine learning and the physical sciences. **Reviews of Modern Physics**, v. 91, n. 4, p. 1-39, 6 dez. 2019. American Physical Society (APS). http://dx.doi.org/10.1103/revmodphys.91.045002.

CHAKRABORTY, Atreya; GAO, Lucia; SHEIKH, Shahbaz. Corporate governance and risk in cross-listed and Canadian only companies. **Management Decision**, v. 57, n. 10, p. 2740-2757, 11 nov. 2019. Emerald. http://dx.doi.org/10.1108/md-10-2017-1052.

CHANG, Te-Min; HSU, Ming-Fu. Integration of incremental filter-wrapper selection strategy with artificial intelligence for enterprise risk management. **International Journal of Machine Learning and Cybernetics**, v. 9, n. 3, p. 477-489, 12 maio 2016. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s13042-016-0545-8.

CHEN, Chih-Chun; CHEN, Chun-Da; LIEN, Donald. Financial distress prediction model: the effects of corporate governance indicators. **Journal of Forecasting**, v. 39, n. 8, p. 1238-1252, 22 abr. 2020. Wiley. http://dx.doi.org/10.1002/for.2684.

CHU, Samuel Kai Wah; CHAN, Kin Hang; WU, Wendy W.y. Charting intellectual capital performance of the gateway to China. **Journal of Intellectual Capital**, v. 12, n. 2, p. 249-276, 19 abr. 2011. Emerald. http://dx.doi.org/10.1108/14691931111123412.

CIACCIO, Agostino di; CIALONE, Giovanni. Insolvency Prediction Analysis of Italian Small Firms by Deep Learning. **International Journal of Data Mining & Knowledge Management Process**, v. 9, n. 6, p. 1-12, 30 nov. 2019. Academy and Industry Research Collaboration Center (AIRCC). http://dx.doi.org/10.5121/ijdkp.2019.9601.

COURONNÉ, Raphael; PROBST, Philipp; BOULESTEIX, Anne-Laure. Random forest versus logistic regression: a large-scale benchmark experiment. **Bmc Bioinformatics**, v. 19, n. 1, p. 1-14, 17 jul. 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s12859-018-2264-5.

CUNNINGHAM, Pádraig; DELANY, Sarah Jane. k-Nearest Neighbour Classifiers - A Tutorial. **ACM Computing Surveys**, v. 54, n. 6, p. 1-25, jul. 2021. Disponível em: https://doi.org/10.1145/3459665.

DAMASCENO, Pedro Igor de Sousa. **Risco de insolvência e sentimento textual bancário**: uma análise dos bancos de capital aberto no brasil. 2021. 55 f. Dissertação (Mestrado) - Curso de Administração, Universidade Federal da Paraíba, João Pessoa, 2021.

DEVI, S. Sarojini; RADHIKA, Y. A Survey on Machine Learning and Statistical Techniques in Bankruptcy Prediction. **International Journal of Machine Learning and Computing**, v. 8, n. 2, p. 133-139, abr. 2018. EJournal Publishing. http://dx.doi.org/10.18178/ijmlc.2018.8.2.676.

DUARTE, Denize Lemos; BARBOZA, Flávio Luiz de Moraes. Forecasting Financial Distress With Machine Learning – A Review. **Future Studies Research Journal**: Trends and Strategies, v. 12, n. 3, p. 528-574, 1 set. 2020. Future Studies Research Journal: Trends and Strategies. http://dx.doi.org/10.24023/futurejournal/2175-5825/2020.v12i3.533.

DžENOPOLJAC, Vladimir; JANOŁEVIC, Stevo; BONTIS, Nick. Intellectual capital and financial performance in the Serbian ICT industry. **Journal of Intellectual Capital**, v. 17, n. 2, p. 373-396, 11 abr. 2016. Emerald. http://dx.doi.org/10.1108/jic-07-2015-0068.

ERDOGAN, Olcay; KONAKLI, Zafer. Corporate Credit Risk Assessment of BIST Companies. **European Scientific Journal, Esj**, v. 14, n. 1, p. 122, 31 jan. 2018. European Scientific Institute, ESI. http://dx.doi.org/10.19044/esj.2018.v14n1p122.

FIJAłKOWSKA, Justyna. Value Added Intellectual Coefficient (VAIC™) as a Tool of Performance Measurement. **Przedsiebiorczosc I Zarzadzanie**, v. 15, n. 1, p. 129-140, 1 jan. 2014. Walter de Gruyter GmbH. http://dx.doi.org/10.2478/eam-2014-0010.

FRIEDMAN, Jerome H. Greedy function approximation: a gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1-39, 1 out. 2001. Institute of Mathematical Statistics. http://dx.doi.org/10.1214/aos/1013203451.

GADGIL, Aashish A. Machine Learning based Intelligent Financial Crisis Prediction Models. **International Journal of Engineering Trends and Applications**: A Review, v. 8, n. 3, p. 30-35, 2021.

GENG, Ruibin; BOSE, Indranil; CHEN, XI. Prediction of financial distress: an empirical study of listed chinese companies using data mining. **European Journal of Operational Research**, v. 241, n. 1, p. 236-247, fev. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.ejor.2014.08.016.

GENTLEMAN, R.; CAREY, V. J. Unsupervised Machine Learning. **Bioconductor Case Studies**, p. 137-157, 2008. Springer New York. http://dx.doi.org/10.1007/978-0-387-77240-0_10.

GOMBOLA, Michael J.; HASKINS, Mark E.; KETZ, J. Edward; WILLIAMS, David D. Cash Flow in Bankruptcy Prediction. **Financial Management**, v. 16, n. 4, p. 55, 1987. Wiley. http://dx.doi.org/10.2307/3666109.

GREGOVA, Elena; VALASKOVA, Katarina; ADAMKO, Peter; TUMPACH, Milos; JAROS, Jaroslav. Predicting Financial Distress of Slovak Enterprises: comparison of selected traditional and learning algorithms methods. **Sustainability**, v. 12, n. 10, p. 3954, 12 maio 2020. MDPI AG. http://dx.doi.org/10.3390/su12103954.

HABIB, Ahsan; COSTA, Mabel D'; HUANG, Hedy Jiaying; BHUIYAN, Md. Borhan Uddin; SUN, Li. Determinants and consequences of financial distress: review of the empirical literature. **Accounting & Finance**, v. 60, n. 1, p. 1023-1075, 12 set. 2018. Wiley. http://dx.doi.org/10.1111/acfi.12400.

HSU, Kuang-Hua; LI, Jian-Fa; FAN, Hon-Jenq. An Application of Intellectual Capital on Financial Distress Models by Using Neural Network. **Proceedings of the 9Th Joint International Conference on Information Sciences (Jcis-06)**, v. 1, n. 1, 2006. Atlantis Press. http://dx.doi.org/10.2991/jcis.2006.150

JADHAV, Sayali D.; CHANNE, H. P. Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. **International Journal of Science and Research**, S.L., v. 5, n. 1, p. 1842-1845, jan. 2016.

JAN, Chyan-Long. Financial Information Asymmetry: using deep learning algorithms to predict financial distress. **Symmetry**, v. 13, n. 3, p. 443, 9 mar. 2021. MDPI AG. http://dx.doi.org/10.3390/sym13030443.

JIN, Hoon; HONG, Jeoung-Pyo; LEE, Kang-Ho; JOO, Dong-Won. Diagnosis of Corporate Insolvency Using Massive News Articles for Credit Management. **2019 Ieee International Conference on Big Data and Smart Computing (Bigcomp)**, p. 1-6, fev. 2019. IEEE. http://dx.doi.org/10.1109/bigcomp.2019.8679267.

KEYA, Maria Sultana; AKTER, Himu; RAHMAN, Md. Atiqur; RAHMAN, Md. Mahbobur; EMON, Minhaz Uddin; ZULFIKER, Md. Sabab. Comparison of Different Machine Learning Algorithms for Detecting Bankruptcy. In: 2021 6TH INTERNATIONAL CONFERENCE ON INVENTIVE COMPUTATION TECHNOLOGIES (ICICT), 6., 2021, **2021 6th International Conference on Inventive Computation Technologies (ICICT).** IEEE, 2021. p. 705-716.

LESÁKOVÁ, Ľubica; GUNDOVÁ, Petra; VINCZEOVÁ, Miroslava. The practice of use of models predicting financial distress in Slovak companies. **Journal of Eastern European and Central Asian Research (Jeecar)**, v. 7, n. 1, p. 122-136, 14 mar. 2020. Journal of Eastern European and Central Asian Research. http://dx.doi.org/10.15549/jeecar.v7i1.369.

LI, Hui; SUN, Jie; SUN, Bo-Liang. Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors. **Expert Systems with Applications**, v. 36, n. 1, p. 643-659, jan. 2009. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2007.09.038.

LIN, Fengyi; LIANG, Deron; CHEN, Enchia. Financial ratio selection for business crisis prediction. **Expert Systems with Applications**, v. 38, n. 12, p. 15094-15102, nov. 2011. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2011.05.035.

MANZANEQUE, Montserrat; MERINO, Elena; PRIEGO, Alba María. The role of institutional shareholders as owners and directors and the financial distress likelihood. Evidence from a concentrated ownership context. **European Management Journal**, v. 34, n. 4, p. 439-451, ago. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.emj.2016.01.007.

MYERS, Stewart C.; MAJLUF, Nicholas S. Corporate financing and investment decisions when firms have information that investors do not have. **Journal of Financial Economics**, v. 13, n. 2, p. 187-221, jun. 1984. Elsevier BV. http://dx.doi.org/10.1016/0304-405x(84)90023-0.

MUÑOZ-IZQUIERDO, Nora; LAITINEN, Erkki K.; CAMACHO-MIÑANO, María-Del-Mar; PASCUAL-EZAMA, David. Does audit report information improve financial distress prediction over Altman's traditional Z -Score model? **Journal of International Financial Management & Accounting**, v. 31, n. 1, p. 65-97, 19 set. 2019. Wiley. http://dx.doi.org/10.1111/jifm.12110.

MSELMI, Nada; LAHIANI, Amine; HAMZA, Taher. Financial distress prediction: the case of french small and medium-sized firms. **International Review of Financial Analysis**, v. 50, p. 67-80, mar. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.irfa.2017.02.004.

MUHAMEDYEV, Ravil I. Machine learning methods: an overview. **Computer Modelling & New Technologies**, v. 6, n. 19, p. 14-29, 2015.

NADEEM, Muhammad; SILVA, Tracy-Anne de; KAYANI, Umar Nawaz. Predicting corporate financial distress for New Zealand listed firms using intellectual capital indicators. **New Zealand Journal of Applied Business Research**, V. 14, N. 2, 2016, p. 1-15.

NAWAZ, Tasawar. Intellectual capital, financial crisis and performance of Islamic banks: Does shariah governance matter? **International Journal of Business and Society.** v. 18, n. 1, 2017, p. 211-226.

OHLSON, James A. Financial Ratios and the Probabilistic Prediction of Bankruptcy. **Journal of Accounting Research**, v. 18, n. 1, p. 109-131, 1980. JSTOR. http://dx.doi.org/10.2307/2490395.

ORESKI, Stjepan; ORESKI, Goran. Genetic algorithm-based heuristic for feature selection in credit risk assessment. **Expert Systems with Applications**, v. 41, n. 4, p. 2052-2064, mar. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2013.09.004.

PARDO-CUEVA, Mariuxi; HERRERA, Reinaldo Armas; GÓMEZ, Ángel Higuerey. La influencia del capital intelectual sobre la rentabilidad de las empresas manufactureras ecuatorianas. **Espacios**, v. 39, n. 51, p. 1-1, dez. 2018.

PETROPOULOS, Anastasios; SIAKOULIS, Vasilis; STAVROULAKIS, Evangelos; VLACHOGIANNAKIS, Nikolaos E.. Predicting bank insolvencies using machine learning techniques. **International Journal of Forecasting**, v. 36, n. 3, p. 1092-1113, jul. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.ijforecast.2019.11.005.

PISNER, Derek A.; SCHNYER, David M.. Support vector machine. **Machine Learning**, p. 101-121, 2020. Elsevier. http://dx.doi.org/10.1016/b978-0-12-815739-8.00006-7.

PULIC, Ante. VAIC™ an accounting tool for IC management. **International Journal of Technology Management**, v. 20, n. 5/6/7/8, 2000. Inderscience Publishers. http://dx.doi.org/10.1504/ijtm.2000.002891.

RAHAYU, Dyah Sulistyowati; SUHARTANTO, Heru. Ensemble Learning in Predicting Financial Distress of Indonesian Public Company. **2020 8Th International Conference on Information And Communication Technology (Icoict)**, p. 1-5, jun. 2020. IEEE. http://dx.doi.org/10.1109/icoict49345.2020.9166246

ROSA, Paulo Sérgio; GARTNER, Ivan Ricardo. Financial distress in Brazilian banks: an early warning model. **Revista Contabilidade & Finanças**, v. 29, n. 77, p. 312-331, 20 dez. 2017. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/1808-057x201803910.

SANTANA FILHO, Júlio César; OLIVEIRA, Elis Regina; SANTOS, Geovane Camilo; OLIVEIRA, Elcio Dihl. Análise dos índices de desempenho econômico-financeiro dos clubes de futebol do campeonato brasileiro de 2014 a 2018: antes e após o profut. **Brazilian Journal of Development**, v. 5, n. 7, p. 9733-9764, 2019. Brazilian Journal of Development. http://dx.doi.org/10.34117/bjdv5n7-149.

SCHONLAU, Matthias; ZOU, Rosie Yuyan. The random forest algorithm for statistical learning. **The Stata Journal**: Promoting communications on statistics and Stata, v. 20, n. 1, p. 3-29, mar. 2020. SAGE Publications. http://dx.doi.org/10.1177/1536867x20909688.

SEHGAL, Sanjay; MISHRA, Ritesh Kumar; DEISTING, Florent; VASHISHT, Rupali. On the determinants and prediction of corporate financial distress in India. **Managerial Finance**, v. 47, n. 10, p. 1428-1447, 5 maio 2021. Emerald. http://dx.doi.org/10.1108/mf-06-2020-0332.

SERASA EXPERIAN. **Indicadores econômicos**. 2021. Disponível em: https://www.serasaexperian.com.br/conteudos/indicadores-economicos/. Acesso em: 05 jun. 2021.

SERASA EXPERIAN. **Recuperação judicial tem queda de 15% em 2020, revela Serasa Experian**. 2021. Disponível em: https://www.serasaexperian.com.br/sala-de-imprensa/noticias/recuperacao-judicial-tem-queda-de-15-em-2020-revela-serasa-experian/. Acesso em: 05 jul. 2021.

SKURICHINA, Marina; DUIN, Robert P. W. Bagging, Boosting and the Random Subspace Method for Linear Classifiers. **Pattern Analysis & Applications**, v. 5, n. 2, p. 121-135, 7 jun. 2002. Disponível em: https://doi.org/10.1007/s100440200011. Acesso em: 18 jan. 2022.

SHAHWAN, Tamer Mohamed. The effects of corporate governance on financial performance and financial distress: evidence from egypt. **Corporate Governance**, v. 15, n. 5, p. 641-662, 5 out. 2015. Emerald. http://dx.doi.org/10.1108/cg-11-2014-0140.

SHAHWAN, Tamer Mohamed; HABIB, Ahmed Mohamed. Does the efficiency of corporate governance and intellectual capital affect a firm's financial distress? Evidence from Egypt. **Journal of Intellectual Capital**, v. 21, n. 3, p. 403-430, 5 mar. 2020. Emerald. http://dx.doi.org/10.1108/jic-06-2019-0143.

SHI, Yin; LI, Xiaoni. A bibliometric study on intelligent techniques of bankruptcy prediction for corporate firms. **Heliyon**, v. 5, n. 12, p. 1-12, dez. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.heliyon.2019.e02997.

SILVA, Cristiano Lima da; ALVES, Joyce Quintino; BRAGA, Oton Crispim; PEREIRA JÚNIOR, José Wellington; ANDRADE, Luiz Odorico Monteiro de; OLIVEIRA, Antônio Mauro Barbosa de. Usando o classificador naive bayes para geração de alertas de risco de óbito infantil. **Revista Eletrônica de Sistemas de Informação**, v. 16, n. 2, p. 1-16, 31 ago. 2017. IBEPES (Instituto Brasileiro de Estudos e Pesquisas Sociais). http://dx.doi.org/10.21529/resi.2017.1602004.

SILVA, Sabrina Espinele da; CAMARGOS, Marcos Antônio de; FONSECA, Simone Evangelista; IQUIAPAZA, Robert Aldo. Determinantes da necessidade de capital de giro e do ciclo financeiro das empresas brasileiras listadas na B3. **Revista Catarinense da Ciência Contábil**, v. 18, p. 1-17, 3 set. 2019. http://dx.doi.org/10.16930/2237-766220192842.

SKURICHINA, Marina; DUIN, Robert P. W. Bagging for linear classifiers. **Pattern Recognition**, v. 31, n. 7, p. 909-930, jul. 1998. https://doi.org/10.1016/s0031-3203(97)00110-6.

SKURICHINA, Marina; DUIN, Robert P. W. Bagging, Boosting and the Random Subspace Method for Linear Classifiers. **Pattern Analysis & Applications**, v. 5, n. 2, p. 121-135, 7 jun. 2002. https://doi.org/10.1007/s100440200011.

SUSS, Joel; TREITEL, Henry. Predicting bank distress in the UK with machine learning. **Staff Working Paper: Bank of England**, p. 1-45, out. 2019

STUPP, Diego Rafael; FLACH, Leonardo; MATTOS, Luísa Karam de. Analysis of the impact of adopting international accounting standards in predicting the insolvency of businesses listed on the BM&FBovespa brazilian stock exchange. **Race - Revista de Administração, Contabilidade e Economia**, v. 17, n. 2, p. 397-422, 28 ago. 2018. Universidade do Oeste de Santa Catarina. http://dx.doi.org/10.18593/race.v17i2.16094.

TARAN, Alina. A Critical Approach to the Corporate Insolvency in Romania. **Florya Chronicles of Political Economy**, v. 3, n. 1, p. 111-129, 2017.

TELES, Germanno; RODRIGUES, Joel J. P. C.; RABêLO, Ricardo A. L.; KOZLOV, Sergei A.. Comparative study of support vector machines and random forests machine learning algorithms on credit operation. **Software**: Practice and Experience, p. 1-9, 12 maio 2020. Wiley. http://dx.doi.org/10.1002/spe.2842.

THINH, Tran Quoc; TUAN, Dang Anh; HUY, Nguyen Thanh; THU, Tran Ngoc Anh. Financial distress prediction of listed companies – empirical evidence on the Vietnamese stock market. **Investment Management and Financial Innovations**, v. 17, n. 2, p. 377-388, 6 jul. 2020. LLC CPC Business Perspectives. http://dx.doi.org/10.21511/imfi.17(2).2020.29.

VIEIRA, Carlos André Marinho; SILVA, Mariana Câmara Gomes e; SILVA, Rafaela Rodrigues da; FLORÊNCIO, Débora Bezerra. Complexidade e Risco dos Conglomerados Financeiros Operantes no Brasil. **Revista Base (Administração e Contabilidade) da Unisinos**, S.I, v. 17, n. 2, p. 1-28, 2020.

VISWANATHAN, P. K.; SRINIVASAN, Suresh; HARIHARAN, N. Predicting Financial Health of Banks for Investor Guidance Using Machine Learning Algorithms. **Journal of Emerging Market Finance**, v. 19, n. 2, p. 226-261, 14. maio.2020. SAGE Publications. http://dx.doi.org/10.1177/0972652720913478.

VODA, Alina Daniela; DOBROTă, Gabriela; ȚÎRCă, Diana Mihaela; DUMITRAșCU, Dănuț Dumitru; DOBROTă, Dan. CORPORATE BANKRUPTCY AND INSOLVENCY PREDICTION MODEL. **Technological and Economic Development of Economy**, v. 27, n. 5, p. 1039-1056, 19 ago. 2021. Vilnius Gediminas Technical University. http://dx.doi.org/10.3846/tede.2021.15106.

WANG, Gang; CHEN, Gang; CHU, Yan. A new random subspace method incorporating sentiment and textual information for financial distress prediction. **Electronic Commerce Research and Applications**, v. 29, p. 30-49, maio 2018. Elsevier BV. http://dx.doi.org/10.1016/j.elerap.2018.03.004.

WANG, Lu; WU, Chong. Business failure prediction based on two-stage selective ensemble with manifold learning algorithm and kernel-based fuzzy self-organizing map. **Knowledge-Based Systems**, v. 121, p. 99-110, abr. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.knosys.2017.01.016.

ZHANG, Shichao. Cost-sensitive K-NN classification. **Neurocomputing**, v. 391, p. 234-242, maio 2020. Elsevier BV. http://dx.doi.org/10.1016/j.neucom.2018.11.101.

ZIEBA, Maciej; TOMCZAK, Sebastian K.; TOMCZAK, Jakub M. Ensemble boosting trees with synthetic features generation in application to bankruptcy prediction. **Expert Systems with Applications**, v. 58, p. 93-101, out. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2016.04.001.

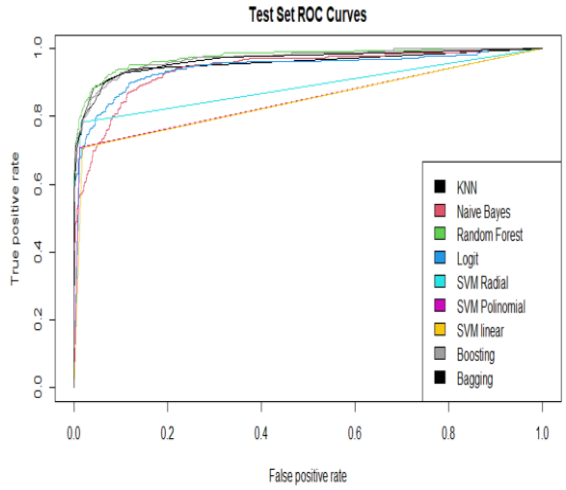
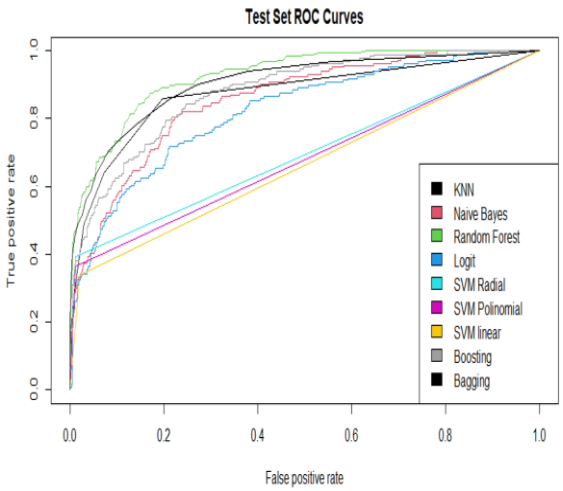
APÊNDICE A: Métricas de desempenho dos modelos preditivos com dados de 2010 a 2019

Tabela A1 – Desempenho dos modelos preditivos com dados com dados de 2010 a 2019

Risco	**Z'' Score**		**σROA**		***Matching***		***Matching (%)***					
Maior risco	1252		738		523		41,77%					
Menor risco	3747		4261		3532		94,26%					
Variável dependente: σROA												
Modelo	**VP**	**FP**	**FN**	**VN**	**ET1**	**ET2**	**Sensib.**	**Espec.**	**ACC**	**AUC**	**MAE**	**RMSE**
K-NN	90	94	32	1033	51,09%	3,00%	48,91%	97,00%	0,8991	0,8734	0,1009	0,3176
Naive Bayes	96	88	91	974	47,83%	8,54%	52,17%	91,46%	0,8567	0,8551	0,1433	0,3786
Logit	61	123	23	1042	66,85%	2,16%	33,15%	97,84%	0,8831	0,8214	0,1169	0,3419
Random Forest	90	94	18	1047	51,09%	1,69%	48,91%	98,31%	0,9103	0,9249	0,0896	0,2994
SVM Base Radial	72	112	13	1052	60,87%	1,22%	39,13%	98,78%	0,8999	0,6895	0,1001	0,3164
SVM Polinomial	67	117	12	1053	63,59%	1,13%	36,41%	98,87%	0,8967	0,6764	0,1033	0,3214
SVM Linear	62	122	22	1043	66,30%	2,07%	33,70%	97,93%	0,8847	0,6581	0,1153	0,3395
Bagging	95	89	27	1038	48,37%	2,54%	51,63%	97,46%	0,9071	0,9052	0,0929	0,3048
Boosting	83	101	30	1035	54,89%	2,82%	45,11%	97,18%	0,8951	0,8815	0,1049	0,3239
Variável dependente: Z'' Score												
Modelo	**VP**	**FP**	**FN**	**VN**	**ET1**	**ET2**	**Sensib.**	**Espec.**	**ACC**	**AUC**	**MAE**	**RMSE**
K-NN	248	65	17	920	20,77%	1,81%	79,23%	98,19%	0,9344	0,9569	0,0656	0,2561
Naive Bayes	227	86	55	882	27,48%	5,87%	72,52%	94,13%	0,8872	0,9395	0,1128	0,3359
Logit	228	85	23	914	27,16%	2,45%	72,84%	97,55%	0,9136	0,9440	0,0864	0,2939
Random Forest	259	54	18	919	17,25%	1,92%	82,75%	98,08%	0,9424	0,9759	0,0576	0,24
SVM Base Radial	245	68	13	924	21,73%	1,39%	78,27%	98,61%	0,9352	0,8844	0,0648	0,2546
SVM Polinomial	222	91	11	926	29,07%	1,17%	70,93%	98,83%	0,9184	0,8488	0,0816	0,2857
SVM Linear	222	91	17	920	29,07%	1,81%	70,93%	98,19%	0,9136	0,8456	0,0864	0,2939
Bagging	256	57	27	910	18,21%	2,88%	81,79%	97,12%	0,9328	0,9651	0,0672	0,2592
Boosting	252	61	25	912	19,49%	2,67%	80,51%	97,33%	0,9312	0,9698	0,0688	0,2623

Fonte: Dados da pesquisa

Figura A1 – Curva ROC dos modelos com dados de 2010 a 2019

Z'' Score	σROA
	

Fonte: Dados da pesquisa

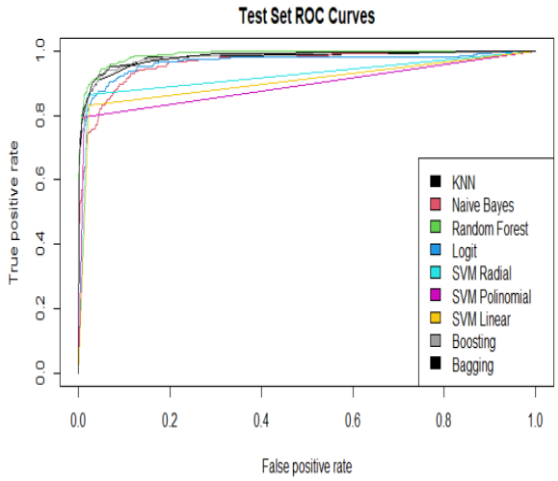
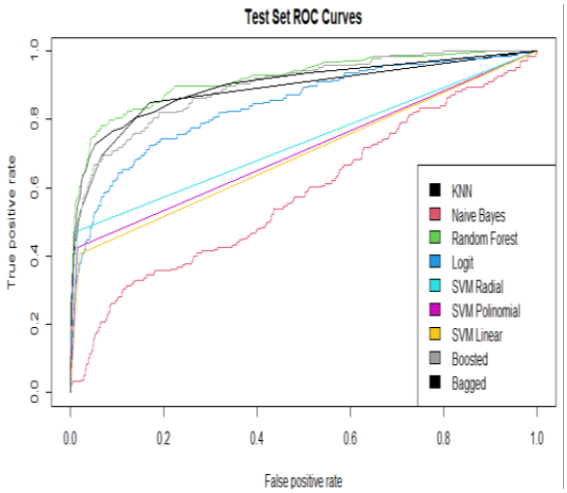
APÊNDICE B: Métricas de desempenho dos modelos preditivos sem empresas financeiras

Tabela B1 – Desempenho dos modelos preditivos sem empresas financeiras

Risco	**Zscore**			**DPROA**			***Matching***			***Matching (%)***		
Maior risco	136			751			560			41,18%		
Menor risco	3774			3583			3583			94,94%		
Variável dependente: σROA												
Modelo	**VP**	**FP**	**FN**	**VN**	**ET1**	**ET2**	**Sensib.**	**Espec.**	**ACC**	**AUC**	**MAE**	**RMSE**
K-NN	99	89	23	1073	47,34%	2,10%	52,66%	97,90%	0,9128	0,8843	0,0872	0,2953
Naive Bayes	114	74	96	1000	39,36%	8,76%	60,64%	91,24%	0,8676	0,8318	0,1324	0,3639
Logit	74	114	27	1069	60,64%	2,46%	39,36%	97,54%	0,8902	0,8406	0,1098	0,3314
Random Forest	105	83	15	1081	44,15%	1,37%	55,85%	98,63%	0,9237	0,9233	0,0763	0,2763
SVM Base Radial	89	99	17	1079	52,66%	1,55%	47,34%	98,45%	0,9097	0,7289	0,0903	0,3006
SVM Polinomial	80	108	17	1079	57,45%	1,55%	42,55%	98,45%	0,9026	0,7050	0,0974	0,3120
SVM Linear	76	112	22	1074	59,57%	2,01%	40,43%	97,99%	0,8956	0,6921	0,1044	0,3231
Bagging	107	81	23	1073	43,09%	2,10%	56,91%	97,90%	0,9190	0,8996	0,0810	0,2846
Boosting	100	88	27	1069	46,81%	2,46%	53,19%	97,54%	0,9104	0,8857	0,0896	0,2993
Variável dependente: Z'' Score												
Modelo	**VP**	**FP**	**FN**	**VN**	**ET1**	**ET2**	**Sensib.**	**Espec.**	**ACC**	**AUC**	**MAE**	**RMSE**
K-NN	292	48	18	926	14,12%	1,91%	85,88%	98,09%	0,9486	0,9789	0,0514	0,2267
Naive Bayes	268	72	41	903	21,18%	4,34%	78,82%	95,66%	0,9120	0,9636	0,0880	0,2967
Logit	284	56	25	919	16,47%	2,65%	83,53%	97,35%	0,9369	0,9670	0,0631	0,2512
Random Forest	299	41	17	927	12,06%	1,80%	87,94%	98,20%	0,9548	0,9892	0,0452	0,2125
SVM Base Radial	294	46	18	926	13,53%	1,91%	86,47%	98,09%	0,9502	0,9228	0,0498	0,2233
SVM Polinomial	270	70	12	932	20,59%	1,27%	79,41%	98,73%	0,9361	0,8907	0,0639	0,2527
SVM Linear	283	57	21	923	16,76%	2,22%	83,24%	97,78%	0,9393	0,9051	0,0607	0,2465
Bagging	299	41	25	919	12,06%	2,65%	87,94%	97,35%	0,9486	0,9813	0,0514	0,2267
Boosting	297	43	22	922	12,65%	2,33%	87,35%	97,67%	0,9494	0,9820	0,0506	0,2250

Fonte: Dados da pesquisa

Figura B1 – Curva ROC dos modelos preditivos sem dados de empresas financeiras

Z'' Score	σROA
	

Fonte: Dados da pesquisa