UNIVERSIDADE DE SÃO PAULO

FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, CONTABILIDADE E ATUÁRIA

DEPARTAMENTO DE ADMINISTRAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO

ANÁLISE DE FALÊNCIA UTILIZANDO IMAGENS E REDES NEURAIS

Luiz Wanderley Tavares

Orientador: Prof. Dr. José Afonso Mazzon.

SÃO PAULO

2024

Prof. Dr. Carlos Gilberto Carlotti Junior
Reitor da Universidade de São Paulo

Prof. Drª. Maria Dolores Montoya Diaz
Diretor da Faculdade de Economia, Administração, Contabilidade e Atuária

Prof. Dr. João Maurício Gama Boaventura
Chefe do Departamento de Administração

Prof. Dr. Felipe Mendes Borini
Coordenador do Programa de Pós-Graduação em Administração

LUIZ WANDERLEY TAVARES

ANÁLISE DE FALÊNCIA UTILIZANDO IMAGENS E REDES NEURAIS

Tese apresentada ao Programa de Pós-Graduação em Administração do Departamento de Administração da Faculdade de Economia, Administração, Contabilidade e Atuária da Universidade de São Paulo, como requisito parcial para a obtenção do título de Doutor em Ciências.

Área de Concentração: Administração.

Orientador: Prof. Dr. José Afonso Mazzon.

Versão corrigida

(versão original disponível na Biblioteca da Faculdade de Economia, Administração, Contabilidade e Atuária)

SÃO PAULO

2024

Tavares, Luiz Wanderley
ANÁLISE DE FALÊNCIA UTILIZANDO IMAGENS E REDES NEURAIS / Luiz Wanderley Tavares. -- São Paulo, 2024.
120 p.

Tese (Doutorado) – Universidade de São Paulo, 2024.
Orientador: José Afonso Mazzon.

1. Redes Neurais, Marketing Financeiro, Falência, Fintech.I. Universidade de São Paulo. Faculdade de Economia, Administração, Contabilidade e Atuária. II. Título.

“The neural network of your brain is the computer of your body, but it is also the computer of your life. It absorbs and registers every experience, however tiny, and compares it with past experiences, then stores it away. You can say, “Spaghetti again? We had it twice last week,” because your brain stores information by constantly comparing today with yesterday. At the same time, you develop likes and dislikes, grow bored, long for variety, and reach the end of one phase of your life, ready for the next. The brain enables it all to take place. It constantly connects new information with what you learned in the past. You remodel and refine your neural network on a second-by-second basis, but so does the world you experience. The largest super computer in existence cannot match this feat, which all of us take for granted” (Chopra e Tanzi, 2013).

AGRADECIMENTOS

Acolhimento.

Este é o sentimento mais presente neste trabalho.

Este trabalho marca o fim de um ciclo muito sonhado desde 1985 quando fui aprovado no vestibular para estudar Matemática no IME-USP. Foi o início do sonho acadêmico: graduação, mestrado e doutorado.

Porém, a necessidade de ir para o mercado de trabalho já nos primeiros anos da graduação mudou completamente o rumo. Entretanto, quando se quer algo, o universo conspira a favor, e após mais de duas décadas retornei para "casa". Fui aprovado para cursar o programa de pós-graduação da FEA-USP.

Quase 25 anos separaram a Graduação do Mestrado e mais uns pares de anos para chegar neste momento. Este somente possível, por eu ter sido acolhido por um dos melhores orientadores da FEA. Ele é uma pessoa de um conhecimento singular e com uma capacidade de ajudar indescritível. Nos momentos mais difíceis desta trajetória, eu tinha a certeza da necessidade em terminar, acima de tudo, por respeito a ele. Obrigado Prof. Dr. Mazzon, sem você não seria possível.

Também não seria possível, se eu não tivesse sido acolhido pelo amigo conhecido por intermédio da USP, Dr. Érico Azevedo. De nossa parceria vieram estudos acadêmicos e uma estrutura para obter os dados utilizados neste trabalho.

Durante estes anos, tive momentos bem difíceis, agravados pela pandemia mundial. Entretanto, os momentos piores foram sendo atenuados pelo acolhimento dado por uma mulher incrível, Paula Coutinho minha esposa, eu amo você. Com ela minha família aumentou, além de ser fã e pai do Matheus e do Felipe, meu núcleo familiar aumentou com os enteados, Gabriel e Manuela. Todos me acolheram nesta difícil fase, com afeto e carinho. Incrível como os filhos sempre nos ensinam a necessidade de continuar aprendendo.

Mas o acolhimento não estaria completo sem citar os fieis companheiros nos momentos “de solidão” da escrita; dois *pets*: uma *Golden* chamada Love e um *Shih Tzu* chamado Ozzy. Gostoso escrever com eles deitados do lado o tempo todo, como se estivessem ali me protegendo, de alguma forma estavam.

Muito obrigado por todos vocês terem sido tão acolhedores. Vocês fizeram todo o esforço desta minha caminhada valer a pena.

Amo todos vocês.

RESUMO

O marketing das instituições financeiras trabalha em criar produtos e serviços voltados às mais variadas necessidades das empresas e dos consumidores. Os serviços ofertados às empresas vão desde uma simples conta corrente com sistemas de pagamentos e cobranças até uma gama de créditos voltados a suportar e ampliar as operações das empresas de todos os portes. Ao analisar melhor estes produtos, é fácil identificar como as instituições financeiras segmentam o mercado, onde empresas de grande porte acabam recebendo uma maior fatia dos valores disponíveis para concessão de empréstimos. Isto ocorre devido ao menor risco de inadimplência e falência destas empresas, e por estas empresas serem um tipo de sociedade onde as informações são abertas, facilitando as análises de risco. Desde o final da década de 60, estudos vêm sendo realizados para identificar o risco de as empresas ficarem inadimplentes ou falirem. Estes estudos normalmente foram feitos sobre os dados das empresas abertas ou listadas em bolsa de valores, sabendo-se muito pouco sobre as empresas de pequeno e médio porte. Este estudo busca mostrar uma forma de analisar estas empresas de pequeno e médio porte através da utilização de redes neurais convolucionais (CNN). Através de uma base de dados contábeis de mais de 100 mil empresas, foi possível a geração de um modelo de *machine learning* para mensurar a probabilidade de empresas falirem ou ficarem inadimplentes. Os modelos CNN (*Convolutional Neural Networking*) têm capacidade de tratar séries-temporais analisando-as como imagens; sendo assim, os dados contábeis mensais foram transformados em imagens com a capacidade da CNN reconhecer padrões comportamentais das empresas. Foram criadas mais de 10.000 imagens e realizados 11 treinamentos para identificar situações onde a CNN consegue ser mais assertiva e com menos possibilidade de ocorrência de erros de treinamento comuns neste método. Os resultados obtidos evidenciam uma capacidade elevada de previsão de falência de empresas, em especial de pequeno e médio porte. A contribuição da tese é fundamentalmente de natureza metodológica e gerencial pela proposição e operacionalização de um método inovador em gestão de risco de falência de empresas.

Palavras-chave: Redes Neurais, Marketing Financeiro, Falência, *Fintech*.

ABSTRACT

The marketing strategies of financial institutions aim to create products and services tailored to the diverse needs of businesses and consumers. The services offered to companies range from simple checking accounts with payment and collection systems to a variety of credit options designed to support and expand the operations of businesses of all sizes. A closer analysis of these products reveals how financial institutions segment the market, with large enterprises receiving a larger share of available loan funds. This is due to their lower risk of default and bankruptcy, as well as the transparency of their information, which facilitates risk analysis. Since the late 1960s, studies have been conducted to identify the risk of businesses becoming delinquent or going bankrupt. These studies have typically focused on publicly traded companies, leaving a significant knowledge gap regarding small and medium-sized enterprises (SMEs). This study aims to demonstrate a method for analyzing SMEs using convolutional neural networks (CNNs). Utilizing a dataset of accounting records from over 100,000 companies, a machine learning model was developed to measure the likelihood of businesses failing or defaulting. CNN models have the capability to process time-series data by analyzing them as images. Therefore, monthly accounting data were transformed into images, enabling the CNN to recognize behavioral patterns of the companies. More than 10,000 images were created, and 11 training sessions were conducted to identify scenarios where the CNN could achieve higher accuracy and minimize common training errors associated with this method. The results obtained indicate a high predictive capability for business failures, especially among SMEs. The contribution of this thesis is fundamentally methodological and managerial, proposing and operationalizing an innovative method in the risk management of business failures.

Keywords: Neural Networks, Financial Marketing, Bankruptcy, Fintech.

SUMÁRIO

FIGURAS, QUADROS E TABELAS

1. INTRODUÇÃO

O mercado financeiro tem experimentado uma significativa transformação, marcada por uma crescente fragmentação que atrai a atenção tanto dos seus participantes quanto da esfera política. Esta fragmentação varia conforme o mercado financeiro específico e pode surgir por diversas razões, incluindo a regulamentação e a supervisão financeiras. As preocupações com a fragmentação incluem a eficiência na prestação de serviços financeiros, seus efeitos na transparência e na proteção dos consumidores e investidores, bem como sua relação com a estabilidade financeira global (Claessens, 2019).

Desde 2010, a fragmentação no mercado financeiro vem sendo impulsionada pelo Banco Mundial e pelo grupo dos G20, com um forte incentivo à inclusão financeira (Ozili, 2020). O Banco Mundial tem promovido a inclusão financeira digital nos países com economias emergentes, visando reduzir a desigualdade social e promover o crescimento econômico global.

A inclusão financeira digital tem sido implementada por meio da criação de bancos digitais e fintechs (tecnologia financeira), os quais promovem o crescimento econômico ao aumentar o volume de transações no sistema financeiro. No entanto, há preocupações sobre se suas atividades podem agravar crises econômicas em cenários de estresse de mercado (Ozili, 2018).

Embora os bancos digitais e as fintechs sejam semelhantes, há distinções importantes entre eles. Os bancos digitais oferecem serviços tradicionais, sendo obrigados a cumprir regulamentações e requisitos legais dos bancos centrais, enquanto as fintechs se concentram em soluções inovadoras para nichos específicos, operando sob regras menos rígidas. Ambos possuem uma vantagem competitiva ao adotar uma experiência de usuário centrada em dispositivos móveis, mas os bancos tradicionais estão acelerando a transformação digital, combinando novas iniciativas com parcerias

tecnológicas para oferecer serviços mais eficientes e seguros (Wewege e Thomsett, 2020).

Segundo Wewege e Thomsett (2020), as fintechs continuarão a impulsionar avanços tecnológicos e a experimentar novas tecnologias, focando na experiência do usuário e na proposta de valor ao cliente. Contudo, apesar do ambiente propício para a conquista de novos clientes no mundo digital, estudos como o de Severiano et al. (2021) revelam uma alta dependência de capital de terceiros e baixa rentabilidade em bancos digitais brasileiros.

O financiamento e os serviços digitais prometem impulsionar o mercado, oferecendo acesso a diversos produtos e serviços para indivíduos e pequenas e médias empresas. Contudo, a baixa rentabilidade observada parece indicar uma falta de capacidade para melhor análise de risco. A prevenção de riscos financeiros é detalhada na seção 2.2.1.

Inicialmente, o marketing bancário focava em atrativos como brindes e sorteios para captação de poupança ou abertura de contas correntes. Com o tempo, percebeu-se a necessidade de um sistema eficaz de planejamento e controle de marketing. O desenvolvimento do relacionamento e a fidelização dos clientes tornaram-se tendências importantes (Verona, 2004). A seção 2.2.2. detalha questões relacionadas ao marketing nas instituições financeiras.

Tanto os bancos tradicionais quanto os digitais continuam buscando aprimorar duas questões importantes: suas campanhas de marketing e a mensuração de riscos operacionais. Muitos estudos têm utilizado uma quantidade massiva de dados disponíveis nas instituições financeiras para analisar o comportamento dos consumidores. No contexto corporativo, Altman (1968) foi pioneiro ao estudar o risco de inadimplência empresarial, abrindo caminho para diversas pesquisas subsequentes. No Brasil, as pesquisas pioneiras realizadas pelo Prof. Stephen Charles Kanitz (Kanitz,

1976; 1978). Empresas de capital aberto são frequentemente utilizadas nesses estudos devido à facilidade de obtenção de dados. No entanto, poucos estudos focam na saúde financeira de pequenas e médias empresas devido à falta de informações públicas.

Mesmo assim, estudos como o de Hung et al. (2020) demonstram que a análise de big data pode melhorar a eficiência da gestão de risco e de campanhas de marketing, utilizando dados internos de instituições financeiras. Este estudo específico considerou dados de empresas clientes de um banco comercial, mostrando melhorias na gestão de risco e no desempenho de marketing por meio de um método de *machine learning* calcado em análise de imagens de dados financeiros de pequenas e médias empresas.

1.1. Problema da pesquisa

A análise de risco dos produtos financeiros leva em consideração a capacidade dos clientes honrarem com os pagamentos negociados. Quando Altman (1968) realizou sua pesquisa, seu objetivo era desenvolver uma forma de prever a adimplência das empresas. Antes de Altman, Myers (1963) já havia conduzido um estudo para calcular uma pontuação de credibilidade para pessoas físicas. Assim como Myers, todos os estudos anteriores e posteriores ao advento das técnicas de *machine learning* concentraram-se na análise de clientes pessoas físicas, utilizando dados como sexo, quantidade de dependentes, estado civil, tempo de moradia, ocupação, renda mensal, entre outros.

Altman (1968) conduziu seu estudo observando o balanço anual das empresas e calculando os índices financeiros correspondentes. Ele iniciou sua pesquisa com sessenta e seis empresas de diferentes setores, eliminou as pequenas empresas e trabalhou apenas com cinco índices financeiros. Assim como Altman, outros pesquisadores enfrentaram dificuldades na obtenção de dados financeiros das empresas. Os dados disponíveis geralmente são de empresas listadas nas bolsas de valores, que

são obrigadas a divulgar seus respectivos demonstrativos financeiros. Em geral, existem dados trimestrais de três a cinco anos, o que pode caracterizar uma base de dados qualificada para análise, dependendo, contudo, da quantidade de empresas analisadas e da metodologia empregada.

No entanto, ainda parece existir uma limitação quanto ao tipo e tamanho das empresas analisadas. Em agosto de 2023, existiam 23.928.954 empresas ativas no Brasil (site Sebrae, novembro de 2023). De acordo com dados da Receita Federal do Brasil (RFB), do total de estabelecimentos registrados até 2023, 13,1% correspondem a "Outros" (3.129.555 estabelecimentos), 52,7% correspondem a Microempreendedor Individual (MEI) (12.604.595 estabelecimentos), 29,3% correspondem a Microempresa (ME) (6.999.225 estabelecimentos) e 5% correspondem a Empresa de Pequeno Porte (EPP) (1.195.579 estabelecimentos) (site Receita Federal do Brasil, novembro de 2023).

As corporações de porte significativo, listadas ou não nas bolsas, sempre foram analisadas por meio de seus balanços. As empresas do tipo MEI são frequentemente avaliadas com base nos dados de seu único proprietário. No entanto, as pequenas e médias empresas caem em uma área cinzenta, onde os bancos possuem pouca informação e acabam fornecendo produtos financeiros "genéricos", sem considerar a real situação dessas empresas. Existem mais de 8 milhões de empresas nessas condições.

Nas duas primeiras décadas do século XXI, foram produzidos pouco menos de 3 mil estudos utilizando *machine learning* para prever inadimplência e falência de empresas, contrastando com os 2 mil estudos realizados desde 2021. Esses trabalhos utilizaram dados de balanços ou índices financeiros de companhias de ampla envergadura. Embora a contabilidade siga regras bem definidas, a questão a ser

colocada é: seria adequado utilizar os modelos criados para analisar grandes empresas, na análise de pequenas e médias empresas?

Os estudos realizados demonstram o potencial do aprendizado de máquina para a análise de organizações de alta relevância, sugerindo que esses métodos poderiam ser aplicáveis a outras empresas. No entanto, analisando os métodos utilizados, pode-se identificar lacunas nos dados e deficiências em alguns métodos devido à forma como são alimentados. Assim, coloca-se outra questão: a formatação dos dados de maneira diferente poderia melhorar a eficiência dos métodos de aprendizado de máquina?

A análise relativamente simplista do risco de crédito das pequenas e médias empresas pelas instituições financeiras apresenta como resultante um tratamento uniforme, com ofertas de taxas e serviços que não refletem a real situação econômico-financeira dessas empresas. Disso decorre uma estratégia de marketing indiferenciado por parte de bancos para esse segmento de mercado. Desse modo, desenvolver um método para analisar a situação dessas empresas permitiria que as instituições financeiras oferecessem taxas e serviços diferenciados, baseados em uma avaliação mais precisa de tomada de risco. Isso se torna possível graças ao aumento da capacidade de processamento dos computadores para o tratamento simultâneo de dados estruturados e não estruturados (Wang et al., 2020). Sem a evolução da capacidade de processamento, este trabalho não seria viável.

1.2. Objetivo da Pesquisa

Este trabalho visa desenvolver uma metodologia inovadora para avaliar a saúde financeira de empresas de pequeno e médio porte utilizando dados financeiros extraídos dos sistemas contábeis, com o objetivo de prever a inadimplência ou falência dessas empresas com base em informações financeiras recentes.

A metodologia proposta envolve a criação de séries temporais a partir dos balancetes mensais das empresas. Esses balancetes, que registram lançamentos mensais nas contas contábeis, permitem monitorar as variações das contas e calcular os índices financeiros mensalmente.

Nos últimos 15 anos, a pesquisa tem se concentrado na criação de modelos preditivos de aprendizado de máquina utilizando dados anuais. Métodos como *random forest* têm sido amplamente bem-sucedidos. No entanto, enquanto o balanço de uma empresa oferece uma visão estática, os balancetes mensais proporcionam uma perspectiva dinâmica, semelhante a uma "animação" financeira.

Embora os métodos tradicionais de séries temporais possam ser aplicados com sucesso relativo, eles ainda exigem um tratamento mínimo dos dados. As redes neurais, conhecidas por sua eficiência no processamento de imagens, oferecem uma abordagem promissora para o tratamento de séries temporais de dados financeiros.

Assim, o objetivo desta pesquisa é desenvolver um modelo que converta dados financeiros em imagens para alimentar uma rede neural, possibilitando a previsão da probabilidade de inadimplência ou falência de empresas de pequeno e médio porte. Este modelo pretende explorar a capacidade das redes neurais de identificar padrões complexos em dados financeiros visualmente representados. Portanto, este estudo pretende evidenciar que a conversão de dados financeiros em imagens, seguida de análise por redes neurais, pode melhorar a precisão na previsão de problemas financeiros futuros em empresas de menor porte. Com isso, a pesquisa visa expandir o conhecimento sobre a aplicação de aprendizado de máquina na análise financeira, introduzindo uma técnica inovadora para a geração e utilização de imagens a partir de dados financeiros.

2. FUNDAMENTAÇÃO TEÓRICA

Teorias sobre novas formas de utilizar as informações econômico--financeiras das empresas para analisar os riscos de elas honrarem com seus compromissos vêm crescendo nos últimos anos com a utilização das técnicas de aprendizado de máquina. Novos algoritmos vêm sendo criados e melhorados a cada ano, gerando novas possibilidades de utilização. Na área de análise da saúde financeira das empresas, os métodos dão suporte para o marketing das instituições financeiras criarem taxas e produtos para os clientes em situações diferenciadas.

Neste capítulo estão listados os métodos atuais de aprendizado de máquina aplicados nos estudos de risco de concordata ou falência de empresas, assim como os índices financeiros utilizados como treinamento para os algoritmos preverem a saúde financeira das empresas.

2.1 Aprendizado de Máquina (*Machine Learning*): Fundamentos e Aplicações

Machine Learning (ML) ou Ensinamento de Máquina é uma subárea da Inteligência Artificial (IA) focada no desenvolvimento de algoritmos para permitir os computadores aprenderem com base em um conjunto de dados. Em vez de serem explicitamente programadas para realizar uma tarefa, as máquinas utilizam estes algoritmos para identificar padrões nos dados e tomar decisões com base nesses padrões (Mahesh, B., 2020).

2.1.1 Tipos de Aprendizado

O aprendizado de máquina é basicamente dividido em 3 tipos, o supervisionado, o não supervisionado e o de reforço.

2.1.1.1 Aprendizado Supervisionado

O aprendizado supervisionado é uma abordagem de *machine learning* onde os modelos são treinados usando um conjunto de dados rotulado (Kotsiantis et al., 2007). Sendo assim, para cada entrada no conjunto de dados, a saída correta (ou rótulo) é conhecida. O principal objetivo desse tipo de aprendizado é fazer um modelo para generalizar e prever a saída de novas entradas não vistas anteriormente a partir dos dados de treinamento.

Componentes principais do aprendizado supervisionado:

- Entrada (*Features* ou Atributos): São as variáveis independentes alimentadas no modelo. Por exemplo, um banco pode criar um modelo para prever o limite de crédito dos correntistas baseado nos dados pessoais e históricos de movimentações.
- Saída (*Labels* ou Rótulos): É a variável dependente a ser prevista pelo modelo. No exemplo anterior, a saída seria o limite de crédito.
- Modelo: É o algoritmo com a função de aprender o mapeamento entre as entradas e saídas. Existem dois modelos, o de regressão e e o de classificação. Este tem como objetivo a previsão de uma classe ou categoria. Por exemplo, determinar se um e-mail é spam ou não é uma tarefa de classificação binária (Dietterich, 1998). Se existem mais de duas classes (como classificar imagens de frutas como maçãs, bananas ou cerejas), é uma classificação multiclasse. Os métodos de regressão buscam prever uma variável dependente (ou variável resposta) com base em uma ou mais variáveis independentes (ou preditoras). A relação entre a variável dependente e as variáveis independentes é modelada matematicamente, permitindo fazer previsões sobre os dados (Choudhary e Gianey, 2017).

- Função de Custo (ou Função de Perda): São funções utilizadas para quantificar quanto o erro das previsões de um modelo dos rótulos verdadeiros e fornece uma medida do desempenho do modelo. O objetivo durante o treinamento é minimizar o valor da função escolhida para analisar o modelo, ou seja, ajustar o modelo para as suas previsões ficarem o mais próximo possível dos valores reais (Burges et al., 2006)

Algoritmos comumente utilizados no aprendizado supervisionado:

- Regressão linear: Choudhary e Gianey (2017) descrevem em sua revisão, "Em termos mais simples, podemos dizer que na regressão linear adicionamos os insumos multiplicados por algumas constantes para obter o resultado. Cria uma correlação entre Y, uma variável dependente, e X, que podem ser múltiplas variáveis independentes, usando uma linha reta (linha de regressão)."
- Regressão logística: Método utilizado para modelar a probabilidade de uma variável de resposta binária (por exemplo, sim/não, verdadeiro/falso) com base em uma ou mais variáveis preditoras (podem ser categóricas, contínuas ou uma combinação de ambas). Este método é frequentemente utilizado para problemas de classificação binária (Dreiseitl e Ohno-Machado, 2002).
- Máquinas de vetores de suporte (SVM): Ferramenta de previsão baseada na teoria de aprendizado de máquina para maximizar a precisão preditiva enquanto evita automaticamente o ajuste excessivo aos dados. Máquinas de vetores de suporte podem ser definidas como sistemas usuários de funções lineares em um espaço de características de alta dimensão, treinado com um algoritmo de aprendizagem da teoria da otimização onde implementa um

viés de aprendizagem derivado da teoria de aprendizagem estatística (Jakkula, 2006).

- Árvores de decisão: Este mecanismo de regressão funciona dividindo um conjunto de dados em subconjuntos de dados menores e, posteriormente, a árvore de decisão relacionada é desenvolvida de forma incremental. Finalmente, é obtida uma árvore com "nós de decisão" e "nós folha". A árvore tem um nó raiz como o nó de decisão superior correspondente ao melhor preditor (Murthy, 1998). Uma árvore de decisão geralmente é construída particionando recursivamente o conjunto de treinamento, de modo que possamos rapidamente obter subconjuntos que estão mais ou menos na mesma classe (Wilton et al., 2022)
- *Random forest*: Florestas aleatórias são uma combinação de preditores de árvores onde cada árvore depende dos valores de um vetor aleatório amostrado de forma independente e com a mesma distribuição para todas as árvores da floresta. O erro de generalização converge até um limite conforme o aumento do número de árvores na floresta. (Breiman, 2001).
- *Naïve Bayes*: É uma técnica de classificação baseada no Teorema de Bayes com pressuposto de independência entre preditores. Para o classificador *Naïve Bayes* a presença de um recurso específico em uma classe não está relacionada à presença de qualquer outro recurso. *Naïve Bayes* tem como alvo principal a indústria de classificação de texto. É usado principalmente para fins de agrupamento e classificação, dependendo de a probabilidade condicional de uma variável acontecer (Mahesh, 2020).
- *LASSO:* é a abreviação de *least absolute shrinkage and selection operator* (operador de redução e seleção mínima absoluta) e é um método de análise

de regressão para realizar seleção e regularização de variáveis, a fim de melhorar a precisão da previsão e a interpretabilidade do modelo estatístico resultante (Hebiri e Lederer, 2012).

- Redes Neurais: Thakur e Konde (2021) descrevem em seu artigo "são sistemas baseados em algoritmos inspirados nas Redes Neurais Biológicas (BNNs). As ANNs oferecem soluções robustas para problemas em diversas áreas, incluindo classificação, previsão, filtragem, otimização, reconhecimento de padrões e aproximação de funções. O sistema nervoso biológico genuíno é extremamente complicado; os algoritmos de redes neurais artificiais buscam abstrair essa complexidade e focar no que pode teoricamente ser mais importante do ponto de vista do processamento de informações."

Desafios no Aprendizado Supervisionado:

- *Overfitting*: Ocorre quando um modelo é excessivamente complexo e começa a memorizar os dados de treinamento em vez de generalizar a partir deles e não consegue mais melhorar sua habilidade de resolver problemas (Jabbar, e Khan, 2015).
- *Underfitting*: Isso ocorre quando o modelo é incapaz de capturar a variabilidade dos dados (Jabbar, e Khan, 2015).
- Viés e Variância: Equilibrar a complexidade do modelo com a quantidade e qualidade dos dados é fundamental. "Um modelo com alta variância e baixo viés é "muito complexo", o que significa que ele aprende padrões "ruidosos" nos dados de treinamento e, assim, não consegue capturar a verdadeira relação por trás dos dados, gerando *overfitting*" (Guan e Burton, 2022, December).

- Dados Rotulados: Em muitos casos, obter um vasto conjunto de dados rotulados pode ser caro, demorado e difícil. Às vezes, é difícil coletar dados rotulados suficientes para treinar um modelo eficaz, especialmente em áreas de aplicação especializadas (Zhou, 2018).

2.1.1.2 Aprendizado Não Supervisionado

Em 2020, Mahesh explica a aprendizagem não supervisionada em contraposição à aprendizagem supervisionada por não haver respostas corretas e não existir "professor para ensinar". Assim, os algoritmos são deixados por conta própria para descobrir e apresentar a melhor estrutura dos dados. Os algoritmos de aprendizagem não supervisionados aprendem poucos recursos dos dados. Quando novos dados são introduzidos, ele utiliza os recursos aprendidos anteriormente para reconhecer a classe dos dados. É usado principalmente para *clustering* e redução de recursos.

No aprendizado não supervisionado, os principais componentes e técnicas incluem:

- Dados não rotulados: Como já explicado, a principal característica do aprendizado não supervisionado é a ausência de rótulos nas amostras de dados usadas para treinamento. Sendo assim, o modelo não tem orientação explícita sobre as classes ou categorias dos dados.
- Agrupamento (*Clustering*): Os algoritmos de agrupamento tentam identificar grupos naturais ou *clusters* de pontos de dados com base em alguma medida de similaridade ou dissimilaridade entre eles. Entre os algoritmos de agrupamento mais utilizados estão o *K-Means,* o *Hierarchical*

Clustering e o DBSCAN (Density-Based Spatial Clustering of Applications with Noise). (Grira et al., 2004).

- Redução de Dimensionalidade: Outra aplicação importante é a redução de dimensionalidade, envolvendo a transformação dos dados em um espaço de menor dimensão e mantendo o máximo de informações possível. Isso é frequentemente realizado com técnicas como a Análise de Componentes Principais (PCA) e a Redução Linear de Dimensionalidade (LDA) (Anowar et al., 2021).
- Detecção de Anomalias (*Outlier Detection*): O aprendizado não supervisionado também é usado para detectar anomalias ou pontos de dados com desvios significativos do comportamento típico do conjunto de dados. Técnicas como o *Isolation Forest* e o *One-Class SVM* são usadas para essa finalidade. (Sadaf e Sultana, 2020).
- *Generative Models*: Alguns métodos de aprendizado não supervisionado envolvem a modelagem da distribuição dos dados de treinamento para gerar novas amostras semelhantes aos dados originais. Isso é chamado de modelagem generativa. Exemplos incluem Redes Generativas Adversariais (GANs) e Modelos Ocultos de Markov (HMMs) (Suk et al. 2016).
- Validação e Avaliação: A avaliação de modelos em aprendizado não supervisionado pode ser mais desafiadora quando comparado com o aprendizado supervisionado por não existir rótulos de classe para comparação direta. Métricas como índice de silhueta e índice Davies-Bouldin são usadas para avaliar a qualidade dos agrupamentos (Cintia Ganesha Putri et al., 2020). No entanto, a avaliação também pode envolver uma análise mais qualitativa.

- Hiper parâmetros: Em seu estudo sobre o ajuste de hiper parâmetros, Hutter et al. (2015) descreve como a escolha de algoritmos e hiper parâmetros apropriados como crucial no aprendizado não supervisionado. Este estudo cita alternativas para otimizar os parâmetros de forma automática.
- Visualização de Dados: Visualização é uma parte importante do aprendizado não supervisionado. Na visualização de dados é comum usar técnicas de redução de dimensionalidade como otimização, aprendizagem de distribuição, separação cega de sinais e análise fatorial (Usama et al., 2019).
- Escalabilidade: Usama et al. (2019) citado acima ainda comenta sobre a escalabilidade dos algoritmos, onde conforme o aumento da quantidade dos dados, a escalabilidade dos algoritmos de aprendizado não supervisionado pode se tornar um problema. Alguns algoritmos de aprendizagem não supervisionados podem tornar-se computacionalmente intensivos conforme o aumento do tamanho do conjunto de dados, tornando-os impraticáveis para aplicações em grande escala. Ao lidar com um extenso conjunto de dados, muitas vezes é necessário empregar técnicas como processamento paralelo, computação distribuída ou uso de hardware especializado para lidar com as demandas computacionais.
- Interpretação: A interpretação no aprendizado não supervisionado pode ser mais desafiadora em comparação com tarefas supervisionadas devido à ausência de dados rotulados para fornecer contexto (Usama et al., 2019).

Ainda sobre a interpretação do resultado, alguns pontos devem ser considerados:

- Falta de Verdade Absoluta: No aprendizado supervisionado são estabelecidos rótulos de “verdade absoluta”, tornando mais fácil avaliar e interpretar o desempenho do modelo. No aprendizado não supervisionado, não há medida objetiva de correção ou precisão porque não existem rótulos predefinidos. Saligkaras e Papageorgiou (2023, August) discutem como a ausência de dados rotulados no aprendizado não supervisionado torna a validação e interpretação dos resultados um desafio.
- Subjetividade: A interpretação dos resultados no aprendizado não supervisionado frequentemente envolve um grau de subjetividade. Os clusters ou padrões identificados pelos algoritmos não supervisionados nem sempre têm significados claros no mundo real, e sua interpretação pode depender da perspectiva do analista (Saligkaras e Papageorgiou, 2023, August).
- Análise Qualitativa: Frequentemente, os resultados do aprendizado não supervisionado requerem análise qualitativa. Chen e Cummings (2024, May) examinaram como as escolhas subjetivas feitas pelos pesquisadores na coleta, processamento e análise de dados podem levar a diferentes resultados. Eles enfatizam que as preferências dos modeladores podem impactar a repetibilidade e a robustez dos estudos de aprendizado de máquina.
- Conhecimento de Domínio: O conhecimento de domínio desempenha um papel fundamental na interpretação dos resultados do aprendizado não supervisionado. A adaptação de domínio não supervisionada visa transferir conhecimento de um domínio de origem rotulado para um domínio de destino não rotulado. Recentemente, o aprendizado adversário de domínio

tornou-se um método cada vez mais popular para abordar essa tarefa, conectando o domínio de origem e o domínio de destino por meio do aprendizado adversário de representações invariantes ao domínio (Du et al., 2021).

- Visualização: As visualizações podem ser essenciais para entender e interpretar os resultados do aprendizado não supervisionado. Técnicas como t-SNE (*t-distributed Stochastic Neighbor Embedding*) e UMAP (*Uniform Manifold Approximation and Projection*) são amplamente utilizadas para reduzir a dimensionalidade e visualizar dados complexos em 2D ou 3D. Essas técnicas ajudam a identificar padrões, agrupamentos e anomalias em grandes conjuntos de dados, facilitando a análise exploratória e a validação de modelos (McInnes, Healy, e Melville, 2018).
- Processo Iterativo: A interpretação dos resultados no aprendizado não supervisionado envolve um processo iterativo de exploração e refinamento. Os analistas podem precisar experimentar diferentes configurações de parâmetros, técnicas de pré-processamento ou algoritmos específicos para obter resultados significativos. Métodos comuns incluem algoritmos de clusterização como *k-means*, onde os centroides são ajustados iterativamente, e técnicas de redução de dimensionalidade como t-SNE e UMAP (McInnes, Healy, e Melville, 2018).
- Validação Externa: Em alguns casos, medidas de validação externa ou métricas específicas do domínio podem ser usadas para avaliar a qualidade e a relevância dos resultados do aprendizado não supervisionado. Entretanto, essas medidas geralmente são características do problema estudado e podem não estar sempre disponíveis. A validação externa no aprendizado de

máquina não supervisionado é um processo crítico para avaliar a qualidade e a generalização dos modelos (Chen e Cummings, 2024, May)

2.1.1.3 Aprendizado de Máquina por Reforço

Em 1996, Kaelbling et al. iniciava sua pesquisa sobre o aprendizado de máquina por reforço (*Reinforcement Learning*, RL) citando a importância de considerar conceitos oriundos da estatística, psicologia, neurociência e ciência da computação. O RL considera uma forma de programar agentes através de recompensas e punições, sem necessidade de especificar como a tarefa deve ser realizada. A aprendizagem por reforço é o problema enfrentado por um agente com a necessidade de aprender o comportamento por meio de interações de tentativa e erro em um ambiente dinâmico. Essa abordagem é inspirada na psicologia do condicionamento operante e é amplamente usada em aplicações onde um agente interage com um ambiente para atingir objetivos, como jogos, robótica, otimização de recursos, entre outros. Abaixo estão listados os elementos e conceitos do aprendizado de máquina por reforço citados por Kaelbling et al. (1996) e Arulkumaran et al. (2017):

- Agente: O agente é a entidade de tomada de decisão em treinamento. Pode ser um programa de computador, um robô físico ou qualquer outra entidade capaz de realizar ações em um ambiente (Sutton e Barto, 2018).
- Ambiente: O ambiente é o contexto no qual o agente opera e toma ações. Pode ser um jogo, um simulador físico, um ambiente virtual ou qualquer sistema no qual o agente pode interagir (Russell e Norvig, 2016).
- Estado (*State*): Em um determinado ponto no tempo, o ambiente e o agente estão em um determinado estado, representando as informações relevantes

do ambiente no momento. O estado pode ser observável ou não observável, dependendo do problema (Joshi et al., 2021).

- Ação (*Action*): O agente escolhe ações, tendo como referência o estado atual. As ações representam as decisões tomadas pelo agente e têm impacto no ambiente. O espaço de ação (*action space*) abrange todas as possíveis ações (Sutton e Barto, 2018).
- Política (*Policy*): A política é a estratégia usada pelo agente para mapear estados para ações. Pode ser determinística (uma ação específica é escolhida para cada estado) ou estocástica (uma distribuição de probabilidade sobre as ações é associada a cada estado) (Sutton e Barto, 2018).
- Recompensa (*Reward*): A recompensa é uma medida numérica recebida pelo agente do ambiente após realizar uma ação em um determinado estado. A recompensa é usada para avaliar o desempenho do agente e é o objetivo do aprendizado por reforço maximizar a recompensa acumulada ao longo do tempo (Sutton e Barto, 2018).
- Horizonte temporal e retorno (*Time Horizon and Return*): O aprendizado por reforço é uma abordagem orientada por tempo. O agente toma decisões em cada passo de tempo e acumula recompensas ao longo de um horizonte temporal. O retorno é a soma das recompensas acumuladas ao longo do tempo, e o objetivo do agente é maximizar o retorno (Sutton e Barto, 2018).
- Função de valor (*Value Function*): A função de valor é uma função para estimar o valor de um estado ou estado-ação, indicando o quanto é bom estar em um estado específico ou tomar uma determinada ação em um estado específico. Duas funções de valor importantes são a função de valor de estado (V) e a função de valor de ação (Q) (Banik et al., 2021).

- Processo de decisão de Markov (*Markov Decision Process,* MDP): Este processo de decisão foi formulado por Puterman (2014). Ele descreve a formalização de um problema de decisão sequencial no qual um agente toma ações em um ambiente para maximizar uma recompensa acumulada ao longo do tempo.
- *Exploration* x *Exploitation*: Um desafio fundamental no aprendizado por reforço é encontrar um equilíbrio entre explorar novas ações para aprender mais sobre o ambiente e explorar ações para maximizar a recompensa atual. Isso é conhecido como o dilema de *exploration* x *exploitation* (Puterman, 2014)
- Algoritmos de Aprendizado: Existem vários algoritmos de aprendizado por reforço, incluindo *Q-learning*, SARSA, métodos baseados em políticas, como *REINFORCE*, e algoritmos mais recentes baseados em redes neurais, como DDPG (*Deep Deterministic Policy Gradient*), A3C (*Asynchronous Advantage Actor-Critic*) e PPO (*Proximal Policy Optimization*) (Banik et al., 2021).
- Treinamento e Aprendizado: O agente aprende através de iterações repetidas com o ambiente. Durante o treinamento, ele ajusta sua política com base nas experiências passadas e nas recompensas recebidas, usando algoritmos de otimização (Sutton e Barto, 2018).

O aprendizado por reforço tem uma ampla gama de aplicações que demonstram sua versatilidade e potencial para resolver problemas complexos e dinâmicos em diversas áreas. No domínio dos jogos, por exemplo, *AlphaGo*, desenvolvido pela *DeepMind*, utiliza técnicas de aprendizado por reforço combinadas com redes neurais profundas para vencer campeões humanos no jogo de Go, evidenciando a capacidade

do RL para resolver problemas de decisão sequencial complexos (Silver et al., 2016). Em robótica, o aprendizado por reforço é amplamente utilizado para ensinar robôs a realizar tarefas complexas, como navegação autônoma, manipulação de objetos e interação com humanos, destacando estudos como o de Kober, Bagnell e Peters (2013), que fornecem uma visão abrangente sobre o uso do RL em robótica.

Na área financeira, o método é empregado para desenvolver estratégias de negociação automatizadas que se adaptam a mudanças de mercado e otimizam portfólios, conforme discutido por Nevmyvaka, Feng e Kearns (2006). No campo da saúde, RL é utilizado para desenvolver planos de tratamento personalizados baseados em dados históricos de saúde e respostas a tratamentos anteriores, exemplificado no estudo de Liu et al. (2017) sobre medicina personalizada. Em transporte e logística, RL é aplicado para otimizar o controle de semáforos e reduzir congestionamentos, como demonstrado por Van der Pol e Oliehoek (2016) em seu trabalho sobre controle de tráfego. Por fim, em marketing e publicidade, RL é usado em sistemas de recomendação para personalizar ofertas e anúncios para usuários com base em seu comportamento e preferências anteriores, conforme descrito por Zhao et al. (2019). Essas aplicações mostram como o aprendizado por reforço pode ser implementado para melhorar a eficiência e a eficácia em diversos setores.

2.1.2 Modelos e Algoritmos Principais

“O aprendizado de máquina (ML) é o estudo científico de algoritmos e modelos estatísticos que os sistemas de computador usam para executar uma tarefa específica sem serem explicitamente programados. Algoritmos de aprendizagem estão em muitas aplicações utilizadas diariamente, o mecanismo de busca do Google usado para pesquisar na internet é um exemplo. Um dos motivos pelos quais funciona tão bem é porque um algoritmo de aprendizagem aprendeu como classificar páginas da web.

Esses algoritmos são usados para diversos fins, como mineração de dados, processamento de imagens, análise preditiva etc. A principal vantagem de usar o aprendizado de máquina é que, uma vez que um algoritmo aprende o que fazer com os dados, ele pode fazer seu trabalho automaticamente.” (Mahesh, 2020).

Traçado um panorama geral das principais características associadas com os tipos de aprendizagem de máquina, passamos a descrever sinteticamente alguns dos modelos e algoritmos utilizados no aprendizado de máquina citados neste trabalho:

- Regressão Linear: Usado para modelar a relação entre uma variável de saída contínua e um ou mais recursos, assumindo uma relação linear.
- Regressão Logística: Utilizada para problemas de classificação binária, onde a saída é uma probabilidade com variação entre 0 e 1.
- Árvores de Decisão: Modelos de árvore usadas para dividir os dados em subconjuntos com base em características para tomar decisões.
- *Random Forests*: Conjunto de árvores de decisão criadas para definir um resultado combinado para melhorar a precisão e reduzir o *overfitting*.
- *Support Vector Machines* (SVM): Usadas para classificação e regressão, baseadas em encontrar o hiperplano com melhor capacidade de separar as classes.
- *Artificial Neural Network* (ANNs): Modelos inspirados no funcionamento do cérebro, compostos por camadas de neurônios interconectados.
- *Convolutional Neural Network* (CNNs): Especializadas no processamento de dados de grade, amplamente usadas em visão computacional.
- *Recurrent Neural Networks* (RNNs): Projetadas para lidar com dados sequenciais, como séries temporais ou texto.

- *K-Means:* Algoritmo de agrupamento classificador de dados em clusters com base na proximidade.
- Agrupamento Hierárquico: Algoritmos para criar árvores de clusters aninhados para representar a estrutura hierárquica dos dados.
- Análise de Componentes Principais (PCA): Técnica de redução de dimensionalidade para preservar a maior parte da variância dos dados.
- *Naïve Bayes*: Usado em tarefas de classificação com base no teorema de Bayes e na suposição de independência entre recursos.
- Algoritmos de Aprendizado por Reforço: Como o *Q-Learning* e o *Deep Q-Network* (DQN), usados para aprender ações sequenciais para maximizar recompensas em um ambiente.
- Algoritmos de Detecção de Anomalias: Como o *Isolation Forest* e o *One-Class SVM*, usados para identificar observações anômalas em um conjunto de dados.
- *Density-Based Clustering Algorithms*: Como o DBSCAN e o *Mean Shift,* usados para encontrar clusters com densidades variáveis.
- *Gradient Boosting Machines*: Algoritmos como *Gradient Boosting, XGBoost e LightGBM,* melhoram iterativamente os modelos fracos para criar um modelo forte.
- *Natural Language Processing* (NLP): Técnicas e modelos específicos para processar texto, como o modelo de linguagem GPT (*Generative Pre-trained Transformer*).
- Aprendizado Semi-Supervisionado: Abordagens com capacidade de combinar dados rotulados e não rotulados para melhorar o desempenho do modelo.

- *Deep Learning*: Subconjunto do aprendizado de máquina concentrado em redes neurais profundas e modelos complexos.

2.1.3 Aplicações Práticas do *Machine Learning*

O aprendizado de máquina desempenha um papel fundamental em uma ampla variedade de setores e campos, trazendo soluções inovadoras e aprimorando a eficiência em diversas áreas da sociedade. Suas aplicações práticas são vastas e impactam diretamente na vida cotidiana. No Quadro 1 estão algumas aplicações do aprendizado de máquina.

Método	Aplicações
Classificação de Texto e Processamento de Linguagem Natural (NLP)	Análise de sentimentos em mídias sociais.
	Categorização de e-mails como spam ou não spam.
	Tradução automática de idiomas.
	Resumo automático de texto.
	Chatbots e assistentes virtuais.
Visão Computacional	Reconhecimento de objetos em imagens e vídeos.
	Detecção de rostos e reconhecimento facial.
	Identificação de placas de veículos.
	Diagnóstico médico por imagens.
	Controle de qualidade em linhas de produção.
Recomendações e Personalização	Recomendação de produtos em e-commerce.
	Sugestões de filmes e músicas em plataformas de streaming.
	Personalização de conteúdo em redes sociais e notícias.
Processamento de Áudio	Reconhecimento de fala e comandos de voz.
	Processamento de áudio para cancelamento de ruído.
	Transcrição automática de áudio.
Previsão e Análise de Dados	Previsão de demanda em cadeias de suprimentos.
	Previsão de preços de ações e mercado financeiro.
	Análise de crédito e detecção de fraudes financeiras.
	Previsão de falhas em equipamentos industriais (Manutenção Preditiva).
Saúde e Medicina	Diagnóstico médico auxiliado por máquina.
	Descoberta de medicamentos e triagem de compostos.
	Monitoramento de pacientes e previsão de riscos de saúde.
Robótica e Automação	Robôs autônomos para navegação e tarefas complexas.
	Controle de drones e veículos autônomos.
Marketing e Publicidade	Segmentação de público-alvo.
	Otimização de campanhas publicitárias.
	Análise de tendências e comportamento do consumidor.
Educação	Personalização de planos de ensino.
	Avaliação automatizada e feedback.
	Detecção de plágio.
Segurança	Detecção de intrusões em redes de computadores.
	Monitoramento de câmeras de segurança.
	Análise de comportamento para prevenção de ameaças.
Meio Ambiente	Previsão de condições climáticas e desastres naturais.
	Monitoramento da qualidade do ar e da água.
	Conservação da vida selvagem.
Setor Automobilístico	Veículos autônomos e assistência ao motorista.
	Otimização de rotas de entrega.
Recursos Humanos e RH	Recrutamento e seleção de candidatos.
	Avaliação de desempenho e engajamento de funcionários.
Esportes	Análise de desempenho de atletas.
	Previsão de resultados de competições.
Agricultura	Monitoramento de safras e otimização de cultivos.
	Detecção de pragas e doenças em plantações.

Quadro 1: Aplicações do aprendizado de máquina. Criado pelo autor.

2.1.4 Desafios e Considerações Éticas

O uso do aprendizado de máquina e da inteligência artificial traz uma série de desafios e considerações éticas necessárias de serem abordados para garantir um desenvolvimento responsável (Bilstrup et al., 2020). Abaixo, destacam-se alguns dos principais desafios e considerações éticas do uso do aprendizado de máquina:

- Viés nos Dados: Piasecki et al. (2018) começa seu artigo citando os riscos do viés nos dados. Os modelos de aprendizagem aprendem a partir de dados históricos, e se esses dados forem enviesados em relação a gênero, raça, classe social ou outros fatores, os modelos podem perpetuar e até ampliar esses vieses. Isso pode resultar em decisões discriminatórias e injustas (Osoba e Welser, 2017).
- Privacidade dos Dados: Muitas vezes são realizados coletas e processamento de quantidades significativas de dados pessoais. A garantia da privacidade desses dados e o cumprimento de regulamentações de privacidade, como a Lei Geral de Proteção de Dados (LGPD), são preocupações importantes (Dourado e Aith, 2022).
- Interpretação dos Modelos: Muitos modelos de aprendizagem, especialmente os baseados em redes neurais profundas, são caixas-pretas, dificultando a compreensão de como eles tomam decisões. Isso é particularmente relevante em áreas como saúde e justiça, onde a capacidade de explicar é considerada importante e em alguns casos obrigatória (McGovern et al., 2019).
- Segurança Cibernética: Os modelos podem ser vulneráveis a ataques para manipular os dados de entrada para enganar o modelo. Garantir a segurança dos sistemas é essencial, especialmente em aplicações críticas (Xin, 2018).

- Responsabilidade Legal e Ética: Quando erros ou danos são causados por sistemas de ML, é difícil atribuir responsabilidade. Questões legais e éticas sobre quem é responsável por decisões tomadas por algoritmos estão em discussão há décadas em muitos países, não se tendo ainda um arcabouço legal em relação a esse aspecto (Matthias, 2004). É provável que a tomada de decisões no futuro envolva tanto técnicos quanto sistemas de IA de alguma forma, exigindo a gestão de discordâncias entre máquina e humano, e a delegação de responsabilidade por decisões e erros (Carter et al., 2020).
- Desemprego Tecnológico: A automação impulsionada pelo ML pode levar ao desemprego em certas indústrias. É importante considerar políticas de reciclagem de habilidades e o impacto econômico dessas mudanças (Nguyen e Vo, 2022).
- Justiça e Equidade: Garantir o acesso às tecnologias de ML de forma equitativa e fazer a distribuição de benefícios de forma justa é uma consideração ética também importante (Floridi e Cowls, 2022).
- Monitoramento e Controle: O monitoramento contínuo de sistemas de ML é essencial para garantir a operabilidade de maneira ética e de acordo com as regulamentações (Dourado e Aith, 2022).
- Transparência e Responsabilidade: As organizações desenvolvedoras e implementadoras de sistemas de ML devem ser transparentes em relação ao uso dessas tecnologias e assumir a responsabilidade por seu impacto na sociedade (Matthias, 2004).
- Ética na Coleta de Dados: A coleta de dados deve ser realizada de maneira ética, respeitando os direitos das pessoas e obtendo consentimento quando necessário (Osoba e Welser, 2017).

A quantidade de dados novos expostos diariamente na internet com ou sem autorização do proprietário é incalculável. A necessidade de haver leis rígidas para inibir os vazamentos e o uso indevido vem sendo feito pelos governos. Entretanto, a ineficiência da fiscalização acaba deixando os usuários completamente ignorantes sobre quais os seus dados estão expostos. Não existem regras quanto ao uso e criação dos modelos de *machine learning*. Isto inclui os modelos utilizados no setor financeiro e pelo marketing das empresas. A questão em não deixar explicita a origem dos dados utilizados pelas empresas aumenta a impunidade. Existe um campo a ser estudado sobre como criar formas para proteger os usuários e empresas do uso de suas informações de forma indevida.

2.2 Riscos Financeiros

O marketing da indústria financeira busca criar produtos e serviços atrativos ao cliente, seja por uma rentabilidade maior ou por formas de emprestar dinheiro. Em ambos os casos, o risco de não conseguir honrar com as rentabilidades ofertadas assim como ter algum tipo de inadimplência pode gerar prejuízo direto decorrente de operações realizadas com empresas clientes.

Sendo assim, a análise de riscos é um trabalho constante e obrigatório.

2.2.1 Prevenção do Risco Financeiro

Prevenir o risco de inadimplência é vital para qualquer negócio ou instituição financeira fornecedora de crédito. A inadimplência ocorre quando um mutuário não cumpre com suas obrigações de pagamento conforme acordado. Um dos primeiros estudos sobre o risco de inadimplência empresarial foi escrito por Altman em 1968. Em seu artigo, Altman inicia questionando a utilidade da análise de índices como técnica analítica na avaliação do desempenho das empresas, utilizando uma metodologia

estatística discriminante múltipla para prever a falência das empresas. Outros autores seguiram Altman e passaram a utilizar outros tipos de modelos de risco de inadimplência e falência. Kanitz (1976, 1978) estabeleceu um modelo de análise discriminante múltipla que calculava o risco de falência em três níveis: solvente, penumbra e insolvente. Elisabetsky (1976) também empregando um modelo de análise calculava o risco de falência em dois níveis: solvente e insolvente. Frydman et al. (1985) utilizaram um algoritmo de particionamento recursivo; Li (1999) utilizou um algoritmo de análise bayesiana e Shumway (2001) utilizou modelos de risco em comparação com modelos estáticos. Portanto, há cerca de 50 anos existem métodos para calcular o risco de inadimplência ou de falência de empresas.

Desde então, os estudos concentram esforços em definir novos modelos e processos para analisar a saúde financeira das empresas. Antes de conceder crédito a um cliente é importante realizar uma análise de crédito completa, incluindo a verificação de referências, a análise de relatórios de crédito e a avaliação da situação financeira do cliente. Entretanto, outros pontos devem ser observados para evitar informações assíncronas entre as partes, levando a um mau funcionamento do mercado. O ganhador do Prêmio Nobel de Economia, George Akerlof, apontou este problema em seu artigo "*The Market for Lemons*" de 2003, onde deixa explícita a necessidade de ter os termos do crédito, incluindo taxas de juros, datas de pagamento e penalidades por atraso, de forma clara e entendida por ambas as partes.

A comunicação com os clientes deve ser mantida regularmente, incluindo lembretes sobre os vencimentos e atualizações de sua conta. Este e os outros procedimentos apresentados a seguir, estão descritos em livros sobre a análise financeira e a gestão do risco de crédito de clientes (Caouette et al., 2011; Fridson e Alvarez, 2022). Esses autores descrevem a necessidade de monitorar regularmente as

contas a receber e estar atento a sinais de alerta, como pagamentos atrasados ou padrões de pagamento inconsistentes.

Outro ponto importante é a definição do limite de crédito com base na análise de crédito e no histórico do cliente. Isso pode prevenir situações de aumento excessivo de endividamento de um cliente. Assim, uma estratégia utilizada tem sido das instituições financeiras buscarem ter uma base maior de clientes com valores compromissados menores para amortecer o impacto se um ou mais clientes se tornarem inadimplentes. Para o caso de clientes inadimplentes, é importante estabelecer uma reserva para devedores duvidosos. Isso ajudará a proteger o fluxo de caixa e a saúde financeira de bancos e fintechs.

Além da proteção contra devedores duvidosos, existe uma necessidade de desenvolver políticas de cobrança firmes e justas. Isso pode incluir a terceirização da cobrança para agências especializadas depois de um certo período ou gerar incentivos para os inadimplentes quitarem seus débitos. Ainda nesta linha, as instituições financeiras ou empresas que operam no mercado B2B podem oferecer descontos ou outros incentivos para os clientes pagarem suas contas em dia ou antecipadamente. Em alguns casos, pode ser apropriado solicitar garantias ou fiadores para proteção contra o risco de inadimplência.

Complementando, as condições de mercado e os padrões de comportamento dos clientes podem mudar. Portanto, é essencial revisar e ajustar regularmente a estratégia de gestão de crédito, assim como gerar modelos com processamento recorrente, garantindo a atualização dos dados dos clientes.

2.2.2 Marketing nas Instituições Financeiras

As instituições financeiras e empresas que operam no mercado B2B têm uma particularidade interessante a ser destacada: a análise de risco de crédito está sob a

responsabilidade da área financeira, enquanto a estratégia de marketing e vendas sob a responsabilidade da área comercial. De modo geral, é comum a primeira ser mais "restritiva" (em que um dos indicadores de avaliação de desempenho é o montante de perdas), enquanto a segunda mais "liberal" (em que um dos indicadores de desempenho é o crescimento das vendas e de market share). O marketing nas instituições financeiras visa atrair e reter clientes, promover produtos e serviços e aumentar a rentabilidade; desempenha também um papel auxiliar importante na gestão e mitigação de riscos de crédito associados aos clientes. A integração bem-sucedida das estratégias de marketing com a gestão de riscos pode ajudar as instituições financeiras e empresas que atuam no mercado B2B a alcançarem seus objetivos de negócios, mantendo a solidez e a estabilidade financeira.

Ele desempenha um papel crucial na aquisição e retenção de clientes, bem como na expansão dos negócios dessas organizações. O risco dos clientes, especialmente em termos de crédito, inadimplência e falência, é uma preocupação primordial. As atividades de marketing e o gerenciamento do risco do cliente estão intimamente interligados de várias maneiras.

Especificamente, o marketing bancário segmenta clientes efetivos e potenciais de acordo com vários critérios, incluindo sua capacidade financeira e perfil de risco (Raiter, 2021). Ao direcionar produtos e serviços diferenciados para segmentos específicos, os bancos podem melhor gerenciar o risco associado a cada grupo.

Raiter (2021) também cita em seu estudo sobre a segmentação dos clientes, como esta classificação pode ajudar as instituições financeiras a desenvolverem e ofertarem produtos financeiros de acordo com o perfil de risco do cliente, como taxas de juros diferenciadas para empréstimos com base na solvabilidade do cliente.

Em tempos de crise, o marketing das instituições financeiras pode intensificar a promoção de produtos de baixo risco, como contas de poupança ou investimentos

conservadores. Esta atitude é descrita por Hellmann et al. (2000), em sua visão de estabilidade de concentração, onde a falta de concorrência no inicio do século atual permitia aos bancos os maiores lucros nestas situações. Nestes casos, a concentração acabava gerando taxas de juros mais elevadas e assim um aumento da inadimplência (Boyd e De Nicolo, 2005).

De acordo com o Banco Central do Brasil, as instituições financeiras fornecem R$ 2,2 trilhões anualmente em crédito para a empresas, onde 50% desse montante é destinado para as pequenas e médias empresas, destacando-se que estas representam mais de 90% das empresas brasileiras. Ainda conforme o Banco Central do Brasil, quando analisadas as taxas de juro aplicadas, a distorção é ainda pior, com as pequenas e médias empresas operando com taxas 4,5 vezes mais altas em comparação com as grandes empresas (https://www.bcb.gov.br/estatisticas/estatisticasmonetariascredito).

As instituições financeiras modernas utilizam análise de dados extensivamente para combinar estratégias de marketing com avaliação de risco. Isso permite a elas a identificação das tendências, prever comportamentos de clientes e ajustar suas estratégias de marketing quanto à oferta de produtos.

O marketing de bancos e das fintechs – novas entrantes na indústria financeira - tem realizado campanhas regulares buscando educar os clientes sobre gestão financeira e responsabilidades de crédito, ajudando com isso a reduzir o risco de inadimplência, fato observado no estudo de Kaiser et al. (2022). Estas atitudes buscam uma boa gestão de relacionamento com o cliente onde os softwares de CRM (*Customer Relationship Management*) e algoritmos de inteligência artificial são utilizados para ajudar a identificar sinais precoces de possíveis problemas de crédito, permitindo uma intervenção proativa para minimizar riscos. Porém, é a interação contínua com os clientes, muitas vezes facilitada por estratégias de marketing, quem fornece o *feedback* valioso para as instituições financeiras. Elas podem usar esse *feedback* para ajustar seus

produtos, serviços e estratégias de gestão de risco. Campanhas de marketing transparentes e informativas sobre os produtos e serviços financeiros podem reduzir mal-entendidos e expectativas irrealistas, minimizando riscos não previstos.

2.2.3 Índices Financeiros

Em 1980, Whittington apontava como a discussão sobre os índices financeiros concentram-se normalmente na definição detalhada deles e na relevância de definições alternativas para diferentes utilizações, em vez de focar nas razões para utilizar os índices em preferência a outros métodos estatísticos. Os índices financeiros comumente são utilizados para avaliar a saúde financeira de uma empresa. Eles são extraídos das demonstrações financeiras, principalmente do balanço patrimonial e da demonstração do resultado. Mesmo os índices sendo calculados com dados recentes, eles ainda revelam o já acontecido e pouco informam sobre o futuro (Haji, 2015). Porém, o agrupamento das informações pode dizer muito sobre a atual situação da empresa, e este estudo traz uma nova forma de analisar estes índices.

Malíková e Brabec (2012) comentam sobre os índices financeiros serem muito populares, nomeadamente pela simplicidade do seu cálculo e pela fácil disponibilidade da obtenção dos dados para o seu cálculo. Os índices financeiros fornecem uma maneira de analisar a posição financeira e o desempenho da empresa escolhida e permitem comparar a saúde financeira e o seu desempenho em séries temporais, bem como com outras empresas, particularmente do mesmo setor de atividade econômica e porte. Os índices financeiros podem ser categorizados em cinco grupos, liquidez, endividamento, rentabilidade, atividade e estrutura de capital (Neto, 2003). Ferreira (2009), disserta sobre os índices financeiros. Abaixo são listados os índices mais utilizados para as análises das empresas utilizando o balanço contábil como base.

Os índices de liquidez mostram a capacidade de uma empresa pagar suas dívidas no curto prazo. Eles são essenciais para avaliar a saúde financeira de uma organização. Entre os índices mais comuns, estão: liquidez corrente (LC), liquidez seca (LS), liquidez imediata (LI) e liquidez geral (LG).

A liquidez corrente (LC) mostra a capacidade da empresa de pagar suas obrigações de curto prazo com seus ativos circulantes. Uma LC maior de 1 indica uma maior quantidade de ativos em comparação com os passivos circulantes.

$$LC = \text{Ativo Circulante} / \text{Passivo Circulante}$$

A liquidez seca (LS) é semelhante à liquidez corrente, mas exclui os estoques do ativo circulante. Isso fornece uma medida mais conservadora da capacidade de pagamento de curto prazo.

$$LS = (\text{Ativo Circulante} - \text{Estoques}) / \text{Passivo Circulante}$$

A liquidez imediata (LI) foca apenas nas disponibilidades (como caixa e equivalentes de caixa) para pagar os passivos de curto prazo.

$$LI = \text{Disponibilidades} / \text{Passivo Circulante}$$

A liquidez geral (LG) considera tanto os ativos e passivos circulantes quanto os de longo prazo para avaliar a capacidade de pagamento da empresa.

$$LG = \frac{(\text{Ativo Circulante} + \text{Realizável a Longo Prazo})}{(\text{Passivo Circulante} + \text{Exigível a Longo Prazo})}$$

Ao avaliar esses índices, é importante analisar se os valores estão acima de 1, pois geralmente indicam uma melhor saúde financeira. Comparar os índices de liquidez de uma empresa com os de outras empresas do mesmo setor pode fornecer percepções adicionais. Acompanhar a evolução da liquidez ao longo do tempo pode ajudar a identificar tendências e mudanças na saúde financeira da empresa.

Os índices de endividamento avaliam a estrutura de capital de uma empresa, ou seja, a relação entre o capital próprio e o capital de terceiros usado no financiamento de suas atividades. Estes índices ajudam a entender o grau de alavancagem financeira de uma empresa e os riscos associados ao seu nível de dívida. Os mais utilizados são os de endividamento geral (EG), comprometimento do patrimônio líquido (CP), imobilização do patrimônio líquido (IP) e cobertura de juros (ICJ).

O índice de endividamento geral (EG) indica a proporção do ativo total financiada por terceiros. Quanto maior o índice, maior o grau de endividamento da empresa.

EG = (Passivo Total / Ativo Total) x 100

O índice de comprometimento do patrimônio líquido (CP) mostra a relação entre as dívidas totais e o patrimônio líquido da empresa. Um índice acima de 100% indica uma maior quantidade de dívidas em comparação com os recursos próprios.

CP = (Passivo Total / Patrimônio Líquido) x 100

O índice de imobilização do Patrimônio líquido (IP) indica qual porcentagem do patrimônio líquido está investida em ativos permanentes (ativos imobilizados). Quanto maior o índice, maior a imobilização do capital próprio.

IP = (Ativo Imobilizado / Patrimônio Líquido) / 100

O índice de cobertura de juros (ICJ) mostra a capacidade da empresa em gerar recursos operacionais suficientes para cobrir suas despesas com juros. Um ICJ abaixo de 1 sugere a possibilidade a empresa ter dificuldade em pagar os juros de sua dívida.

ICJ = EBIT / Despesas com Juros

Ao avaliar os índices de endividamento é necessário considerar os padrões do setor. Observar a tendência ao longo do tempo, aumentos significativos no endividamento podem indicar riscos crescentes. Analisar em conjunto com outros indicadores, como os índices de liquidez, rentabilidade e atividade, para obter uma visão holística da saúde financeira da empresa. Ainda sobre a dívida, deve ser considerado o perfil da dívida: prazos, taxas de juros e moedas nas quais as dívidas foram contraídas também são relevantes na análise, lembrando que a interpretação dos índices pode variar de acordo com o contexto econômico, a natureza da empresa e o setor de atuação.

Os índices de rentabilidade são fundamentais para analisar a capacidade de uma empresa gerar lucro em relação a um determinado parâmetro, como vendas, ativos ou patrimônio líquido. Eles fornecem informações sobre a eficiência operacional e

financeira da empresa. Os principais índices de rentabilidade são a margem bruta, a margem operacional, a margem líquida, o retorno sobre o ativo (ROA), o retorno sobre o patrimônio líquido (ROE) e o retorno sobre o investimento (ROI).

A margem bruta mede a rentabilidade média do negócio, ou seja, qual a porcentagem de lucro a empresa ganha com asa venda.

Margem Bruta = (Lucro Bruto / Vendas Totais) x 100

A margem operacional mede a porcentagem de lucro obtido a partir da receita bruta, esse índice avalia a geração de lucro através de suas operações, a eficiência da gestão e manutenção do lucro. A margem operacional deduz todas as despesas operacionais, exceto juros e impostos.

Margem Operacional = (Lucro Operacional / Vendas Totais) * 100

A margem líquida é o indicador financeiro para verificar a porcentagem de lucro em relação às receitas. A margem líquida deduz todas as despesas, incluindo juros e impostos.

Margem Líquida = (Lucro Líquido / Vendas Totais) / 100

O retorno sobre o ativo (ROA) mede odesempenho da empresa em usar seus ativos para gerar lucro e é utilizado para medir a eficiência das ações ou para quem quer descobrir qual retorno podem ter ao investir em ações de uma companhia.

ROA = Lucro Líquido / Ativo Total Médio

O retorno sobre o patrimônio líquido (ROE) mede a capacidade da empresa em gerar retorno para os proprietários ou acionistas com base nos recursos próprios da empresa.

ROE = Lucro Líquido / Patrimônio Líquido Médio

O retorno sobre o investimento (ROI) mede o retorno financeiro gerado em relação ao investimento realizado, considerando tanto os recursos próprios quanto os de terceiros.

ROI = Lucro Operacional / Investimento Total

Os índices de atividade, também conhecidos como índices de eficiência ou índices operacionais, medem odesempenho de uma empresa utilizar seus ativos no processo operacional. Esses índices são úteis para avaliar a capacidade da empresa de gerenciar seus ativos de forma eficiente. Os mais utilizados são o giro de estoques, o prazo médio de estocagem, o giro do contas a receber, o prazo médio de recebimento, o giro do contas a pagar, o prazo médio de pagamento e o giro do ativo total. Matias (2007) afirma como o objetivo da gestão do capital de giro é o mesmo objetivo da empresa: gerar valor econômico para os *stakeholders*.

O giro de estoque mede a quantidade de vezes que o estoque foi renovado ou vendido durante um período. Um número mais alto geralmente indica boa gestão do estoque, enquanto um número baixo pode sugerir estoques excessivos ou vendas lentas.

Giro de Estoque = Custo das Mercadorias Vendidas / Estoque Médio

O prazo médio de estocagem indica, em média, quantos dias a empresa leva para vender seu estoque.

Prazo Médio de Estocagem = 365 / Giro de Estoque

O giro do contas a receber mostra quantas vezes o “contas a receber” foi recebido durante o período.

Giro do Contas a Receber = Vendas a Prazo / Contas a Receber Médio

O prazo médio de recebimento indica quantos dias, em média, a empresa leva para receber.

Prazo Médio de Recebimento = 365 / Giro do Contas a Receber

O giro do contas a pagar mostra quantas vezes a empresa pagou suas “contas a pagar” durante um período.

Giro do Contas a Pagar = Compras a Crédito / Contas a Pagar Médio

O prazo médio de pagamento representa a média do número de dias que a empresa leva para pagar seus fornecedores.

Prazo Médio de Pagamento = 365 / Giro de Contas a Pagar

O giro do ativo total mede a eficiência da empresa utilizar todos os seus ativos para gerar vendas. Um número mais alto indica uma utilização mais eficaz dos ativos.

Giro do Ativo Total = Vendas / Ativo Total Médio

Os índices de estrutura de capital, também conhecidos como índices de endividamento ou alavancagem, medem quanto uma empresa financia suas operações por meio de dívida em relação ao capital próprio. Esses índices são fundamentais para avaliar a solidez financeira e a capacidade de uma empresa cumprir suas obrigações decorrentes de dívidas contraídas. Os principais índices são os de endividamento, de imobilização do patrimônio líquido, de cobertura de juros e de alavancagem financeira.

O índice de endividamento indica a proporção do total de ativos financiados por credores. Um índice mais alto pode indicar maior risco para os credores e investidores.

Índice de Endividamento = Dívida Total / Ativo Total

O índice de imobilização do patrimônio líquido mede a proporção dos ativos permanentes financiados pelos proprietários da empresa.

Índice de Imobilização do PL = Ativo Permanente / Patrimônio Líquido

O índice de cobertura de juros avalia a capacidade da empresa de cumprir suas obrigações de juros com os lucros gerados. Um índice mais alto indica um bom posicionamento da empresa para cumprir suas obrigações de juros.

$$\text{Índice de Cobertura de Juros} = \frac{\text{Lucro Antes dos Juros e Impostos (LAJIR)}}{\text{Despesas de Juros}}$$

O índice de alavancagem financeira mede o grau de utilização de dívida em relação ao capital próprio. Quanto maior o índice, maior é a alavancagem financeira e o risco financeiro.

Índice de Alavancagem = Capital de Terceiros / Patrimônio Líquido

Além dos índices descritos acima, este estudo leva em considerações os valores lançados diretamente nas principais contas analíticas do balanço. Ou seja, as contas contábeis que compões os índices financeiros foram utilizadas pelas suas análises horizontais e verticais, por exemplo o Patrimônio Líquido (utilizado no índice de alavancagem) também foi utilizado como fonte de treinamento do modelo.

2.3 Aprendizado de Máquina Aplicado em Falência e Inadimplência.

Como já salientado, um dos primeiros trabalhos a utilizar métodos estatísticos para prever a falência das empresas foi realizado por Altman (1968), o qual utilizou a análise discriminante como método principal. Frydman et al. (1985) aplicaram o algoritmo de particionamento recursivo para avaliar o caso da dificuldade financeira das empresas em comparação à análise discriminante utilizada em 1968 por Altman. O algoritmo utilizado é uma árvore de decisão com suas "folhas" analisando os índices financeiros escolhido pelos autores.

O modelo de risco utilizado por Shumway (2001) para prever a falência das empresas também utiliza o estudo de Altman como base para comparar os seus resultados, porém com uma acurácia abaixo de 70%.

Kassai e Onusic (2004) escolheram a análise envoltória de dados para fazer a predição da insolvência das empresas e os índices de endividamento geral, endividamento de longo prazo e a composição do endividamento como entradas do modelo. O modelo atingiu o índice de acerto de 90,0% entre as empresas insolventes e de 74,0% entre as empresas solventes, resultando em 76,6% de acertos de classificação no total. Um ponto de observação quanto ao estudo é o tamanho da base utilizada com apenas 60 empresas, sendo 10 insolventes e 50 solventes, aspecto este altamente limitador para generalização do resultado.

Em 2004, Carton criou um modelo preditivo utilizando análise de regressão nos índices financeiros previamente separados em seis categorias de medidas, rentabilidade, crescimento, eficiência, fluxo de caixa, sobrevivência e baseadas no mercado. Os agrupamentos dos índices foram validados por análises fatoriais e o modelo estrutural conseguiu uma acurácia pouco acima de 80%. Independente da precisão do modelo, a relevante contribuição do estudo é a utilização da separação dos índices financeiros em categorias e comprovação da eficiência desta separação pela análise fatorial.

Min et al. (2006) propuseram uma abordagem híbrida, integrando algoritmos genéticos e SVM para maximizar o desempenho preditivo do SVM. Foram utilizados dois aspectos, seleção de subconjuntos de recursos e otimização de parâmetros. A aplicação do modelo híbrido proposto melhorou a eficácia na previsão de falências, entretanto, a amostra de validação conseguiu uma acurácia de 80,3%.

A preocupação com o impacto das condições macroeconômicas para a modelagem do risco de crédito foi estudada por Carling et al. (2007). Ao utilizar o método de simulação de Monte-Carlo para tratar o tipo de valor em risco e o déficit esperado, os autores fazem previsões do risco de inadimplência condicionadas às características da empresa e do empréstimo, bem como com as condições macroeconômicas.

Outro estudo voltado à identificação de falência foi realizado por Hung e Chen (2009). Os autores compararam três métodos de aprendizado de máquina: árvore de decisão, rede neural de retro propagação (BNN) e a máquina de vetores de suporte (SVM). Foi utilizada uma base de dados com 30 índices financeiros de 120 empresas. O resultado apresentado mostrou uma acurácia máxima de 72,5% no método BNN.

Em um estudo comparativo entre os métodos de *machine learning*, Aktan (2011) testou a eficiência de 8 algoritmos de aprendizado de máquina como *Naive Bayes, Bayesian Network*, k-NN, ANN, SVM, C4.5, CHAID e CART em empresas

com dificuldades financeiras. Foram usados 53 índices financeiros e o estudo mostrou uma maior acuracidade na árvore de decisão CART (95,3%). Entretanto, os métodos *Naive Bayes* (80,9%) e *Bayesian Network* (84,6%) também tiveram uma performance melhor em comparação ao ANN (Rede Neural Artificial) (76,3%).

O estudo de Barboza, Kimura e Altman (2017) utilizou as categorias de Carton (2004) em modelos de aprendizado de máquina. Uma melhora estrutural foi a inclusão de novas categorias, como a margem operacional e a mudança no retorno sobre o patrimônio líquido, para melhorar a predição de falência. Na amostra utilizada, o método de aprendizado de máquina *random forest* liderou o estudo com uma acurácia de 87,06%. Porém, outros métodos, como *bagging* e *boosting,* tiverem também resultados melhores em comparação ao estudo de Carton (2004).

Em estudo para verificar a performance dos métodos de redes neurais, Addo et al. (2018) comparou sete modelos: rede elástica (regressão logística com regularização), *random forest*, *gradient boosting* e uma abordagem de rede neural com quatro complexidades diferentes. Inicialmente, foram utilizadas 181 variáveis em uma base de dados com mais de 100 mil empresas e em um segundo momento o trabalho compara os modelos com as 10 variáveis mais importantes definidas pelos próprios modelos. Como conclusão, nenhum dos quatro modelos de redes neurais proporcionou melhor desempenho em comparação com os modelos *random forest* e *gradient boosting*. A acurácias dos dois melhores modelos foi superior a 97% em todos os casos, chegando a passar de 99% em alguns deles; contudo, um aspecto fundamental a ser ressaltado é que não ficou claro no estudo a não incidência de *overfitting* nos modelos que leva a um aumento irreal no indicador de acurácia.

Hosaka (2019) utilizou redes neurais convolucionais (CNN) para prever falências usando imagens monocromáticas representando os índices financeiros. A razão apresentada cita como as redes neurais convolucionais são mais adequadas para

aplicação em imagens e menos adequadas para dados numéricos gerais, incluindo demonstrações financeiras. Foi utilizado um total de 7.520 imagens para as classes de empresas falidas e continuadas para treinar a rede neural convolucional baseada no GoogleNet. As previsões de falências por meio da rede treinada demonstraram ter um desempenho superior em comparação com métodos de árvores de decisão, análise discriminante linear, máquinas de vetores de suporte e perceptron multicamadas. Este trabalho utilizou somente os índices financeiros de um único balanço, sendo uma visão estática no tempo, não analisando a evolução da empresa em um intervalo maior.

O modelo supervisionado de aprendizado de máquina *extreme gradient boosting* foi utilizado por Bussmann et al. (2021) em comparação com a regressão logística para analisar a possibilidade de inadimplência de empresas de pequeno e médio porte. A amostra utilizada contém 15.000 empresas deste perfil com a identificação da inadimplência como uma variável booleana, além dos índices financeiros e contábeis tradicionais. Estas empresas estavam nesta base de dados por terem solicitado algum tipo de empréstimo. A amostra foi separada randomicamente em treinamento (80%) e teste (20%) e o método *extreme gradient boosting* chegou à acurácia de 93%. Esta alta acurácia pode ter acontecido pelo fato de todas as empresas da base serem devedoras de alguma forma, sendo um modelo bastante limitado pelo fato de não ter sido testado em empresas não solicitantes de empréstimos.

As limitações observadas nos estudos realizados para prever a inadimplência e ou a falência de empresas vão desde a utilização de amostras pequenas, de considerar apenas empresas falidas ou insolventes, de empresas de um único país, de possível *overfitting* do modelo, de não testar os *assumptions* requeridos pelo uso de métodos estatísticos como o de regressão ou discriminante, dentre outros aspectos. Ainda sobre as amostras, na expressiva maioria dos casos são de empresas abertas e os dados

contábeis utilizados são anuais (Balanços Contábeis), o que também caracteriza uma limitação metodológica de validade e confiabilidade.

2.4 *Universal Sentence Encoder* (Codificador de Frase Universal).

Cer et al. (2018) criaram um modelo chamado de *Universal Sentence Encoder*, utilizado para codificação de sentenças em vetores de incorporação visando a transferência de aprendizagem para outras tarefas ou situações.

O *Universal Sentence Encoder* foi projetado para codificar frases em vetores de alta dimensão. Isto facilita a realização de diversas tarefas de processamento de linguagem, como similaridade semântica, classificação de texto e agrupamento.

O modelo se destaca por sua habilidade em capturar significados semânticos complexos das frases. Ele é treinado em uma variedade de fontes de dados e utiliza aprendizado profundo (*Deep Learning*) para gerar representações vetoriais eficazes para muitas tarefas de NLP. Os dados de treinamento não supervisionados para os modelos de codificação de frases foram extraídos de uma variedade de fontes da web, como Wikipedia, sites de notícias, páginas de perguntas e respostas e fóruns de discussão. Os autores aumentaram a aprendizagem não supervisionada com treinamento em dados supervisionados do corpus *Stanford Natural Language Inference* (Bowman et al., 2015). O corpus *Stanford Natural Language Inference* é uma coleção disponível gratuitamente de pares de frases rotuladas com 570 mil pares escrita por humanos e permitiu uma maior captura de um amplo espectro de nuances linguísticas e contextos.

Foram utilizadas duas arquiteturas no desenvolvimento das versões, uma baseada em *Transformer* e a outra em *Deep Averaging Network* (DAN).

A arquitetura *Transformer* é um tipo de modelo de processamento de linguagem natural introduzido no artigo "*Attention Is All You Need*" de Vaswani et al. (2017), e

difere de arquiteturas anteriores por depender exclusivamente de mecanismos de atenção, dispensando a necessidade de redes recorrentes. Entre as principais características estão a atenção autorregressiva onde o modelo pesa a importância de uma entrada no processamento de cada palavra na sequência, sendo fundamental para entender o contexto e as relações dentro do texto; ser composto de codificadores e decodificadores empilhados, onde cada codificador consiste em camadas de atenção seguidas de redes neurais *feed-forward* e cada decodificador também inclui uma camada de atenção adicional focado nas saídas do codificador; codificações posicionais adicionadas às incorporações de entrada, permitindo o modelo saber a posição das palavras na sequência.

A arquitetura *Deep Averaging Network* (DAN) é um tipo de rede neural para o processamento de linguagem natural caracterizada por sua simplicidade e eficiência. As palavras de uma frase são inicialmente convertidas em vetores através de incorporações de palavras pré-treinadas. Em seguida, o modelo calcula a média desses vetores de palavra para criar uma representação vetorial única da frase ou documento. O vetor médio é então passado através de uma ou mais camadas *feed-forward* de uma rede neural profunda. Entre essas camadas, são aplicadas funções de ativação não linear e técnicas de regularização, como *dropout*, para melhorar a capacidade do modelo de generalizar a partir dos dados de treinamento. A última camada da rede produz a saída, podendo ser configurada para diferentes tarefas, como classificação ou regressão.

Como comentado acima, as arquiteturas usam vetores de palavras pré-treinados, onde cada palavra é mapeada para um vetor de alta dimensão com a capacidade de capturar seu significado semântico. Estes vetores são obtidos por meio de métodos como o *Word2Vec* por Mikolov et al. (2013) ou *GloVe* por Pennington et al. (2014, October) e são treinados em um vasto corpus de texto.

Word2Vec é uma técnica popular de processamento de linguagem natural (NLP) usada para aprender representações vetoriais de palavras, chamadas de *word embeddings* (vetores ou incorporações de palavras). Esses vetores procuram capturar o significado semântico, as relações sintáticas e o contexto das palavras dentro de enormes conjuntos de dados de texto. As arquiteturas utilizadas foram *Skip-Gram* e *Continuous Bag of Words* (CBOW).

Skip-Gram: Prediz o contexto (palavras ao redor) dada uma palavra. É útil quando a quantidade de dados de treinamento é pequena e as representações de palavras raras são necessárias.

CBOW: Prediz uma palavra com base em seu contexto. É mais rápido e tem melhor desempenho com palavras frequentes.

GloVe (*Global Vectors for Word Representation*) é uma técnica de modelagem de espaço vetorial para representação de palavras em processamento de linguagem natural (NLP). Esta técnica usa a abordagem baseada na ideia de as relações entre as palavras poderem ser compreendidas pela agregação global das estatísticas co-ocorrestes de um corpus de texto.

A função de custo do *GloVe* é projetada para forçar o produto escalar $w_i^T w_j$ a ser igual ao logaritmo da probabilidade de co-ocorrência das palavras *i* e *j*, proporcional ao logaritmo da contagem de co-ocorrência X_{ij}. A função de custo é dada pela seguinte equação:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T w_j + b_i + b_j - \log X_{ij})^2$$

Onde, w_i e w_j são os vetores de palavras para as palavras *i* e *j*, b_i e b_j são termos de viés para cada palavra, *V* é o tamanho do vocabulário, e $f(X_{ij})$ é uma função de ponderação para ajudar a lidar com palavras raras.

2.5 Visualização de Dados

McCormick (1988), define a visualização de informação como um método de computação para transformar o simbólico em geométrico, permitindo observar as simulações e cálculos e oferecendo um método para “observar” o invisível. Esta transformação enriquece o processo de descoberta científica e promove insights profundos e inesperados.

A visualização é o processo de transformar dados, informações e conhecimento em apresentações gráficas para apoiar tarefas como análise de dados, exploração de informações, explicação de informações, previsão de tendências, detecção de padrões, descoberta de ritmo e assim por diante (Zhang, 2007).

Zhang (2007) ainda reforça que, sem o auxílio da visualização, há menos percepção ou compreensão dos dados, informações ou conhecimento pelas pessoas por diversos motivos. Estas razões podem incluir as limitações da visão humana ou a invisibilidade e abstração dos dados, informações e conhecimento.

A conversão dos dados brutos em um formato interpretável e exibível para os usuários poderem interpretar passa por algoritmos e transformações dos dados; entretanto, esta conversão não precisa ser feita obrigatoriamente para um humano interpretar.

A seção 2.1.3 mostra como existem modelos de inteligência artificial capazes de buscar padrões em imagens, desde a identificação de um animal, até a identificação de doenças como o câncer de pele. Os benefícios da aplicação da visualização de informações podem variar desde o uso da capacidade perceptiva humana até a redução da carga de trabalho cognitivo e o aumento da eficácia de novas recuperações de informações (Zhang, 2007).

A visualização aplica a utilização das imagens a problemas práticos de análise de dados, sendo assim, pode-se esperar o desenvolvimento de uma ciência aplicada à

visualização de dados. Conforme o crescimento da importância da visualização, crescem também os benefícios de uma abordagem científica. Novos sistemas de símbolos e imagens estão sendo desenvolvidos para atender às necessidades de uma sociedade com uma quantidade de dados com crescimento exponencial. Estes novos sistemas e imagens desenvolvidos associados aos novos algoritmos podem coexistir por muito tempo (Ware, 2019).

Assim como o olho humano e seus neurônios são capazes de perceber as pequenas variações de cores, as informações traduzidas em imagens podem transmitir a mesma sensibilidade para as redes neurais artificiais. A luminescência e o brilho podem transmitir a intensidade e importância de uma informação, assim como as posições podem definir conteúdos. Uma imagem criada a partir de dados pode não ser compreendida por humanos. Uma estrutura criada a partir de variáveis pode identificar desde variáveis nominais e ordinais, até as intervaladas e proporcionais. As variáveis ganham formas e intensidades diferentes, podendo ter nuances e padrões imperceptíveis ao olho humano. O *QR Code* (*Quick Response Code*) pode ser citado como um destes casos, os textos ou informações são convertidos em uma imagem 2D (bidimensional), sendo incapaz de ser lida sem um algoritmo para convertê-la de volta.

Figura 1: Exemplo de *QR Code*. Criado pelo autor.

A rede neural voltada ao tratamento de imagens analisa os padrões, decompondo a imagem aplicando filtros projetados para detectar tipos específicos de características, como bordas, ângulos ou texturas. Por exemplo, um filtro pode ser configurado para reagir fortemente a áreas com variações bruscas de intensidade de luz, indicando a presença de uma borda. Estes padrões são classificados por proximidade, similaridade, continuidade, simetria e contornos, entre outros aspectos. Isto gera um mapa de características, onde outras funções introduzem a ponderação e a não-linearidade, ajudando a rede a aprender padrões complexos.

Estas camadas criadas pelos algoritmos são estruturadas de forma a gerar respostas diretas sobre o conteúdo e significado das imagens de acordo com um pré-treinamento ou agrupamento previamente definido.

Apresentado o arcabouço teórico que fundamenta a presente tese, apresenta-se, a seguir, a descrição dos procedimentos metodológicos utilizados para a criação do método de classificação de risco e predição da falência de empresas de pequeno e médio porte.

3. ASPECTOS METODOLÓGICOS

O método proposto nesta tese leva em consideração a utilização dos dados contábeis extraídos diretamente da contabilidade, fornecidos pelos contadores das empresas amostradas. A base atual tem mais de 105 mil empresas com mais de 400 milhões de lançamentos contábeis mensais. O modelo preditivo proposto está fundamentado na utilização de métodos de aprendizado de máquina por imagens para analisar a solvabilidade ou não de empresas de pequeno e médio porte.

3.1. Tipologia da Pesquisa

Primeiro segue a descrição da base dados utilizada na pesquisa. Este estudo utilizou os *backups* de 580 empresa de contabilidades com o conteúdo dos dados contábeis completos de pouco mais de 105 mil empresas de pequeno e médio porte, abrangendo diversos setores e regiões da economia brasileira. Ao todo, tem-se uma base expressiva de dados abrangendo mais de 20 milhões de registros de planos de contas contábeis e mais de 400 milhões de lançamentos contábeis dos últimos 10 anos.

Os *backups* lidos são de 15 sistemas contábeis diferentes, com formas distintas de estruturação dos dados. A contabilidade utiliza padrões de nomenclatura de suas contas contábeis, porém nem sempre os contadores utilizam as mesmas numerações e os mesmos tipos de estrutura para indicar os lançamentos. É comum um plano de contas ter, por exemplo, uma conta Banco em um código e em outro plano de contas a mesma conta Banco estar identificada por um outro código ou estrutura.

Este problema de falta de um padrão único gera inúmeros desafios a serem transpostos para analisar os lançamentos contábeis, simplesmente pelo fato de não ser possível comparar diretamente os dados de uma empresa com os de outra.

Desta forma, a utilização de dados consolidados sempre ajudou os pesquisadores na realização de comparação criteriosa de dados entres as empresas. Os dados dos Balanços e Balancetes são informações consolidadas e com uma estrutura mais homogênea. Entretanto, se perde muito dos dados originais. Os lançamentos contábeis falam muito sobre a situação da empresa, enriquecendo em muito a análise comparativa.

Sendo assim, a primeira parte deste trabalho consiste em normalizar os dados dos planos de contas contábeis de todas as empresas, atualizando todos os lançamentos contábeis destas empresas para um plano de contas **único**. Neste estudo foram consolidadas as contas contábeis sintéticas em 208 contas contábeis normalizadas.

Com os lançamentos contábeis das empresas atualizados em um único plano de contas, foi gerado um balancete para cada mês com dados de cada empresa em forma de um vetor com todos os dados contábeis da empresa.

Como último passo antes de submeter os dados aos métodos de *machine learning,* foram gerados os índices financeiros e suas memórias de cálculo em um único registro para cada mês de cada empresa. Também foram criados os índices de variação dos dados de um mês para o outro.

Para evitar erros em analisar dados absolutos, todos os dados foram reduzidos em um percentual proporcional ao tamanho da empresa, facilitando assim a comparação direta das empresas de diferentes portes.

3.2. Procedimento de Normalização

Como apresentado, o método inicia com os procedimentos de normalização dos Planos de Contas e estruturação destas contas nos lançamentos contábeis. Este passo é necessário para poder homogeneizar os dados e assim poder compará-los.

3.2.1. Plano de Contas

Gans et al. (2007) descrevem o plano de contas como a base para os sistemas de informação contábil nas organizações e indústrias. Planos de contas são listas de títulos de contas e estruturas de codificação numérica correspondentes usadas para registrar dados financeiros, como receitas e despesas, bem como para descrever ativos e passivos. Usando um plano de contas, os dados financeiros podem ser classificados e agregados uniformemente em classificações operacionais, como linhas de produtos, centros de custo, funções operacionais ou outras categorias para necessidades específicas da organização e do setor das empresas. Essas contas servem de núcleo para a compilação de demonstrações financeiras padronizadas e relatórios gerenciais benéficos para os *stakeholders* internos e externos. O plano de contas é um componente fundamental dos sistemas de informação contábil das indústrias públicas e privadas.

Existem padrões de planos de contas pré-definidos, como o definido pelo Sistema Público de Escrituração Digital (SPED), uma solução fornecida pela Receita Federal do Brasil para oficializar os arquivos digitais das escriturações fiscal e contábil dos sistemas empresariais dentro de um formato específico e padronizado. Este plano de contas padrão foi criado para normalizar as contas contábeis para a fiscalização pelos órgãos públicos; entretanto, as empresas não são obrigadas a utilizá-lo, sendo somente feita a conversão de seu plano de contas para o SPED no momento do envio do documento à Receita Federal.

Outro modelo padronizado conhecido internacionalmente é o *International Financial Reporting Standards* (IFRS), este criado pelo *International Accounting Standards Board* (IASB), com o intuito de normalizar os lançamentos contábeis de empresas do setor financeiro. Entretanto, a sua utilização não é tão simples (Trimble, 2017). Geralmente as diferenças dos procedimentos locais para os internacionais geram variações do IFRS, evitando sua total implantação. No Brasil, o Comitê de

Pronunciamentos Contábeis (CPC) é responsável por emitir normas contábeis que buscam alinhar-se ao IFRS, adaptando as diretrizes internacionais às especificidades locais.

Portanto, pode-se observar na prática uma vasta quantidade de planos de contas parecidos, porém, as diferenças impossibilitam a comparação direta entre os códigos das contas contábeis das empresas.

Plano Contábil com 4 Níveis de Detalhamento				
Conta Contábil				Descrição da Conta Contábil
1				Ativo
1	1			Ativo Circulante
1	1	1		Caixa e Equivalentes de Caixa
1	1	1	001	Caixa
1	1	1	002	Bancos
1	1	2		Contas a Receber
1	1	2	001	Clientes
1	1	2	002	Duplicatas
1	1	3		Estoque
1	1	3	001	Mercadorias
1	1	3	002	Produtos Acabados
1	1	4		Outros Créditos
1	1	4	001	Impostos a Recuperar
1	1	4	002	Outros Valores a Receber
1	1	5		Aplicações Financeiras
1	1	5	001	Ações
1	1	5	002	Debêntures
1	2			Ativo Não Circulante
1	2	1		Realizável a Longo Prazo
1	2	1	001	Clientes
1	2	1	002	(-) Perdas Estimadas com Créditos de Liquidação Duvidosa
1	2	2		Investimentos
1	2	2	001	Participações Societárias
2				Passivo
2	1			Passivo Circulante
2	1	1		Fornecedores
2	1	1	001	Nacional
2	1	1	002	Estrangeiro
2	1	2		Empréstimos e Financiamentos
2	1	2	001	Emprestimos
2	1	2	002	Financiamento
2	1	3		Obrigações Fiscais
2	1	3	001	Simples Nacional
2	1	3	002	ICMS a Recolher
2	1	4		Obrigações Trabalhistas e Sociais
2	1	4	001	Salários a Pagar
2	1	4	002	FGTS a Recolher
2	1	5		Contas a Pagar
2	1	5	001	Telefone a Pagar
2	1	5	002	Aluguel
2	1	5	002	Energia a Pagar
2	2			Passivo Não Circulante
2	2	1		Financiamentos
2	2	1	001	Financiamentos Bancos
2	2	1	002	Financiamento Estrangeiro
2	2	2		Passivo Exigível a Longo Prazo
2	2	2	001	Emprestimos
2	3			Patrimonio Líquido
2	3	1		Capital Social
2	3	1	001	Capital Subscrito
2	3	1	002	(-) Capital a Integralizar
2	3	2		Reservas
2	3	2	001	Reservas de Capital
2	3	2	002	Reservas de Lucros

Quadro 2: Exemplo de Plano de Contas. Criado pelo autor.

Como já observado, os planos de contas são códigos agrupados pela conta superior. No exemplo do Quadro 2 existem 4 níveis de contas, sendo o último nível considerado o nível analítico, onde são lançados os valores de crédito e débito e os três superiores são acumuladores ou sintéticos.

Nesta estrutura existem contas analíticas criadas conforme a necessidade da empresa, porém as contas sintéticas de nível 1 a 3 costumam ser mais padronizadas. Em toda a contabilidade existe o Ativo e o Passivo, assim como os Ativos e Passivos, Circulante e Não Circulante. Dentro da estrutura do Passivo está o Patrimônio Líquido e algumas outras contas acabam tendo uma nomenclatura muito próxima. Entretanto, a codificação entre os planos de contas costuma ser diferentes.

3.2.2. Movimentação Contábil

A normalização do plano de contas é importante pelo fato dos lançamentos contábeis utilizarem apenas os códigos dos planos de conta e não suas descrições.

Conta Débito	Conta Crédito	Data do Lançamento	Valor	Observação
1.1.1.001	2.1.1.001	20/11/2023	R$ 150,00	Copo de água
2.1.1.001	1.1.3.001	20/11/2023	R$ 150,00	Copo de água
1.1.2.001	1.1.1.001	21/11/2023	R$ 500,00	NF 0032

Tabela 1: Exemplo de movimentações contábeis. Criado pelo autor.

Os códigos dos lançamentos sozinhos não são compreendidos sem a leitura deles no plano contábil gerador. No exemplo da Tabela 1 podemos observar três lançamentos contábeis. Os dois primeiros pertencem a uma mesma movimentação, no caso a compra de Copos de Água. No primeiro lançamento temos a saída do dinheiro do Caixa da empresa para o pagamento do Fornecedor Nacional. Em seguida, é lançado a entrada do produto no Estoque de Mercadorias da empresa, completando a operação. Sendo assim, a questão é normalizar a codificação dos planos de contas em uma única estrutura ajustada a todos os planos de contas existentes.

Neste ponto a primeira decisão a ser tomada é o nível a ser considerado na normalização. Para este trabalho, foi levado em consideração somente os níveis sintéticos, uma vez ser possível calcular todos os índices financeiros e comparar as empresas sem utilizar as contas analíticas. No exemplo acima, a compra sairia do

“Caixa e Equivalente de Caixa” para a conta Fornecedor e após entraria no Estoque. Mesmo não especificando exatamente de onde saiu e entrou, estão claras as movimentações financeiras realizadas e a contabilização dos lançamentos.

3.2.3. Conceito Utilizado para a Normalização das Contas Contábeis

Inicialmente foi criado um plano de contas padrão, com quatro níveis sintéticos, com ao todo 208 contas, 4 contas nível um, 12 contas nível dois, 42 contas nível três e 150 contas nível quatro.

O problema passou a ser a ação de comparação e troca dos códigos das contas de forma confiável. Contas nível um, normalmente são “Ativo”, “Passivo”, “Receitas, Custos e Despesas” e “Contas de Apuração”, sendo a padronização relativamente fácil de ser feita. Porém, no nível quatro temos situações bem mais complexas, por ter descrições possíveis de ser repetidas, como a conta “Cliente” poder existir tanto em “Ativo Circulante” como em “Ativo Não Circulante”. Entretanto, pode-se observar um padrão comparativo ao concatenar a descrição de todos os níveis. No Quadro 3 pode-se observar melhor estes padrões.

Plano Contábil com 4 Níveis de Detalhamento					Descrição Completa do Plano de Contas
Conta Contábil				Descrição da Conta Contábil	
1				Ativo	Ativo
1	1			Ativo Circulante	Ativo Ativo Circulante
1	1	1		Caixa e Equivalentes de Caixa	Ativo Ativo Circulante Caixa e Equivalentes de Caixa
1	1	1	001	Caixa	Ativo Ativo Circulante Caixa e Equivalentes de Caixa Caixa
1	1	1	002	Bancos	Ativo Ativo Circulante Caixa e Equivalentes de Caixa Bancos
1	1	2		Contas a Receber	Ativo Ativo Circulante Contas a Receber
1	1	2	001	Clientes	Ativo Ativo Circulante Contas a Receber Clientes
1	1	2	002	Duplicatas	Ativo Ativo Circulante Contas a Receber Duplicatas

Quadro 3: Padrões de concatenação de descrição dos planos de contas. Elaborado pelo autor.

No Quadro 3 a concatenação das descrições forma frases comparáveis. Comparando “Ativo Ativo Circulante Contas a Receber Clientes” ou “Ativo Ativo Circulante Contas a Receber Duplicatas” com “Ativo Ativo Circulante Contas a Receber” teremos um grau de proximidade muito alto, este não observado se forem

comparadas as duas frases com “Ativo Ativo Circulante Caixa e Equivalentes de Caixa”.

O problema pode ser resumido em criar um plano de contas compacto para recepcionar todas as contas dos clientes. Esta transformação de um código em outro seria realizada pelo treinamento das descrições completas com os novos códigos. Este plano de contas compacto tem como objetivo possibilitar o cálculo padronizado de todos os índices financeiros e poder comparar a evolução mensal da empresa em relação a outras empresas do mesmo setor.

A formulação destas 208 contas levou em consideração estas necessidades e refletem a atual situação das empresas presentes na base de dados analisada. Mesmo sendo apenas 208 contas, o treinamento de textos foi feito com 338 descrições completas; sendo assim, uma conta pode ter mais de uma descrição treinada para ela, assegurando levar em conta as variações criadas pelos contadores.

Este treinamento foi realizado com o método descrito na seção 2.4 e foram criados os vetores para cada uma das 338 descrições completas definidas para as 208 contas destino. Sendo assim, um plano de contas contábil pode ser rapidamente convertido nas 208 contas normalizadas levando em consideração a descrição completa mais próxima, sendo a conta normalizada inserida em todo o sistema de movimentações financeiras.

A base de treinamento de descrições completas pode aumentar de acordo com a acuracidade verificada nas contas contábeis dos planos de contas analisados. O treinamento atual garantiu uma acuracidade mínima de 86,5%.

3.3. Geração dos Vetores Contábeis.

Com os planos de contas normalizados, foram criados os vetores contendo os dados contábeis e os índices financeiros calculados para cada mês. Além destas informações, os vetores possuem o setor de atividade econômica e o estado da federação da sede da empresa.

O vetor principal contém as informações de um mês contábil. Estas informações foram geradas com o processamento de todas as movimentações contábeis de cada empresa e acumulando nas contas contábeis mensalmente, gerando assim o balancete mensal para cada mês de operação de cada empresa. Com este balancete gerado, foram calculados todos os índices financeiros do mês, assim como as análises horizontal e vertical das contas contábeis.

Os índices escolhidos para a construção do vetor a ser utilizado nos modelos preditivos estão listados no Quadro 4. São ao todo vinte e um índices financeiros normalmente utilizados pelo mercado. Os índices são calculados utilizando contas contábeis específicas e estas contas acabam sendo utilizadas mais de uma vez, sendo possível a geração de vieses ao analisar os dados em estruturas estatísticas ou de treinamento de máquina.

As contas contábeis formadoras do balancete normalmente são analisadas de forma horizontal e vertical. A análise horizontal compara as contas ao longo do tempo, levando em consideração a variação existente de um período para outro. Por sua vez, a análise vertical é realizada expressando cada conta contábil como uma porcentagem do total de ativos existentes na empresa. Estas análises ajudam a avaliar tendências, estrutura financeira, eficiência e as comparações com outras empresas.

Índices Financeiros Escolhidos	
Liquidez Seca	
Liquidez Imediata	
Ativo Permanente	% Ativo Permanente
Liquidez Corrente	
Imobilizado Recursos	
Imobilizado Patrimônio	
Composição do Endividamento	
Capital Circulante Líquido	% Capital Circulante Líquido
Capital Circulante Próprio	% Capital Circulante Próprio
Realizável Longo Prazo	% Realizável Longo Prazo
Liquidez Geral	
EBITDA	% EBITDA
Custo Operacional	% Custo Operacional
Custo Financeiro	% Custo Financeiro
Margem Líquida	
Margem Contribuição	
Carga Tributária	
Margem Bruta	
Despesas	% Despesas
Crescimento	
Equilíbio	

Quadro 4: Índices selecionados. Criado pelo autor.

Sendo assim, foi criado um vetor com as contas contábeis mais utilizadas nos índices financeiros e com maior poder de informações sobre a saúde financeira da empresa. Para cada conta foi calculado os percentuais das análises horizontal e vertical. No Quadro 5 estão listadas as contas contábeis utilizadas.

Contas Padronizadas	
Número	Descrição
11100	Ativo circulante disponível em caixa
11200	Ativo circulante clientes
11300	Ativo circulante estoque
11500	Ativo circulante aplicações financeiras
13200	Ativo não-circulante investimentos
13300	Ativo não-circulante imobilizado
13400	Ativo não circulante intangível
21100	Passivo circulante fornecedores
21200	Passivo circulante empréstimos e financiamentos
22000	Passivo não-circulante
22300	Passivo não-circulante passivo exigível a longo prazo
23000	Passivo patrimônio líquido
31100	Receitas de venda
32100	Custos dos produtos, mercadorias e serviços vendidos
32200	Despesas operacionais
32300	Despesas financeiras
32400	Outras despesas operacionais
32500	Provisão para imposto de renda pessoa jurídica e contribuição social sobre o lucro líquido
32700	Custos diretos da produção de serviços
32800	Custos dos serviços prestados
32900	Despesas operacionais despesas administrativas

Quadro 5: Contas contábeis selecionadas. Criado pelo autor.

3.4.Amostragem

Para validar o método proposto neste estudo, três divisões ou setores de atividade econômica foram selecionados; “comércio varejista”, “atividades de atenção à saúde humana” e “serviços de escritório, de apoio administrativo e outros serviços prestados principalmente às empresas”. Também foram escolhidos dois grupos bastante distintos do comércio varejista para aprofundar a análise; 472 (bens de consumo não durável: comércio. varejista de produtos alimentícios, bebidas e fumo) e 475 (bens de consumo durável: comércio varejista de equipamentos de informática e comunicação; equipamentos e artigos de uso doméstico). As quantidades das empresas do comércio varejista por grupo estão listadas no Quadro 6 e a escolha foi realizada pelo fato de ser um dos setores com maior quantidade de empresas na base trabalhada.

Comércio Varejista			
Divisão	Grupo	Empresas	Descrição
47	471	1.051	Comércio varejista não-especializado
47	472	1.131	Comércio varejista de produtos alimentícios, bebidas e fumo
47	473	389	Comércio varejista de combustíveis para veículos automotores
47	474	1.266	Comércio varejista de material de construção
47	475	1.545	Comércio varejista de equipamentos de informática e comunicação; equipamentos e artigos de uso doméstico
47	476	377	Comércio varejista de artigos culturais, recreativos e esportivos
47	477	1.111	Comércio varejista de produtos farmacêuticos, perfumaria e cosméticos e artigos médicos, ópticos e ortopédicos
47	478	2.659	Comércio varejista de produtos novos não especificados anteriormente e de produtos usados
Total 47		9.529	

Quadro 6: Comércio varejista com a quantidade de empresas. Pelo autor.

A divisão 82 (Serviços: de escritório, apoio administrativo e outros serviços prestados principalmente às empresas) e seus números estão na Quadro 7. Esta divisão foi escolhida para servir de teste de treinamento como uma divisão diferente.

Serviços de Escritório, de Apoio Administrativo e Outros Serviços Prestados Principalmente às Empresas			
Divisão	Grupo	Empresas	Descrição
82	821	1.373	Atividades de atendimento hospitalar
82	822	16	Serviços móveis de atendimento a urgências e de remoção de pacientes
82	823	216	Atividades de atenção ambulatorial executadas por médicos e odontólogos
82	829	403	Atividades de atenção à saúde humana não especificadas anteriormente
Total 82		2.008	

Quadro 7: Serviços de Escritório, de Apoio Adm. E Outros. Pelo autor.

As quantidades das empresas de atividades de atenção à saúde humana por grupo estão listadas no Quadro 8. Esta divisão foi escolhida por ser de um setor muito

diferente do primeiro setor escolhido e também por possuir uma boa quantidade de empresas na base de dados.

Atividades de Atenção à Saúde Humana			
Divisão	Grupo	Empresas	Descrição
86	861	241	Atividades de atendimento hospitalar
86	862	6	Serviços móveis de atendimento a urgências e de remoção de pacientes
86	863	1.746	Atividades de atenção ambulatorial executadas por médicos e odontólogos
86	864	257	Atividades de serviços de complementação diagnóstica e terapêutica
86	865	582	Atividades de profissionais da área de saúde, exceto médicos e odontólogos
86	866	31	Atividades de apoio à gestão de saúde
86	869	56	Atividades de atenção à saúde humana não especificadas anteriormente
Total 86		2.919	

Quadro 8: Atividades de atenção à saúde humana com a quantidade de empresas. Pelo autor.

Desse modo, a seleção desses três setores distintos de atividade econômica contribui para aumentar a capacidade de generalização do modelo proposto na tese.

Por último, há que se destacar que os dados utilizados para a construção do modelo preditivo correspondem a 416.710 vetores de dados mensais para cada estrutura de informação (índices financeiros e análises horizontal e vertical) na divisão 47, 116.487 vetores na divisão 86 e 75.373 vetores na divisão 82. Estes vetores foram alimentados em uma base de dados padrão SQL de alta performance.

3.5. Construção do Modelo Preditivo

O objetivo do estudo é propor e testar um modelo preditivo capaz de identificar empresas com probabilidade de entrar em falência. Esta ação pode ajudar o setor financeiro a analisar o risco de crédito e assim definir produtos mais customizados para os clientes.

Os estudos existentes utilizaram os dados de empresas abertas pelo fato de os dados estarem disponíveis. O balanço anual é tratado utilizando os índices financeiros ou em comparação com o balanço anterior. Este tipo de abordagem consegue fornecer um grupo de variáveis possíveis de treinar um modelo com *random forest* e *gradient boosting,* sendo utilizados em muitos dos estudos já citados. Porém, estes pacotes de dados não levam em consideração a evolução da empresa em períodos menores dentro do ano, podendo a defasagem da informação gerar resultados falsos.

Quando os dados estão em transformação ao longo do tempo, é comum a utilização de séries temporais. Entretanto, os estudos mais recentes utilizaram redes neurais como *convolutional neural network* (CNN) (Jin et al., 2020), *recurrent neural network* (RNN) (Madan e Mangipudi, 2018) e *artificial neural network* (ANN) com *multilayer perceptron* (MLP) (Aktan, 2011) conseguindo bons resultados.

No caso do estudo de Aktan (2011), a rede neural foi utilizada com dados de um, dois e três anos anteriores e o modelo foi perdendo acurácia, obtendo 90,0%, 85,7% e 76,3% respectivamente. Esta perda de performance é um dos problemas encontrado em modelos temporais e é chamado “efeito de recência” ou “viés de recência”; isto ocorre quando os dados mais recentes acabam tendo um maior peso na decisão do modelo. A ideia de ter uma “fotografia” com a evolução da empresa passou a ser levada em consideração. Pois, como a base de dados possui mês a mês a evolução contábil da empresa por anos, a construção do modelo ganha complexidade.

Zhao et al. (2017) demonstraram a vantagem da utilização das redes neurais convolucionais (CNN) para tratamento das séries temporais e, como citado anteriormente, Hosaka (2019) sugeriu que CNN são mais adequadas para aplicação em imagens e trabalhou com uma imagem em tons de cinza com as informações diretas dos índices gerados pelo balanço anual, sendo outro estudo realizado com bons resultados. Os estudos de Jin et al. (2020) e Hosaka (2019) mostram em seus estudos a

necessidade de utilizar muitas imagens para o treinamento do modelo CNN, sendo considerada uma limitação na utilização desse método.

A proposta do modelo é identificar as empresas com risco de falência, portanto as 14.456 empresas foram rotuladas como 1 (falidas) ou 0 (empresas abertas e em operação). As empresas abertas, mas não em operação, foram classificadas como 1. Esta classificação foi usada no treinamento do modelo em todos os casos testados.

Com os balancetes disponíveis e classificados, as duas formas de comparar as empresas através dos balancetes são pelos índices obtidos ou pelas análises horizontal e vertical. Balanços são dados financeiros em um determinado momento da vida de uma empresa e apresentam problemas para representar as variações ocorridas nos meses antecedentes (Krugman, 1999). Isto gera distorções nas análises dos balanços de uma empresa quando comparados um ano com o outro, justamente a falta do entendimento das ocorrências entre dois relatórios. Entretanto, a amostra utilizada neste estudo possui os dados mensais das empresas, fornecendo uma fotografia da empresa a cada mês.

O desafio deste trabalho passou a ser encontrar uma forma de identificar a série temporal dos 12 últimos meses operacionais das empresas para entender os eventos antecessores ao último balanço ou a parada das operações.

Utilizando o conceito das redes neurais tratarem imagens com uma acuracidade maior quando comparado ao tratamento de números, as variáveis selecionadas dos 12 últimos meses das 14.456 empresas foram utilizadas na criação de uma imagem contendo as informações financeira onde cada pixel representa uma variável no mês e a variação da tonalidade da cor o valor. Com a utilização de uma imagem representando a série temporal, a utilização da CNN passou a ser considerada como a forma mais para analisar as imagens. A Figura 2 é um exemplo de uma imagem representando doze meses das contas contábeis de uma empresa.

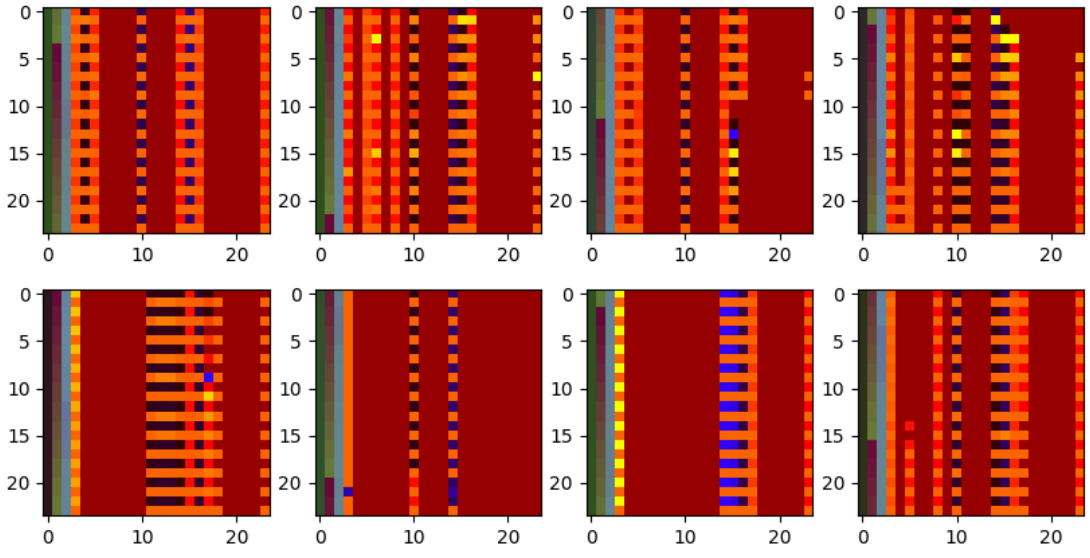

Figura 2: Imagens dos doze meses das contas contábeis de oito empresas. Pelo autor.

A imagem pode alimentar uma rede neural convolucional com o intuito desta rede passar a ter a capacidade de prever a falência das empresas.

Foi escolhida a estrutura de *deep learning* do *framework Keras*, por ser uma estrutura de fácil acesso e utilizando a linguagem de programação amplamente utilizada, Python (Chollet, 2021).

Keras foi desenvolvido por François Chollet e é mantido por ele e outros colaboradores como código aberto. François Chollet é um dos colaboradores no desenvolvimento do *TensorFlow*, a plataforma de aprendizado de máquina de código aberto do Google integrada com o *Keras*.

Um dos usos do *Keras* é o processamento de imagens, onde o sistema é alimentado com imagens rotuladas na fase de treinamento e tem um alto nível de acuracidade. Um dos exemplos muito utilizado como aprendizado de *deep learning* é o MNIST (https://keras.io/examples/vision/mnist_convnet/), um banco de dados de imagens de dígitos numéricos manuscritos.

Neste estudo foi utilizado o modelo sequencial do *Keras* com CNN. Na criação da imagem foram considerados alguns cuidados como a imagem a criada não poder

sofrer transformação. Cada pixel tem uma posição bem definida e uma cor obrigatória. Sendo assim, ela não pode ser rotacionada, redimensionada, invertida ou inclinada. Estas transformações são normais em imagens comuns, como as de animais e objetos. Servem para fazer uma imagem gerar outras imagens com variações, aumentando a base de treinamento. Ainda nesta abordagem, as imagens serão quadradas e com tamanho definido.

Além dos dados contábeis, foi definida a inserção das seguintes informações para influenciar na tomada de decisão do método:

- divisão e grupo da empresa
- número do estado da empresa fornecido pelo IBGE
- período dos dados
- código do país
- índice da inflação (IPCA do mês e dos últimos doze meses).

Na seção 3.3 foram definidos 21 índices financeiros e 21 contas contábeis, sendo estas 21 com a análise horizontal e 21 com a análise vertical. Como comentado na seção 3.3, os índices decorrem de fórmulas definidas com as contas contábeis e dados financeiros; sendo assim, foi definida a criação de uma imagem para os índices financeiros e outra imagem para as contas contábeis.

3.5.1. Criação das Imagens

As imagens construídas têm 24 linhas e 24 colunas com os pixels em padrão RGB (forma de representar as cores, o acrônimo vem do nome das cores em inglês: *red, green* e *blue*) onde cada cor é definida por um trio de números entre 0 e 255 e cada um deles define a intensidade da cor vermelha, verde e azul. Por exemplo a cor azul pura é definida como (0, 0, 255)

O primeiro pixel representa a divisão e grupo da empresa, sendo o primeiro número a divisão, o segundo número a unidade do grupo multiplicada por 10 e o terceiro número o estado. Por exemplo, uma empresa de São Paulo do grupo 472 tem o primeiro pixel (47, 20, 35).

O segundo pixel representa o período e o país, sendo o primeiro pixel o ano subtraído de 1970 e multiplicado por 2 (forma adotada para ficar entre 0 e 255), o segundo o mês multiplicado por 10 e o terceiro o código do país da União Internacional de Telecomunicações (UIT). Por exemplo, o pixel de uma empresa brasileira em janeiro de 2020 tem o segundo pixel (100, 10, 55).

O terceiro pixel representa a inflação do período, sendo o primeiro pixel o tipo de índice (IPCA foi definido como 100) o segundo pixel a raiz quadrada da variação percentual do mês multiplicada por 100 e o terceiro a variação dos últimos 12 meses com o mesmo cálculo. Foi definido o 125 como zero, sendo assim os resultados são somados a 125 e se considera a parte inteira do número. Por exemplo, uma inflação do mês em 0,54 e a anual em 7,02064 (inflação de junho de 1997) gera o pixel (100, 132, 151).

Os dois tipos de imagens têm os outros pixels definidos pelos conteúdos das variáveis. Para poder ter mais informação por imagem foi utilizado o tamanho 24x24, sendo assim o período possui 2 linhas.

Na imagem referente aos índices financeiros, as duas linhas do período são iguais. Porém, na imagem das contas contábeis uma linha contém a análise vertical e a outra a análise horizontal. No caso da imagem dos índices, esta poderia conter dois anos de informação; neste estudo foram utilizados 12 meses para poder comparar o resultado entre as duas imagens.

Efetuada a descrição dos procedimentos metodológicos empregados na pesquisa empírica, apresenta-se, no capítulo subsequente, os resultados alcançados no processamento de dados realizado.

4. RESULTADOS

Para a realização dos procedimentos de validade e confiabilidade, as imagens totais foram separadas em dois grupos; um com 80% para construir o modelo de treinamento e outro com 20% para o modelo de validação. A seleção das empresas em um ou outro grupo foi feita de forma aleatória. Esta separação foi feita para o modelo não conhecer o resultado dos dados utilizados para validação.

As imagens do treinamento (80%) foram separadas pelo *Keras* em 90% para treino e 10% para validação (neste caso o modelo conhece o resultado de todos os dados). Nos resultados estão publicadas as perdas e acuracidade do modelo em fase de treino e a matriz de confusão das imagens separas para o teste.

A seguir estão os resultados dos 11 treinamentos e 4 testes cruzados realizados. Todos os processamentos realizados levaram em consideração as seguintes configurações:

```
rescale=1./255.,
rotation_range=0,
zoom_range=0.0,
width_shift_range=0.0,
height_shift_range=0.0,
shear_range=0.0,
horizontal_flip=False,
fill_mode="nearest",
validation_split=0.1
target_size=(24, 24)
class_mode="categorical",
subset='validation',
shuffle=True,
seed=42
batch_size = 8
epochs = 5
```

4.1. Resultados do Treinamento

Para a construção do modelo de treinamento foram utilizados os dados dos índices financeiros e das contas contábeis das empresas das diferentes divisões analisadas (86, 47 e 82) e grupos (478, 475 e 472). Ressalte-se que somente foram utilizados os dados válidos de cada empresa.

As tabelas apresentadas em seguida correspondem aos resultados alcançados para cada divisão e grupo considerados, levando-se em consideração inclusive a união entre divisões.

Divisão 86 com Índices Financeiros				
Total de imagens geradas: 2.043				
Imagens usadas no treinamento:		1.475		
Imagens usadas na validação:		163		
Imagens separadas para o teste:		405		
Treinamento				
Perda do modelo:		77,21292%		
Acuracidade do modelo:		58,89571%		
Matriz de Confusão				
Divisão 86 Indices		Real		
		0	1	% Acerto
Predito	0	194	28	87,4%
	1	91	92	50,3%
	% Acerto	68,1%	76,7%	70,6%

Divisão 86 com Contas Contábeis				
Total de imagens geradas: 2.043				
Imagens usadas no treinamento:		1.470		
Imagens usadas na validação:		163		
Imagens separadas para o teste:		410		
Treinamento				
Perda do modelo:		65,75513%		
Acuracidade do modelo:		65,03068%		
Matriz de Confusão				
Divisão 86 Contas Contábeis		Real		
		0	1	% Acerto
Predito	0	183	36	83,6%
	1	81	110	57,6%
	% Acerto	69,3%	75,3%	71,5%

Tabela 2: Resultados da divisão 86 com IF e CC. Pelo autor.

Divisão 47 com Índices Financeiros				
Total de imagens geradas: 6.417				
Imagens usadas no treinamento:		4.618		
Imagens usadas na validação:		513		
Imagens separadas para o teste:		1.286		
Treinamento				
Perda do modelo:		24,23577%		
Acuracidade do modelo:		92,20273%		
Matriz de Confusão				
Divisão 47 Indices		Real		
		0	1	% Acerto
Predito	0	612	69	89,9%
	1	20	585	96,7%
	% Acerto	96,8%	89,4%	93,1%

Divisão 47 com Contas Contábeis				
Total de imagens geradas: 6.417				
Imagens usadas no treinamento:		4.661		
Imagens usadas na validação:		517		
Imagens separadas para o teste:		1.239		
Treinamento				
Perda do modelo:		8,53512%		
Acuracidade do modelo:		97,48549%		
Matriz de Confusão				
Divisão 47 Contas Contábeis		Real		
		0	1	% Acerto
Predito	0	624	29	95,6%
	1	4	582	99,3%
	% Acerto	99,4%	95,3%	97,3%

Tabela 3: Resultados da divisão 47 com IF e CC. Pelo autor.

Divisão 47 e 86 com Índices Financeiros				
Total de imagens geradas: 8.460				
Imagens usadas no treinamento:		6.075		
Imagens usadas na validação:		675		
Imagens separadas para o teste:		1.710		
Treinamento				
Perda do modelo:		10,86181%		
Acuracidade do modelo:		95,70370%		
Matriz de Confusão				
Divisão 47 + 86 Indices		Real		
		0	1	% Acerto
Predito	0	913	27	97,1%
	1	33	757	95,8%
	% Acerto	96,5%	96,6%	96,5%

Divisão 47 e 86 com Contas Contábeis				
Total de imagens geradas: 8.460				
Imagens usadas no treinamento:		6.057		
Imagens usadas na validação:		673		
Imagens separadas para o teste:		1.730		
Treinamento				
Perda do modelo:		10,42625%		
Acuracidade do modelo:		96,43388%		
Matriz de Confusão				
Divisão 47 + 86 Contas Contábeis		Real		
		0	1	% Acerto
Predito	0	896	30	96,8%
	1	8	787	99,0%
	% Acerto	99,1%	96,3%	97,8%

Tabela 4: Resultado conjunto das divisões 47 e 86 com IF e CC. Pelo autor.

Grupo 478 com Índices Financeiros				
Total de imagens geradas: 1.788				
Imagens usadas no treinamento:		1.260		
Imagens usadas na validação:		140		
Imagens separadas para o teste:		388		
Treinamento				
Perda do modelo:		62,64799%		
Acuracidade do modelo:		60,71429%		
Matriz de Confusão				
Grupo 478 Indices		Real		
		0	1	% Acerto
Predito	0	179	22	89,1%
	1	109	78	41,7%
	% Acerto	62,2%	78,0%	66,2%

Grupo 478 com Contas Contábeis				
Total de imagens geradas: 1.788				
Imagens usadas no treinamento:		1.291		
Imagens usadas na validação:		143		
Imagens separadas para o teste:		354		
Treinamento				
Perda do modelo:		60,27632%		
Acuracidade do modelo:		64,33566%		
Matriz de Confusão				
Grupo 478 Contas Contábeis		Real		
		0	1	% Acerto
Predito	0	140	35	80,0%
	1	51	128	71,5%
	% Acerto	73,3%	78,5%	75,7%

Tabela 5: Resultados do grupo 478 com IF e CC. Pelo autor.

Grupo 475 com Contas Contábeis				
Total de imagens geradas: 1.021				
Imagens usadas no treinamento:			726	
Imagens usadas na validação:			80	
Imagens separadas para o teste:			215	
Treinamento				
Perda do modelo:			66,29423%	
Acuracidade do modelo:			61,25000%	
Matriz de Confusão				
Grupo 475 Contas Contábeis		Real		
		0	1	% Acerto
Predito	0	99	44	69,2%
	1	47	25	34,7%
	% Acerto	67,8%	36,2%	57,7%

Grupo 472 com Contas Contábeis				
Total de imagens geradas: 1.244				
Imagens usadas no treinamento:			496	
Imagens usadas na validação:			55	
Imagens separadas para o teste:			142	
Treinamento				
Perda do modelo:			66,70751%	
Acuracidade do modelo:			56,36364%	
Matriz de Confusão				
Grupo 472 Contas Contábeis		Real		
		0	1	% Acerto
Predito	0	48	23	67,6%
	1	31	40	56,3%
	% Acerto	60,8%	63,5%	62,0%

Tabela 6: Resultados dos grupos 475 e 472 com CC. Pelo autor.

Divisão 82 com Contas Contábeis				
Total de imagens geradas: 1.351				
Imagens usadas no treinamento:			971	
Imagens usadas na validação:			108	
Imagens separadas para o teste:			270	
Treinamento				
Perda do modelo:			71,10168%	
Acuracidade do modelo:			50,00000%	
Matriz de Confusão				
Divisão 82 Contas Contábeis		Real		
		0	1	% Acerto
Predito	0	101	47	68,2%
	1	32	90	73,8%
	% Acerto	75,9%	65,7%	70,7%

Tabela 7: Resultados da divisão 82 com contas contábeis. Pelo autor.

4.2. Teste Cruzado

A realização deste teste é importante para entender se a CNN tem a capacidade de ser treinada em uma divisão e este modelo treinado ser utilizado para fazer a predição em dados de empresas de outra divisão. Este teste utilizou no treinamento os dados das contas contábeis das empresas da divisão 47 e foi testado com a divisão 82 e 86. Assim como a mistura do treinamento com mais de uma divisão e a utilização dos

testes com apenas os dados de uma delas e por fim o treinamento com uma divisão e a utilização de um grupo pertencente a esta divisão como teste.

Treino Divisão 47 / Teste Divisão 82 com Contas Contábeis				
Total de imagens utilizadas: 5.448				
Imagens usadas no treinamento:	4.661			
Imagens usadas na validação:	517			
Imagens separadas para o teste:	270			
Treinamento				
Perda do modelo:	8,53512%			
Acuracidade do modelo:	97,48549%			
Matriz de Confusão				
Divisão 47 / Teste 82		Real		
Contas Contábeis		0	1	% Acerto
Predito	0	16	132	10,8%
	1	0	122	100,0%
	% Acerto	100,0%	48,0%	51,1%

Treino Divisão 47 / Teste Divisão 86 com Contas Contábeis				
Total de imagens utilizadas: 5.588				
Imagens usadas no treinamento:	4.661			
Imagens usadas na validação:	517			
Imagens separadas para o teste:	410			
Treinamento				
Perda do modelo:	8,53512%			
Acuracidade do modelo:	97,48549%			
Matriz de Confusão				
Divisão 47 / Teste 86		Real		
Contas Contábeis		0	1	% Acerto
Predito	0	15	204	6,8%
	1	0	191	100,0%
	% Acerto	100,0%	48,4%	50,2%

Tabela 8: Resultados treino divisão 47 e teste divisão 82 com contas contábeis. Pelo autor.

Treino Divisão 47 e 86 / Teste Divisão 86 com Contas Contábeis				
Total de imagens utilizadas: 7.140				
Imagens usadas no treinamento:	6.057			
Imagens usadas na validação:	673			
Imagens separadas para o teste:	410			
Treinamento				
Perda do modelo:	10,42625%			
Acuracidade do modelo:	96,43388%			
Matriz de Confusão				
Divisão 47 e 86 / Teste 86		Real		
Contas Contábeis		0	1	% Acerto
Predito	0	213	6	97,3%
	1	23	168	88,0%
	% Acerto	90,3%	96,6%	92,9%

Treino Divisão 47 / Teste Grupo 472 com Contas Contábeis				
Total de imagens utilizadas: 5.320				
Imagens usadas no treinamento:	4.661			
Imagens usadas na validação:	517			
Imagens separadas para o teste:	142			
Treinamento				
Perda do modelo:	8,53512%			
Acuracidade do modelo:	97,48549%			
Matriz de Confusão				
Divisão 47 / Teste 472		Real		
Contas Contábeis		0	1	% Acerto
Predito	0	69	3	95,8%
	1	2	68	97,1%
	% Acerto	97,2%	95,8%	96,5%

Tabela 9: Resultados treino divisão 47 e teste grupo 472 com contas contábeis. Pelo autor

5. ANÁLISE DOS RESULTADOS

Todos os treinamentos realizados utilizaram 80% das imagens para treinamento e validação e 20% das imagens para o teste. O treinamento utilizou 90% para o treino do modelo e 10% para a validação dele. A separação das imagens para treino e teste foi realizada de forma aleatória. O código utilizado para a criação das imagens está no Apêndice I. No caso do treinamento, a separação foi realizada de forma aleatória pela própria plataforma do Keras.

Os resultados dos treinamentos foram analisados utilizando as premissas da análise das redes neurais; ou seja, o teste de perda e a acuracidade do modelo. Os testes foram analisados através da "Matriz de Confusão".

Treinamento	Tipo Imagem	Treino			Teste		Perda	Acuracidade	Acerto Teste
		Divisão / Grupo	Qtd Treino	Qtd Validação	Divisão / Grupo	Quantidade			
1	Índices Financeiro	86	1.475	163	86	405	0,77212924	0,58895707	70,6%
2	Contas Contábeis	86	1.470	163	86	410	0,65751289	0,65020676	71,5%
3	Índices Financeiro	47	4.618	513	47	1.286	0,24235767	0,92202729	93,1%
4	**Contas Contábeis**	**47**	**4.661**	**517**	**47**	**1.239**	**0,08535116**	**0,97485495**	**97,3%**
5	Índices Financeiro	47 e 86	6.075	675	47 e 86	1.710	0,10861805	0,95703703	96,5%
6	**Contas Contábeis**	**47 e 86**	**6.057**	**673**	**47 e 86**	**1.730**	**0,10426246**	**0,96433878**	**97,8%**
7	Índices Financeiro	478	1.260	140	478	388	0,62647986	0,60714287	66,2%
8	Contas Contábeis	478	1.291	143	478	354	0,60276318	0,64335662	75,7%
9	Contas Contábeis	475	726	80	475	215	0,66294229	0,61230001	57,7%
10	Contas Contábeis	472	496	55	472	142	0,66707510	0,56363636	62,0%
11	Contas Contábeis	82	973	108	82	270	0,71101683	0,50000000	70,7%
12	Contas Contábeis	47	4.661	517	82	270	0,08535116	0,97485495	51,1%
13	Contas Contábeis	47	4.661	517	86	410	0,08535116	0,97485495	50,2%
14	**Contas Contábeis**	**47 e 86**	**6.057**	**673**	**86**	**410**	**0,10426246**	**0,96433878**	**92,9%**
15	**Contas Contábeis**	**47**	**4.661**	**517**	**472**	**142**	**0,08535116**	**0,97485495**	**96,5%**

Tabela 10: Dados consolidados dos testes realizados (Apêndice II). Pelo autor

O teste de perda em uma Rede Neural Convolucional serve para mensurar a eficácia e a confiabilidade de um modelo. O teste de perda é uma função matemática e tem a tarefa de quantificar a diferença entre as previsões feitas pelo modelo e os valores reais ou esperados. O resultado é um valor numérico mensurando a precisão do modelo através das perdas, onde valores menores indicam uma maior precisão do modelo. Valores acima de 0,20 já indicam a possibilidade de ter ocorrido *overfitting* (Yu et al., 2020). O teste de perda deve ser avaliado em conjunto com a acuracidade do modelo.

A função acurácia é utilizada como métrica de desempenho para avaliar a proporção de previsões corretas em relação com o total de previsões realizadas (LeCun et al., 2015).

A matriz de confusão utilizada para a análise das imagens de teste é uma tabela contendo a quantidade de imagens classificadas corretamente e erroneamente. Neste trabalho foi adotada uma classificação binária, sendo assim, a matriz apresenta a quantidade de verdadeiro positivo, falso positivo, verdadeiro negativo e falso negativo. Na Tabela 10, o acerto do teste foi obtido pela divisão do total dos acertos (positivos e negativos) pelo total de testes realizados.

$$\text{Acerto Teste} = \frac{(\text{Verdadeiro Positivo} + \text{Verdadeiro Negativo})}{(\text{Total de Testes})}$$

Tabela 11: Cálculo do acerto do teste.

5.1. Influência da quantidade de imagens na fase de treinamento

Pode-se observar pelos resultados como a quantidade de imagens no treinamento da rede neural convolucional afeta sensivelmente o resultado. Independente dos dados geradores das imagens, a acuracidade e o teste de perdas melhoram conforme o aumento da quantidade de imagens utilizadas no treinamento.

Este tipo de observação é recorrente em todos os estudos realizados utilizando a CNN. Os melhores resultados deste estudo utilizaram mais de 4.610 imagens na fase de treinamento.

Esta análise pode ser observada pela perda do modelo nos casos com o treinamento efetuado com menos de 1.480 imagens. A quantidade de imagens mínima para a análise com menor perda não foi o foco desta avaliação.

Todos os treinamentos com no máximo a quantidade de 1.480 imagens tiveram perdas acima de 60% e acuracidade em treino de no máximo 65%. No caso dos treinamentos

com mais de 4.610 imagens a acuracidade menor foi de 92,2%, porém não foi garantia de ter um índice pequeno de perdas do modelo. No caso do treinamento da divisão 47 com índices financeiros, pode-se observar que a perda do modelo foi de 24,2%.

Entretanto, fica claro como a quantidade de imagens pode gerar melhores resultados. Nos casos da utilização das contas contábeis, os modelos treinados com quantidade de imagens superior às 4.610 apresentaram uma elevada capacidade de predição.

A quantidade de imagens pode ser considerada uma limitação do uso da CNN para identificação de objetos ou tendência em imagens; entretanto, não é um limitador para este estudo pela quantidade de dados trabalhados e onde busca-se mostrar o quanto a CNN é capaz de classificar corretamente as imagens, mesmo estas sendo consideradas abstratas aos olhos humanos.

5.2. Comparação entre a utilização dos índices financeiros e as contas contábeis

Para as divisões 47 e 86 e para o grupo 478 foram criadas imagens tanto para os índices financeiros quanto para as contas contábeis (análises vertical e horizontal). O objetivo foi identificar se poderia haver diferenças nos resultados entre elas.

Os índices financeiros utilizam os dados contábeis em sua composição, e muitos destes dados são utilizados em mais de um índice. A questão a ser verificada é se a possível correlação entre os índices geraria algum tipo de variação na acuracidade em comparação ao uso dos dados contábeis.

O cálculo das correlações dos índices financeiros assim como as análises vertical e horizontal dos dados contábeis estão no Apêndice III, onde pode-se observar uma baixa correlação entre a maiorias das variáveis nos três casos. Foram utilizados os dados da divisão 47 nos testes das correlações.

Uma explicação plausível para o fato foi encontrada com a análise dos dados contábeis. Como são empresas de pequeno e médio porte, a contabilidade, em geral, não costuma seguir os rigores encontrados nas empresas de grande porte e de capital aberto. Esta simplificação contábil parece acabar alterando o resultado dos índices financeiros. Muitos dos dados das contas contábeis importantes para os cálculos estão zerados ou com valores não correspondentes à realidade da empresa, dificultando a análise destas empresas através dos índices financeiros calculados.

A criação das imagens dos índices financeiros com estas distorções acaba dificultando a identificação dos padrões pela CNN. Entretanto, a criação das imagens com as análises horizontal e vertical acaba gerando um padrão normalmente repetido em todas as empresas destes portes.

Esta análise pode ser observada principalmente pela perda do modelo em fase de treinamento, onde as perdas dos modelos das imagens geradas com os índices financeiros foram maiores em comparação às imagens geradas pelas análises vertical e horizontal das contas contábeis.

O caso com maior distorção aconteceu no treinamento da divisão 47, onde as perdas do modelo de índices financeiros chegaram a 24,2%. A divisão 47 é justamente a com maior número de imagens e teve uma acuracidade de 92,2%. Esta acuracidade poderia ser considerada excelente, ainda mais quando observado o acerto do teste de 93,1%. Entretanto, esta perda do modelo aponta problemas no treinamento da CNN e, portanto, um modelo pouco confiável.

O treinamento da divisão 47 com as imagens geradas com os dados das análises vertical e horizontal das contas contábeis devolveu um resultado muito melhor. A acuracidade do modelo chegou a 97,5% e o acerto do teste à 97,3%. Porém, a perda do modelo foi de 8,5%, indicando uma capacidade de predição superior aos modelos

testados em outros estudos descritos na seção 2.3 envolvendo as redes neurais com dados quantitativos.

Como citado anteriormente, uma informação não confiável nos dados das contas contábeis pode afetar diretamente vários índices financeiros; entretanto, esta informação isolada não afeta de forma tão direta as outras contas contábeis, não prejudicando a classificação da CNN das imagens criadas pelos dados destas.

Sendo assim, foi feita a opção de gerar somente as imagens das análises vertical e horizontal das contas contábeis para os outros treinamentos.

5.3. Treinamentos e testes da mesma divisão

Foram realizados treinamentos com três divisões: 47, 82 e 86.

A quantidade de imagens foi decisiva para a qualidade do modelo.

As divisões 47 e 86 foram usadas para a verificação das imagens geradas tanto por índices financeiros como para as geradas pelas contas contábeis pelo fato de possuir mais imagens, ao todo 6.417 e 2.043 imagens respectivamente. O resultado obtido pelos modelos da divisão 86 poderiam ser considerados razoáveis, se fosse analisado somente o acerto do teste; porém, a acuracidade e a alta perda indica um modelo fraco.

No caso da divisão 47, a seção 5.2 explicou como o modelo criado com as imagens geradas com os dados das contas contábeis forneceu um modelo confiável e com alta acuracidade e acerto do teste.

No caso da divisão 82, foram utilizadas 1.244 imagens totais e não foi suficiente para gerar um modelo confiável.

5.4. Treinamentos e testes do mesmo grupo

Foram realizados treinamentos com três grupos de empresas: 472, 475 e 478.

Os resultados obtidos nos modelos dos três casos refletem a baixa quantidade de imagens para realizar o treinamento.

O grupo 478 foi usado para a verificação das imagens geradas tanto por índices financeiros como para as geradas pelas contas contábeis pelo fato de possuir mais imagens, ao todo 1.788 imagens. Mesmo assim, o resultado nos dois modelos não ultrapassou a acuracidade de 65,0% e ainda teve perdas de mais de 60,0%.

Nenhum dos modelos pode ser considerado válido para a realização de uma predição confiável.

5.5. Treinamentos e testes com a intersecção de duas divisões

Separadamente, as divisões 47 e 86 obtiveram resultados bem diferentes, onde a divisão 47 conseguiu um modelo de alta qualidade para as imagens das contas contábeis; porém, fraco para as imagens geradas pelos índices financeiros.

Já a divisão 86 obteve modelos fracos para os dois tipos de imagens geradas.

Entretanto, a união de dois grupos tão antagônicos, “comércio varejista” e “atividades de atenção à saúde humana”, poderia gerar um modelo melhor e confiável?

A contabilidade é uma técnica universal e independente do tipo da empresa, foi criada como uma forma de documentar os lançamentos financeiros de uma empresa para poderem ser analisados por qualquer pessoa conhecedora da técnica.

As imagens geradas possuem a identificação da divisão e grupo dos dados e a dúvida a ser esclarecida era se a CNN teria condição de ter um modelo de predição melhor para a divisão 86 caso esta fosse treinada em conjunto com os dados de outra divisão com maior quantidade de imagens.

Os resultados mostraram o aumento da capacidade preditiva tanto para o modelo com imagens de dados financeiros como para o modelo com imagens dos dados das contas contábeis.

No caso do modelo das imagens das contas contábeis a acuracidade chegou à 96,4% e o acerto do teste à 97,8% (o maior obtido em todos os treinamentos) com uma perda de teste de 10,4%. Sendo assim, pode ser considerado um bom modelo preditivo.

Para verificar se o modelo realmente ajudou a melhorar o acerto de novas imagens, foram utilizadas as mesmas 410 imagens empregadas no teste do modelo criado somente para a divisão 86 (onde foi obtido um resultado de 71,5% de acerto do teste) no treinamento da intersecção das duas divisões. O resultado do acerto nesta condição foi de 92,9%, podendo ser considerado um bom modelo preditivo e melhor em comparação ao modelo gerado apenas pela divisão 86.

5.6. Treinamentos em uma divisão e testes com divisão diferente

Como forma de analisar se os dados contábeis poderiam ser tratados de forma única, independente da divisão de treinamento, foi utilizado o treinamento da divisão 47 para aplicar as imagens utilizadas nos testes das divisões 82 e 86.

Os resultados obtidos mostram um acerto fraco. As imagens da divisão 82 obtiveram um acerto de 51,1% (no treinamento 11 obteve 70,7%) e as imagens da divisão 86 obteve um acerto de 50,2% (no treinamento 2 obteve 71,5%).

A perda da qualidade do acerto mostra como a CNN leva em consideração a informação da divisão e setor colocados na imagem. Observa-se na seção 5.5 que o mesmo grupo de imagens obteve 92,9% em um treinamento misto com a mesma divisão 47.

5.7. Treinamentos em uma divisão e testes com um grupo da mesma divisão

Completando o caso descrito no capítulo 4 em que foi utilizado o treinamento da divisão 47 e as imagens de teste do grupo 472, obteve-se um acerto do teste de 62,0%.

O treinamento em uma divisão e testes com um grupo da mesma divisão utilizando as mesmas 142 imagens alcançou um acerto de 96,5%. Esta melhora no resultado podia ser esperada, pois o grupo 472 está contido no treinamento da divisão 47.

CONCLUSÕES

O mercado financeiro sempre conviveu com o problema de inadimplência e falência das empresas de todos os portes. Esta situação acaba gerando distorções nos serviços prestados pelas instituições financeiras, onde a incerteza sobre a real condição das empresas acaba prejudicando principalmente as empresas de pequeno e médio porte.

Os trabalhos realizados com o objetivo de predizer as inadimplências e falências sempre foram muito focados em empresas de capital aberto, normalmente empresas de grande porte com dados disponíveis devido às normas do mercado acionário.

Outro fator identificado como limitante é a utilização dos dados dos balanços anuais destas empresas. A utilização de balancetes trimestrais ajudou vários estudos sobre o tema, mas ainda assim são dados de corporações de porte significativo.

As pequenas e médias empresas sempre foram uma "área cinzenta", onde a pouca informação disponível ao sistema financeiro acaba prejudicando a criação de produtos com um custo mais acessível para as empresas destes portes. As pequenas e médias empresas dificilmente conseguem capital de forma fácil; e quando conseguem, acabam tendo o seu risco analisado de forma rasa, resultando na fixação de taxas elevadas para a concessão de empréstimos e financiamentos, quando comparadas às taxas praticadas com empresas abertas.

Os estudos antecessores conseguiram bons avanços utilizando os métodos de aprendizado de máquina, onde os dados com as variações dos índices financeiros sempre foram utilizados como dado principal. Estes estudos normalmente criam uma visão com variações de dados em um determinado período, não tratando a evolução da informação. Mesmo os estudos com dados trimestrais foram feitos desta maneira.

Este trabalho foi realizado com o intuito de criar uma forma de analisar as empresas de pequeno e médio porte através de sua contabilidade utilizando redes

neurais e considerando o uso dos dados dentro de um conceito de análise temporal. A rede neural convolucional foi escolhida para analisar as imagens com os dados contábeis por ser um método de treinamento de máquina com capacidade de tratamento de séries temporais.

Este tipo de rede neural é utilizado em várias aplicações para a identificação de objetos através das imagens, tendo a habilidade de aprender automaticamente e de forma hierárquica as características dos dados. As CNN iniciam seu aprendizado com pequenos detalhes e gradualmente reconhecem padrões mais amplos e abstratos.

O envio de uma imagem com os dados da divisão e grupo de atuação da empresa, assim como a identificação do país, inflação do mês e dos últimos doze meses forneceu subsídios para a CNN fazer uma análise mais ampla dos dados. Outro ponto é a identificação das variáveis. Elas são colocadas em linhas, e a evolução de seus conteúdos ficam explícitos, facilitando a identificação dos padrões. A utilização da divisão e grupo da empresa pela CNN fica explicito pelo resultado obtido nas seções 3.5.3 e 3.5.15. Na primeira o treinamento somente utilizou uma divisão com poucas imagens obtendo um resultado com boa acurácia; entretanto, na segunda seção observa-se como a inclusão do processamento de uma divisão com mais imagens melhora a análise das imagens do setor com menos imagens, mesmo sendo de divisões completamente diferentes.

A construção da imagem é um avanço importante para futuros estudos, ficando visível como o uso das variações das variáveis brutas pela CNN acabam sendo melhores em comparação com o uso das variações de índices criados a partir dos dados brutos. A CNN mostra ser capaz de analisar com mais profundidade os dados não tratados e deles tirar as suas conclusões. Com isso, o estudo apresentou uma nova forma de “tabular” dados temporais para serem analisados por uma CNN. A construção

da imagem consegue captar as nuances contidas nos dados, abrindo uma nova forma de estruturar as informações para outros estudos.

Entretanto, o resultado mostrou algumas limitações, como o caso da necessidade de uma considerável quantidade de imagens para a realização de um treinamento válido; porém, os resultados obtidos mostraram como a CNN consegue classificar as empresas de forma correta e com uma acuracidade superior a 90%.

Desta forma, o estudo apresenta ao mercado financeiro uma metodologia viável para analisar o risco de inadimplência e falência das empresas, garantindo para as pequenas e médias empresas uma análise justa sobre seu real risco financeiro. Com um treinamento realizado com empresas de todos os setores, as novas empresas podem ser analisadas com uma comparação direta com a base treinada.

Um sub produto deste estudo mostra uma forma de normalizar os planos de contas contábeis. A criação de um plano de contas padrão e a utilização do Codificador de Frase Universal (seção 2.4) para buscar a proximidade entre os planos de contas é uma estratégia passível de ser utilizada para outras finalidades. No caso deste trabalho o método ainda fornece a capacidade do sistema para processar a contabilidade em outras línguas, além do Português, por comparar as proximidades das frases de forma multilíngue. Sendo assim, a metodologia de criar imagens com os dados pode ser considerada universal, com a única observação de buscar as divisões e grupos similares das empresas de outros países. Fica a oportunidade de novos estudos envolvendo a contabilidade de empresas de mais de um país, buscando entender se é possível fazer o modelo trabalhar com um treinamento capaz de analisar empresas de outros países em conjunto.

REFERÊNCIAS BIBLIOGRÁFICAS

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, 6(2), 38.

Akerlof, G. A. (2003). Writing the "The market for lemons": A personal and interpretive essay. *Available at nobelprize.org.*

Aktan, S. (2011). Application of machine learning algorithms for business failure prediction. *Investment Management and Financial Innovations*, (8, Iss. 2), 52-65.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4), 589-609.

Altman, E. I. (1971). Railroad bankruptcy propensity. *The Journal of Finance*, 26(2), 333-345.

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review*, 40, 100378.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6), 26-38.

Banik, S., Loeffler, T. D., Batra, R., Singh, H., Cherukara, M. J., & Sankaranarayanan, S. K. (2021). Learning with delayed rewards—a case study on inverse defect design in 2D materials. *ACS Applied Materials & Interfaces*, 13(30), 36455-36464.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications,* 83, 405-417.

Bilstrup, K. E. K., Kaspersen, M. H., & Petersen, M. G. (2020, July). Staging reflections on ethical dilemmas in machine learning: A card-based design workshop for high school students. In *Proceedings of the 2020 ACM designing interactive systems conference* (pp. 1211-1222).

Borovykh, A., Bohte, S., & Oosterlee, C. W. (2017). Conditional time series forecasting with convolutional neural networks. *arXiv preprint arXiv:1703.04691*.

Boyd, J. H., & De Nicolo, G. (2005). The theory of bank risk taking and competition revisited. *The Journal of finance*, 60(3), 1329-1343.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.

Burges, C., Ragno, R., & Le, Q. (2006). Learning to rank with nonsmooth cost functions. *Advances in neural information processing systems*, 19.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57, 203-216.

Caouette, J. B., Altman, E. I., Narayanan, P., & Nimmo, R. (2011). *Managing credit risk: The great challenge for global financial markets*. John Wiley & Sons.

Carling, K., Jacobson, T., Lindé, J., & Roszbach, K. (2007). Corporate credit risk modeling and the macroeconomy. *Journal of banking & finance*, 31(3), 845-868.

Carter, S. M., Rogers, W., Win, K. T., Frazer, H., Richards, B., & Houssami, N. (2020). The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. *The Breast*, 49, 25-32.

Carton, R. B. (2004). Measuring organizational performance: An exploratory study.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Kurzweil, R. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175.*

Chen, W., & Cummings, M. (2024, May). Subjectivity in Unsupervised Machine Learning Model Selection. In *Proceedings of the AAAI Symposium Series* (Vol. 3, No. 1, pp. 22-29).

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Chopra, D., & Tanzi, R. Super Brain: Unleashing the Explosive Power of Your Mind to Maximize Health, Happiness, and Spiritual Well-Being, 2013.

Choudhary, R., & Gianey, H. K. (2017, December). Comprehensive review on supervised machine learning algorithms. In 2017 *International Conference on Machine Learning and Data Science* (MLDS) (pp. 37-43). IEEE.

Cintia Ganesha Putri, D., Leu, J. S., & Seda, P. (2020). Design of an unsupervised machine learning-based movie recommender system. *Symmetry,* 12(2), 185.

Claessens, S. (2019). Fragmentation in global financial markets: good or bad for financial stability?.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics, 35(5-6),* 352-359.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation, 10(7),* 1895-1923.

Dourado, D. D. A., & Aith, F. M. A. (2022). The regulation of artificial intelligence for health in Brazil begins with the General Personal Data Protection Law. *Revista de Saúde Pública*, 56.

Du, Y., Tan, Z., Zhang, X., Yao, Y., Yu, H., & Wang, C. (2021). Unsupervised domain adaptation with unified joint distribution alignment. In *Database Systems for*

Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part II 26 (pp. 449-464). Springer International Publishing.

ELIZABETSKY, R. Um modelo matemático para decisões de crédito no banco comercial. 1976. 190 f. Dissertação (Mestrado em Engenharia da Produção) – Escola Politécnica da Universidade de São Paulo. São Paulo, 1976.

Ferreira, R. J. (2009). Contabilidade básica. *Teoria e mais de, 1*.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, 535-545.

Fridson, M. S., & Alvarez, F. (2022). *Financial statement analysis: a practitioner's guide*. John Wiley & Sons.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Frydman, H., Altman, E. I., & Kao, D. L. (1985). Introducing recursive partitioning for financial classification: the case of financial distress. The journal of finance, 40(1), 269-291.

Gans, D. N., Piland, N. F., & Honoré, P. A. (2007). Developing a chart of accounts: historical perspective of the Medical Group Management Association. *Journal of Public Health Management and Practice, 13(2)*, 130-132.

Grira, N., Crucianu, M., & Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: a brief survey. *A review of machine learning techniques for processing multimedia content, 1*(2004), 9-16.

Guan, X., & Burton, H. (2022, December). Bias-variance tradeoff in machine learning: Theoretical formulation and implications to structural engineering applications. In *Structures* (Vol. 46, pp. 17-30). Elsevier.

Haji Nkuhi, A. (2015). *The role of financial statements in investment decision Making A case of Tanga port authority* (Doctoral dissertation, Mzumbe University).

Hebiri, M., & Lederer, J. (2012). How correlations influence lasso prediction. *IEEE Transactions on Information Theory,* 59(3), 1846-1854.

Hellmann, T. F., Murdock, K. C., & Stiglitz, J. E. (2000). Liberalization, moral hazard in banking, and prudential regulation: Are capital requirements enough? *American economic review*, 91(1), 147-165.

Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert systems with applications*, 117, 287-299.

Hung, C., & Chen, J. H. (2009). A selective ensemble based on expected probabilities for bankruptcy prediction. *Expert systems with applications*, 36(3), 5297-5303.

Hung, J. L., He, W., & Shen, J. (2020). Big data analytics for supply chain relationship in banking. *Industrial Marketing Management*, 86, 144-153.

Hutter, F., Lücke, J., & Schmidt-Thieme, L. (2015). Beyond manual tuning of hyperparameters. *KI-Künstliche Intelligenz*, 29, 329-337.

Jabbar, H., & Khan, R. Z. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices,* 70(10.3850), 978-981.

Jakkula, V. (2006). Tutorial on support vector machine (svm). *School of EECS, Washington State University*, 37(2.5), 3.

Jin, X., Yu, X., Wang, X., Bai, Y., Su, T., & Kong, J. (2020). Prediction for Time Series with CNN and LSTM. In *Proceedings of the 11th international conference on modelling, identification and control (ICMIC2019)* (pp. 631-641). Springer Singapore.

Joshi, D. J., Kale, I., Gandewar, S., Korate, O., Patwari, D., & Patil, S. (2021). Reinforcement learning: a survey. In *Machine Learning and Information Processing: Proceedings of ICMLIP 2020* (pp. 297-308). Springer Singapore.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4, 237-285.

Kaiser, T., Lusardi, A., Menkhoff, L., & Urban, C. (2022). Financial education affects financial knowledge and downstream behaviors. *Journal of Financial Economics*, 145(2), 255-272.

KANITZ, S. C. Como prever falências de empresas. São Paulo: Mcgraw- Hill, 1978.

KANITZ, S. C. Indicadores contábeis financeiros – previsão de insolvência: a experiência da pequena e média empresa brasileira. Tese (Livre- Docência). – Departamento de Contabilidade da FEA/USP, São Paulo, 1976.

Kassai, S., & Onusic, L. M. (2004). Modelos de Previsão de Insolvência utilizando a Análise por Envoltório de Dados: aplicação a empresas brasileira. In *Congresso USP de Controladoria e Contabilidade* (Vol. 4).

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research, 32*(11), 1238-1274.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering,* 160(1), 3-24

Krugman, P. (1999). Balance sheets, the transfer problem, and financial crises. *International tax and public finance*, 6, 459-472.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436-444.

Li, K. (1999). Bayesian analysis of duration models: an application to Chapter 11 bankruptcy. *Economics Letters*, 63(3), 305-312.

Li, Y., & Wang, Y. (2017). Machine learning methods of bankruptcy prediction using accounting ratios. *Open Journal of Business and Management*, 6(1), 1-20.

Liu, Y., Li, L., Hao, Z., Zhang, X., Li, J., & Zhang, T. (2017). A deep reinforcement learning approach to personalized medicine. In *Proceedings of the 2017 SIAM International Conference on Data Mining* (pp. 153-161). SIAM.

Madan, R., & Mangipudi, P. S. (2018, August). Predicting computer network traffic: a time series forecasting approach using DWT, ARIMA and RNN. In 2018 *Eleventh International Conference on Contemporary Computing (IC3)* (pp. 1-5). IEEE.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).* [Internet], 9(1), 381-386.

Malíková, O., & Brabec, Z. (2012). The influence of a different accounting system on informative value of selected financial ratios. *Technological and economic development of economy,* 18(1), 149-163.

Matias, A. B. (2007). Finanças corporativas de curto prazo: a gestão do valor do capital de giro. *São Paulo: Atlas*, 1, 285.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and information technology, 6, 175-183.

McCormick, B. H. (1988). Visualization in scientific computing. *Acm Sigbio Newsletter*, 10(1), 15-21.

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society,* 100(11), 2175-2199.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781.

Min, S. H., Lee, J., & Han, I. (2006). Hybrid genetic algorithms and support vector machines for bankruptcy prediction. *Expert systems with applications*, 31(3), 652-660.

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data mining and knowledge discovery*, 2, 345-389.

Myers, J. H. (1963). Predicting credit risk with a numerical scoring system. *Journal of Applied Psychology*, 47(5), 348.

Neto, A. A. (2003). Finanças corporativas e valor. Atlas.

Nevmyvaka, Y., Feng, Y., & Kearns, M. (2006). Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd international conference on Machine learning* (pp. 673-680). ACM.

Nguyen, Q. P., & Vo, D. H. (2022). Artificial intelligence and unemployment: An international evidence. *Structural Change and Economic Dynamics*, 63, 40-55.

Osoba, O. A., & Welser, W. (2017). *The risks of artificial intelligence to security and the future of work.* Santa Monica, CA: RAND.

Ozili, P. K. (2018). Impact of digital finance on financial inclusion and stability. *Borsa Istanbul Review,* 18(4), 329-340.

Ozili, P. K. (2020). Theories of financial inclusion. In *Uncertainty and challenges in contemporary economic behaviour* (pp. 89-115). Emerald Publishing Limited.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Piasecki, J., Waligora, M., & Dranseika, V. (2018). Google search as an additional source in systematic reviews. *Science and engineering ethics*, 24, 809-810.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons.

Raiter, O. (2021). Segmentation of bank consumers for artificial intelligence marketing. *International Journal of Contemporary Financial Issues*, 1(1), 39-54.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

Sadaf, K., & Sultana, J. (2020). Intrusion detection based on autoencoder and isolation forest in fog computing. *IEEE Access, 8*, 167059-167068.

Saligkaras, D., & Papageorgiou, V. E. (2023, August). Seeking the truth beyond the data. An unsupervised machine learning approach. In *AIP Conference Proceedings* (Vol. 2812, No. 1). AIP Publishing

Severiano, M. C., da Silva Dantas, C. E., ALMEIDA, C. R. D. C., Valdevino, R. Q. S., de Oliveira, A. M., & de Paula, B. S. (2021). Avaliação de desempenho nos bancos digitais: uma abordagem na perspectiva gerencial. In *Anais do Congresso Brasileiro de Custos-ABC*.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The journal of business*, 74(1), 101-124.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529* (7587), 484-489.

Singh, M. (2012). Marketing mix of 4P's for competitive advantage. IOSR *Journal of Business and Management*, 3(6), 40-45.

Suk, H. I., Wee, C. Y., Lee, S. W., & Shen, D. (2016). State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *NeuroImage*, 129, 292-307.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tabak, B. M., Fazio, D. M., & Cajueiro, D. O. (2012). The relationship between banking market competition and risk-taking: Do size and capitalization matter? *Journal of Banking & Finance*, 36(12), 3366-3381.

Thakur, A., & Konde, A. (2021). Fundamentals of neural networks. *International Journal for Research in Applied Science and Engineering Technology,* 9(VIII), 407-426.

Trimble, M. (2017). The historical and current status of IFRS adoption around the world. *University of Mannheim*, 1-56.

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. L. A., Elkhatib, Y., ... & Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE access*, 7, 65579-65615.

Van der Pol, E., & Oliehoek, F. A. (2016). Coordinated deep reinforcement learners for traffic light control. In *Proceedings of the NIPS Workshop on Learning, Inference and Control of Multi-Agent Systems*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Verona, M. M. (2004). Marketing bancário. *Revista do Curso de Administração da Faculdade da Serra Gaúcha. Caxias do Sul,* 4(6), 67-79.

Wang, J., Yang, Y., Wang, T., Sherratt, R. S., & Zhang, J. (2020). Big data service architecture: a survey. *Journal of Internet Technology*, 21(2), 393-405.

Ware, C. (2019). *Information visualization: perception for design*. Morgan Kaufmann.

Wewege, L., Lee, J., & Thomsett, M. C. (2020). Disruptions and digital banking trends. *Journal of Applied Finance and Banking,* 10(6), 15-56.

Whittington, G. (1980). Some basic properties of accounting ratios. *Journal of business finance and accounting,* 7(2), 219-232.

Wilton, J., Koay, A., Ko, R., Xu, M., & Ye, N. (2022). Positive-unlabeled learning using random forests via recursive greedy risk minimization. *Advances in Neural Information Processing Systems, 35,* 24060-24071.

Xie, T., & Zhang, J. (2022). Financial Default Risk Prediction Algorithm Based on Neural Network under the Background of Big Data. *Mobile Information Systems*, 2022.

Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., ... & Wang, C. (2018). Machine learning and deep learning methods for cybersecurity. *Ieee access*, 6, 35365-35381.

Yu, R., Wang, Y., Zou, Z., & Wang, L. (2020). Convolutional neural networks with refined loss functions for the real-time crash risk analysis. *Transportation research part C: emerging technologies,* 119, 102740.

Zhang, J. (2007). *Visualization for information retrieval* (Vol. 23). Springer Science & Business Media.

Zhao, B., Lu, H., Chen, S., Liu, J., & Wu, D. (2017). Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics*, 28(1), 162-169.

Zhao, X., Fu, C., King, I., & Lyu, M. R. (2019). Deep reinforcement learning for page-wise recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 95-103). ACM.

Zhou, Z. H. (2018). A brief introduction to weakly supervised learning. *National science review,* 5(1), 44-53.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology,* 67(2), 301-320.

Internet links

https://keras.io/

https://sebrae.com.br/

https://www.gov.br/receitafederal/

https://www.bcb.gov.br/estatisticas

https://keras.io/examples/vision/mnist_convnet/

Apêndice I – Código gerador das imagens e ambiente de processamento

```
from PIL import Image
import pandas as pd
import random
import csv
import os

def criar_diretorio(diretorio_destino):
    if not os.path.exists(diretorio_destino):
        os.makedirs(diretorio_destino)

def cria_imagem(cor_y, row_y, varn_y, conta_img, diretorio):
    criar_diretorio(diretorio)

    nome_img = str(conta_img) + ".png"
    caminho_completo = os.path.join(diretorio, nome_img)
    FILENAME = (caminho_completo, "PNG")

    col_y = varn_y
    fim_imp = len(cor_y)
    linha = fim_imp - 576
    coluna = varn_y
    pillow_obj = Image.new("RGB", (24, 24))
    pixel_set = pillow_obj.load()

    for linpri in range(linha, fim_imp):
        cada = cor_y[linpri]
        lin_imp = cada[0] - row_y
        color = (cada[2], cada[3], cada[4])
        pixel_set[cada[1], lin_imp] = color

    pillow_obj.save(*FILENAME)
    return(caminho_completo)

rel_indicators_customers_cab_div_rat = spark.sql(f"""
select * from rel_indicators_customers_cab_div_rat
where code_group = 47
  and (cab_21000 + cab_22000 + cab_23000) <> 0
order by identifier, period
"""
)
df_rel_indicators_customers_cab_div_rat = rel_indicators_customers_cab_div_rat.toPandas()
ax = df_rel_indicators_customers_cab_div_rat.count()

ipca_rgb = spark.sql(f""" select * from ipca_rgb """)
df_ipca_rgb = ipca_rgb.toPandas()

estado_numero =
{'RO':11,'AC':12,'AM':13,'RR':14,'PA':15,'AP':16,'TO':17,'MA':21,'PI':22,'CE':23,'RN':24,'PB':25,'PE':26,
'AL':27,'SE':28,'BA':29,'MG':31,'ES':32,'RJ':33,'SP':35,'PR':41,'SC':42,'RS':43,'MS':50,'MT':51,'GO':52,'
DF':53}
```

```
imagem_tamanho = 0
imagem_periodo = 12
imagem_contagem = 0
imagem_diretorio = "../archives"

imagem_data = []
imagem_treino = []
imagem_teste = []

cor_t = []
row = 0
col = 36
identifier_image = 0
variavel = 6
mes_image = df_rel_indicators_customers_cab_div_rat['period'][0].month
ano_image = df_rel_indicators_customers_cab_div_rat['period'][0].year

for registro in range(ax[0]):
    if df_rel_indicators_customers_cab_div_rat['cab_10000'][registro] == 0: continue

    identifier_loop = df_rel_indicators_customers_cab_div_rat['identifier'][registro]
    ind_id_log = df_rel_indicators_customers_cab_div_rat['indicator_id'][registro]
    identifier_log = df_rel_indicators_customers_cab_div_rat['identifier'][registro]
    ipca_ano = df_rel_indicators_customers_cab_div_rat['period'][registro].year
    ipca_mes = df_rel_indicators_customers_cab_div_rat['period'][registro].month

    if identifier_image == 0 or identifier_image != identifier_loop:
        # print(identifier_image, identifier_loop)
        identifier_image = identifier_loop
        row = 0
        cor_t.clear()
        cor_t = []
    if df_rel_indicators_customers_cab_div_rat['code_division'][registro]:
        rgb1 = int(df_rel_indicators_customers_cab_div_rat['code_division'][registro])
    else: rgb1 = 0

    if df_rel_indicators_customers_cab_div_rat['code_group'][registro]:
        rgb2 = (int(df_rel_indicators_customers_cab_div_rat['code_group'][registro]) - rgb1 * 10 ) * 10
    else:
        rgb2 = 0

    est_uf = df_rel_indicators_customers_cab_div_rat['uf'][registro]
    rgb3 = estado_numero[est_uf]

    aux = [row, 0, rgb1, rgb2, rgb3]
    cor_t.append(aux)

    aux = [row+1, 0, rgb1, rgb2, rgb3]
    cor_t.append(aux)

    ipca_rgb = df_ipca_rgb[(df_ipca_rgb["ano"] == ipca_ano) & (df_ipca_rgb["mes"] ==
ipca_mes)].reset_index()
```

```
    rgb1 = (df_rel_indicators_customers_cab_div_rat['period'][registro].year - 1970) * 2
    rgb2 = df_rel_indicators_customers_cab_div_rat['period'][registro].month * 10
    rgb3 = 55

    if imagem_periodo != 12:
        if imagem_periodo == 13 and rgb2 < 120:
            continue
        elif (rgb2 % 3) != 0:
            continue

    aux = [row, 1, rgb1, rgb2, rgb3]
    cor_t.append(aux)

    aux = [row+1, 1, rgb1, rgb2, rgb3]
    cor_t.append(aux)

    if ipca_rgb["rgb_ipca"].shape[0] == 0:
        rgb1 = 100
        rgb2 = 125
        rgb3 = 125
    else:
        rgb1 = ipca_rgb["rgb_ipca"][0]
        rgb2 = ipca_rgb["rgb_mes"][0]
        rgb3 = ipca_rgb["rgb_acum"][0]
    aux = [row, 2, rgb1, rgb2, rgb3]
    cor_t.append(aux)

    aux = [row+1, 2, rgb1, rgb2, rgb3]
    cor_t.append(aux)

    varn = 3
    cab_rgb = 100
    for col_num in ['cab_11100', 'cab_11200', 'cab_11300', 'cab_11500', 'cab_13200', 'cab_13300',
'cab_13400', 'cab_21100', 'cab_21200', 'cab_22000', 'cab_22300', 'cab_23000', 'cab_31100', 'cab_32100',
'cab_32200', 'cab_32300', 'cab_32400', 'cab_32500', 'cab_32700', 'cab_32800', 'cab_32900']:

        cab_rgb = cab_rgb + 5

        div_num = col_num.replace("cab", "div")
        rat_num = col_num.replace("cab", "rat")

        for ic_num in [div_num, rat_num]:

            fresult = 0
            result = df_rel_indicators_customers_cab_div_rat[ic_num][registro] * 100

            if result < 0: rgb1 = 50
            else: rgb1 = 250
```

```
        if pd.isna(result):
            result = 0
            fresult = 0
        else:
            fresult = result

        if pd.isna(fresult):
            fresult = 0

        fresult = int(fresult)
        fresult = abs(fresult)

        if fresult > 255: fresult = 255
        if fresult < 0: fresult = 0

        rgb2 = 0
        rgb3 = 0

        if rgb1 == 50:
            rgb3 = int(fresult)
        else:
            rgb2 = int(fresult)

        if fresult == 0: rgb1 = 150

        if ic_num[0:3] == "div":
            aux = [row, varn, rgb1, rgb2, rgb3]
            cor_t.append(aux)
        else:
            aux = [row+1, varn, rgb1, rgb2, rgb3]
            cor_t.append(aux)
            varn = varn + 1

    row = row + 2

    if row >= 24:
        prx_reg = registro +1
        if (prx_reg) == ax[0]:
            imprime = 0
        else:
            imprime = 1
            if identifier_log != df_rel_indicators_customers_cab_div_rat['identifier'][prx_reg]:
                imprime = 0

        if imprime == 0:
            train_test = random.randint(0, 9)
            if train_test < 2:
                imagem_diretorio = ../archives/Images/test/"
            else:
                imagem_diretorio = ../archives/Images/train/"
```

```
            imagem_criada = cria_imagem(cor_t, row, varn, imagem_contagem, imagem_diretorio)
            print(imagem_criada, row, varn, imagem_contagem)

            ind_id_file = 0
            if ind_id_log > 1:
                ind_id_file = 1

            aux = [imagem_contagem, ind_id_file, identifier_log, ipca_ano, ipca_mes]
            imagem_data.append(aux)

            if train_test < 2:
                aux = [imagem_criada]
                imagem_teste.append(aux)
            else:
                aux = [imagem_criada, ind_id_file]
                imagem_treino.append(aux)

            imagem_contagem = imagem_contagem + 1

treino_field = ['img_code', 'target']
teste_field = ['img_code']

with open('../archives/Images/test_data.csv', 'w') as f:
    csv_writer = csv.writer(f)
    csv_writer.writerow(teste_field)
    csv_writer.writerows(imagem_teste)

with open('../archives/Images/train_data.csv', 'w') as f:
    csv_writer = csv.writer(f)
    csv_writer.writerow(treino_field)
    csv_writer.writerows(imagem_treino)

with open('../archives/lista_data.csv', 'w') as f:
    csv_writer = csv.writer(f)
    csv_writer.writerow(treino_field)
    csv_writer.writerows(imagem_data)
```

Apêndice II – Dados consolidados dos testes realizados

Treinamento	Tipo Imagem	Treino			Teste		Perda	Acuracidade	Acerto Teste
		Divisão / Grupo	Qtd Treino	Qtd Validação	Divisão / Grupo	Quantidade			
1	Índices Financeiro	86	1.475	163	86	405	0,77212924	0,58895707	70,6%
2	Contas Contábeis	86	1.470	163	86	410	0,65751289	0,65020676	71,5%
3	Índices Financeiro	47	4.618	513	47	1.286	0,24235767	0,92202729	93,1%
4	**Contas Contábeis**	**47**	**4.661**	**517**	**47**	**1.239**	**0,08535116**	**0,97485495**	**97,3%**
5	Índices Financeiro	47 e 86	6.075	675	47 e 86	1.710	0,10861805	0,95703703	96,5%
6	**Contas Contábeis**	**47 e 86**	**6.057**	**673**	**47 e 86**	**1.730**	**0,10426246**	**0,96433878**	**97,8%**
7	Índices Financeiro	478	1.260	140	478	388	0,62647986	0,60714287	66,2%
8	Contas Contábeis	478	1.291	143	478	354	0,60276318	0,64335662	75,7%
9	Contas Contábeis	475	726	80	475	215	0,66294229	0,61230001	57,7%
10	Contas Contábeis	472	496	55	472	142	0,66707510	0,56363636	62,0%
11	Contas Contábeis	82	973	108	82	270	0,71101683	0,50000000	70,7%
12	Contas Contábeis	47	4.661	517	82	270	0,08535116	0,97485495	51,1%
13	Contas Contábeis	47	4.661	517	86	410	0,08535116	0,97485495	50,2%
14	**Contas Contábeis**	**47 e 86**	**6.057**	**673**	**86**	**410**	**0,10426246**	**0,96433878**	**92,9%**
15	**Contas Contábeis**	**47**	**4.661**	**517**	**472**	**142**	**0,08535116**	**0,97485495**	**96,5%**

Apêndice III – Correlações

Correlação dos Índices Financeiros

Índices	AP	Seca II	Seca III	Liq. Corr.	Imob Rec.	Imob Pat Liq	comp_End	CCL	CCP	RLP	Liq Geral	EBITDA	Custo Oper	Custo Financeiro	Margem Liq	Margem Contr	Carga Trib	Margem Bruta	Despesas	Crescimento	Equilíbrio
AP	1,000000	-0,000963	-0,000853	-0,000964	-0,002319	-0,002302	0,010263	0,017812	0,017842	-0,000671	-0,000765	0,000708	0,000621	0,000845	-0,004222	-0,004944	0,000279	-0,005015	-0,010432	-0,000360	-0,001641
Seca II	-0,000963	1,000000	0,990371	0,999990	-0,000169	-0,000170	-0,000879	-0,000180	-0,000177	-0,000166	0,138369	-0,000187	0,000174	0,000173	-0,000210	-0,000204	0,000154	-0,000199	0,000242	-0,000073	0,000337
Seca III	-0,000853	0,990371	1,000000	0,990932	-0,000149	-0,000150	-0,000909	-0,000159	-0,000156	-0,000147	0,000055	-0,000165	0,000154	0,000152	-0,000185	-0,000180	0,000136	-0,000176	0,000214	-0,000064	0,000298
Liq. Corr.	-0,000964	0,999990	0,990932	1,000000	-0,000169	-0,000169	-0,000881	-0,000180	-0,000177	-0,000166	0,134273	-0,000186	0,000174	0,000172	-0,000209	-0,000203	0,000154	-0,000199	0,000242	-0,000073	0,000335
Imob Rec.	-0,002319	-0,000169	-0,000149	-0,000169	1,000000	0,999736	0,000151	-0,000160	-0,000157	-0,000150	-0,000138	-0,000166	0,000155	0,000153	-0,000251	-0,000242	0,000217	-0,000238	0,000215	-0,000065	-0,000150
Imob Pat Liq	-0,002302	-0,000170	-0,000150	-0,000169	0,999736	1,000000	0,000148	-0,000161	-0,000158	-0,000148	-0,000139	-0,000166	0,000155	0,000154	-0,000252	-0,000243	0,000216	-0,000238	0,000216	-0,000065	-0,000098
comp_End	0,010263	-0,000879	-0,000969	-0,000881	0,000151	0,000148	1,000000	0,000064	0,000082	0,020300	-0,000623	-0,000906	0,000917	0,000919	0,000178	0,000209	-0,000201	0,000220	0,000654	-0,010438	-0,004414
CCL	0,017812	-0,000180	-0,000159	-0,000180	-0,000160	-0,000161	0,000064	1,000000	0,999795	-0,000174	-0,000147	0,132171	-0,058213	-0,045699	-0,000198	-0,000193	0,000145	-0,000188	-0,656217	-0,000069	0,001239
CCP	0,017842	-0,000177	-0,000156	-0,000177	-0,000157	-0,000158	0,000082	0,999795	1,000000	-0,000168	-0,000145	0,112068	-0,037982	-0,025458	-0,000195	-0,000189	0,000143	-0,000185	-0,640800	-0,000068	0,001258
RLP	-0,000671	-0,000166	-0,000147	-0,000166	-0,000150	-0,000148	0,020300	-0,000174	-0,000168	1,000000	-0,000136	-0,000162	0,000151	0,000150	-0,000319	-0,000343	-0,000024	-0,000339	0,000211	-0,000472	0,000311
Liq Geral	-0,000765	0,138369	0,000055	0,134273	-0,000138	-0,000139	-0,000623	-0,000147	-0,000145	0,000000	1,000000	-0,000153	0,000143	0,000141	-0,000171	-0,000167	0,000127	-0,000163	0,000198	-0,000060	0,000272
EBITDA	0,000708	-0,000187	-0,000165	-0,000186	-0,000166	-0,000166	-0,000906	0,132171	0,112068	-0,000162	-0,000153	1,000000	-0,997240	-0,996231	-0,000205	-0,000199	0,000150	-0,000195	-0,834685	-0,000071	-0,000770
Custo Oper	0,000621	0,000174	0,000154	0,000174	0,000155	0,000155	0,000917	-0,058213	-0,037982	0,000151	0,000143	-0,997240	1,000000	0,999220	0,000191	0,000186	-0,000141	0,000181	0,791491	0,000066	0,000868
Custo Financeiro	0,000845	0,000173	0,000152	0,000172	0,000153	0,000154	0,000919	-0,045699	-0,025458	0,000150	0,000141	-0,996231	0,999920	1,000000	0,000189	0,000184	-0,000139	0,000180	0,783772	0,000066	0,000884
Margem Liq	-0,004222	-0,000210	-0,000185	-0,000209	-0,000251	-0,000252	0,000178	-0,000198	-0,000195	-0,000319	-0,000171	-0,000205	0,000191	0,000189	1,000000	0,995784	-0,050194	0,995164	0,000266	-0,000080	0,000344
Margem Contr	-0,004944	-0,000204	-0,000180	-0,000203	-0,000242	-0,000243	0,000209	-0,000193	-0,000189	-0,000343	-0,000167	-0,000199	0,000186	0,000184	0,995784	1,000000	-0,043754	0,999400	0,000259	-0,000078	0,000350
Carga Trib	0,000279	0,000154	0,000136	0,000154	0,000217	0,000216	-0,000201	0,000145	0,000143	-0,000024	0,000127	0,000150	-0,000141	-0,000139	-0,050194	-0,043754	1,000000	-0,044803	-0,000195	0,000059	-0,000260
Margem Bruta	-0,005015	-0,000199	-0,000176	-0,000199	-0,000238	-0,000238	-0,000220	-0,000188	-0,000185	-0,000339	-0,000163	-0,000195	0,000181	0,000180	0,995164	0,999400	-0,044803	1,000000	0,000253	-0,000076	0,000339
Despesas	-0,010432	0,000242	0,000214	0,000242	0,000215	0,000216	0,000654	-0,656217	-0,640800	0,000211	0,000198	-0,834685	0,791491	0,783772	0,000266	0,000259	-0,000195	0,000253	1,000000	0,000092	-0,000102
Crescimento	-0,000360	-0,000073	-0,000064	-0,000073	-0,000065	-0,000065	-0,010438	-0,000069	-0,000068	-0,000472	-0,000060	-0,000071	0,000066	0,000066	-0,000080	-0,000078	0,000059	-0,000076	0,000092	1,000000	-0,000232
Equilíbrio	-0,001641	0,000337	0,000298	0,000335	-0,000150	-0,000098	-0,004414	0,001239	0,001258	0,000311	0,000272	-0,000770	0,000868	0,000884	0,000344	0,000350	-0,000260	0,000339	-0,000102	-0,000232	1,000000

Correlação da Análise Vertical das Contas Contábeis

Contas	11100	11200	11300	11500	13200	13300	13400	21100	21200	22000	22300	23000	31100	32100	32200	32300	32400	32500	32700	32800	32900
11100	1,000000	-0,000145	-0,000019	0,000003	0,000014	-0,001895	0,000009	-0,000022	-0,000022	-0,000020	0,000007	0,000021	-0,000002	-0,001427	0,000003	0,000006	-0,000021	0,000033	-0,000002	-0,000021	-0,000020
11200	-0,000145	1,000000	-0,000035	0,423505	-0,041480	-0,006568	0,072442	-0,000002	-0,000570	0,000045	0,000146	-0,000019	-0,000094	-0,000034	-0,000059	0,022186	0,000195	0,000489	-0,000122	0,000019	-0,000092
11300	-0,000019	-0,000035	1,000000	0,000049	0,000012	0,000036	0,000018	-0,000011	-0,000011	-0,000010	0,000003	0,000010	-0,000001	0,000003	0,000001	0,000000	-0,000010	0,000007	-0,000001	-0,000010	-0,000010
11500	0,000003	0,423505	0,000049	1,000000	-0,000001	0,009316	0,137780	0,000002	0,000026	0,000028	-0,000001	-0,000033	-0,000283	-0,000007	-0,000001	0,000002	0,000028	0,000002	0,000003	0,000028	0,000027
13200	0,000014	-0,041480	0,000012	-0,000001	1,000000	0,001104	0,000084	0,000037	-0,000114	0,000064	0,000233	-0,000037	0,000011	0,000000	-0,000002	0,007072	0,000036	-0,000015	0,000012	0,000035	0,000039
13300	-0,001895	-0,006568	0,000036	0,009316	0,001104	1,000000	0,001126	-0,000063	-0,000088	-0,000062	0,000006	0,000062	-0,000007	0,000003	0,000002	0,000139	-0,000063	0,000037	0,000037	-0,000063	-0,000050
13400	0,000009	0,072442	0,000018	0,137780	0,000084	0,001126	1,000000	0,000032	0,000012	0,000034	-0,000013	-0,000035	-0,000058	0,000001	-0,000003	-0,000005	0,000036	0,009784	0,000003	0,000037	0,000046
21100	-0,000022	-0,000002	-0,000011	0,000002	0,000037	-0,000063	0,000032	1,000000	0,995405	0,989911	0,000015	-0,989344	-0,023560	-0,002285	-0,000507	0,000015	0,995808	0,000030	-0,004847	0,995804	0,896719
21200	-0,000022	-0,000570	-0,000011	0,000026	-0,000114	-0,000088	0,000012	0,995405	1,000000	0,993675	0,000014	-0,993107	-0,023649	-0,000487	0,000005	-0,000255	0,999594	0,000030	-0,000004	0,999591	0,900253
22000	-0,000020	0,000045	-0,000010	0,000028	0,000064	-0,000062	0,000034	0,989911	0,993675	1,000000	0,108676	-0,986676	0,033522	-0,034199	-0,054805	0,000017	0,994077	0,000029	-0,079923	0,994073	0,852927
22300	0,000007	0,000146	0,000003	-0,000001	0,000233	0,000006	-0,000013	0,000015	0,000014	0,108676	1,000000	0,008717	0,524942	-0,310268	-0,504421	0,000017	0,000015	0,000002	-0,735491	0,000015	-0,389799
23000	0,000021	-0,000019	0,000010	-0,000033	-0,000037	0,000062	-0,000035	-0,989344	-0,993107	-0,986676	0,008717	1,000000	0,079408	-0,059161	-0,056312	-0,005587	-0,993509	-0,000030	-0,006418	-0,993505	-0,909720
31100	-0,000002	-0,000094	-0,000001	-0,000283	0,000011	-0,000007	-0,000058	-0,023560	-0,023649	0,033522	0,524942	0,079408	1,000000	-0,965970	-0,996875	-0,000015	-0,023659	-0,000016	-0,386091	-0,023660	-0,388769
32100	-0,001427	-0,000034	0,000003	-0,000007	0,000000	0,000003	0,000001	-0,002285	-0,000487	-0,034199	-0,310268	-0,059161	-0,965970	1,000000	0,968986	0,000001	-0,000487	-0,000025	0,230633	-0,000487	0,301227
32200	0,000003	-0,000059	0,000001	-0,000001	-0,000002	0,000002	-0,000003	-0,000507	0,000005	-0,054805	-0,504421	-0,056312	-0,996875	0,968987	1,000000	-0,000002	0,000006	-0,000003	0,419705	0,000007	0,362567
32300	0,000006	0,022186	0,000000	0,000002	0,007072	0,000139	-0,000005	0,000015	-0,000255	0,000017	0,000017	-0,005587	-0,000015	0,000001	-0,000002	1,000000	0,000018	0,000160	0,000001	0,000015	0,000006
32400	-0,000021	0,000195	-0,000010	0,000028	0,000036	-0,000063	0,000036	0,995808	0,999594	0,994077	0,000015	-0,993509	-0,023659	-0,000487	0,000006	0,000018	1,000000	0,000034	-0,000004	0,999996	0,900618
32500	0,000033	0,000489	0,000007	0,000002	-0,000015	0,000037	0,009784	0,000030	0,000030	0,000029	0,000002	-0,000030	-0,000016	-0,000025	-0,000003	0,000160	0,000034	1,000000	0,000003	-0,001247	0,000028
32700	-0,000002	-0,000122	-0,000001	0,000003	0,000012	0,000037	0,000003	-0,004847	-0,000004	-0,079923	-0,735491	-0,006418	-0,386091	0,230633	0,419705	0,000001	-0,000004	0,000003	1,000000	-0,000004	0,298498
32800	-0,000021	0,000019	-0,000010	0,000028	0,000035	-0,000063	0,000037	0,995804	0,999591	0,994073	0,000015	-0,993505	-0,023660	-0,000487	0,000007	0,000015	0,999996	-0,001247	-0,000004	1,000000	0,900615
32900	-0,000020	-0,000092	-0,000010	0,000027	0,000039	-0,000050	0,000046	0,896719	0,900253	0,852927	-0,389799	-0,909720	-0,388769	0,301227	0,362567	0,000006	0,900618	0,000028	0,298498	0,900615	1,000000

Correlação da Análise Horizontal das Contas Contábeis

Contas	11100	11200	11300	11500	13200	13300	13400	21100	21200	22000	22300	23000	31100	32100	32200	32300	32400	32500	32700	32800	32900
11100	1,000000	-0,091687	-0,026987	-0,035451	-0,000042	-0,290005	0,000105	0,110485	-0,267997	-0,001885	0,313801	-0,000766	0,015259	-0,038246	0,005070	0,520125	0,003634	-0,000511	-0,000641	0,100712	-0,034354
11200	-0,091687	1,000000	0,010259	-0,000871	0,000141	-0,003716	-0,000009	0,022869	0,008896	-0,001624	0,345436	0,000411	0,001456	0,009521	0,000779	0,000071	0,000263	0,000041	0,007213	0,058114	-0,020586
11300	-0,026987	0,010259	1,000000	0,000002	0,000056	0,013564	0,029601	0,004421	0,000487	-0,030678	-0,063763	0,000031	-0,000162	-0,005546	0,000494	0,000082	0,000268	0,000339	-0,014311	-0,000156	0,003051
11500	-0,035451	-0,000871	0,000002	1,000000	0,028923	0,029637	0,003630	0,031383	-0,004794	0,029475	0,000119	-0,001531	-0,002312	-0,000001	0,000001	-0,032123	-0,000818	-0,000005	-0,000363	0,003335	0,000004
13200	-0,000042	0,000141	0,000056	0,028923	1,000000	0,000000	0,030973	0,000050	0,000082	-0,000112	-0,000125	-0,000357	0,000864	-0,000203	0,000007	-0,000004	-0,000762	-0,003848	0,000262	-0,008506	-0,000769
13300	-0,290005	-0,003716	0,013564	0,029637	0,000000	1,000000	0,000086	-0,013293	-0,577173	0,973108	0,000029	-0,049215	-0,000203	-0,009027	-0,000018	-0,000091	0,012617	-0,000741	0,000561	0,000335	0,000012
13400	0,000105	-0,000009	0,029601	0,003630	0,030973	0,000086	1,000000	0,000178	-0,000045	0,000457	0,026019	0,000092	-0,007720	0,002086	0,091922	0,033375	0,006849	0,040331	-0,033419	-0,153396	-0,000100
21100	0,110485	0,022869	0,004421	0,031383	0,000050	-0,013293	0,000178	1,000000	0,457906	0,056553	0,591928	-0,008871	-0,008427	-0,008139	0,002590	-0,139166	-0,000172	0,000607	-0,004841	-0,015317	0,000037
21200	-0,267997	0,008896	0,000487	-0,004794	0,000082	-0,577173	-0,000045	0,457906	1,000000	-0,559256	0,000055	0,220249	-0,019087	0,002229	0,000033	-0,483795	-0,000252	0,000204	0,000897	-0,000002	-0,000004
22000	-0,001885	-0,001624	-0,030678	0,029475	-0,000112	0,973108	0,000457	0,056553	-0,559256	1,000000	0,981296	-0,047991	-0,000046	0,000008	-0,000003	-0,000033	-0,000457	-0,000076	0,011999	-0,000112	0,003519
22300	0,313801	0,345436	-0,063763	0,000119	-0,000125	0,000029	0,026019	0,591928	0,000055	0,981296	1,000000	-0,000008	0,014469	-0,123114	-0,000032	-0,000042	-0,000579	-0,000149	0,000230	-0,000080	0,043844
23000	-0,000766	0,000411	0,000031	-0,001531	-0,000357	-0,049215	0,000092	-0,008871	0,220249	-0,047991	-0,000008	1,000000	0,000046	-0,000570	0,000010	0,000842	0,000146	0,000168	0,000008	0,000356	-0,000005
31100	0,015259	0,001456	-0,000162	-0,002312	0,000864	-0,000203	-0,007720	-0,008427	-0,019087	-0,000046	0,014469	0,000046	1,000000	0,011456	-0,049987	-0,197915	0,000766	0,029867	0,163824	-0,162771	-0,113434
32100	-0,038246	0,009521	-0,005546	-0,000001	-0,000203	-0,009027	0,002086	-0,008139	0,002229	0,000008	-0,123114	-0,000570	0,011456	1,000000	-0,069827	0,016552	-0,000233	-0,255580	0,005838	-0,037150	-0,083304
32200	0,005070	0,000779	0,000494	0,000001	0,000007	-0,000018	0,091922	0,002590	0,000033	-0,000003	-0,000032	0,000010	-0,049987	-0,069827	1,000000	0,000010	-0,000150	-0,000137	0,606563	-0,001410	0,011019
32300	0,520125	0,000071	0,000082	-0,032123	-0,000004	-0,000091	0,033375	-0,139166	-0,483795	-0,000033	-0,000042	0,000842	-0,197915	0,016552	0,000010	1,000000	-0,001216	-0,000189	-0,579487	0,000071	-0,007132
32400	0,003634	0,000263	0,000268	-0,000818	-0,000762	0,012617	0,006849	-0,000172	-0,000252	-0,000457	-0,000579	0,000146	0,000766	-0,000233	-0,000150	-0,001216	1,000000	0,001303	-0,000761	0,000313	-0,000232
32500	-0,000511	0,000041	0,000339	-0,000005	-0,003848	-0,000741	0,040331	0,000607	0,000204	-0,000076	-0,000149	0,000168	0,029867	-0,255580	-0,000137	-0,000189	0,001303	1,000000	0,107576	-0,000149	-0,000171
32700	-0,000641	0,007213	-0,014311	-0,000363	0,000262	0,000561	-0,033419	-0,004841	0,000897	0,011999	0,000230	0,000008	0,163824	0,005838	0,606563	-0,579487	-0,000761	0,107576	1,000000	-0,000456	0,005118
32800	0,100712	0,058114	-0,000156	0,003335	-0,008506	0,000335	-0,153396	-0,015317	-0,000002	-0,000112	-0,000080	0,000356	-0,162771	-0,037150	-0,001410	0,000071	0,000313	-0,000149	-0,000456	1,000000	-0,128320
32900	-0,034354	-0,020586	0,003051	0,000004	-0,000769	0,000012	-0,000100	0,000037	-0,000004	0,003519	0,043844	-0,000005	-0,113434	-0,083304	0,011019	-0,007132	-0,000232	-0,000171	0,005118	-0,128320	1,000000